Pure Exploration with Feedback Graphs

Alessio Russo[1] **Yichen Song**[1] **Aldo Pacchiano**[1,2]
[1]Boston University [2] Broad Institute of MIT and Harvard

Abstract

We study the sample complexity of pure exploration in an online learning problem with a feedback graph. This graph dictates the feedback available to the learner, covering scenarios between full-information, pure bandit feedback, and settings with no feedback on the chosen action. While variants of this problem have been investigated for regret minimization, no prior work has addressed the pure exploration setting, which is the focus of our study. We derive an instance-specific lower bound on the sample complexity of learning the best action with fixed confidence, even when the feedback graph is unknown and stochastic, and present unidentifiability results for Bernoulli rewards. Additionally, our findings reveal how the sample complexity scales with key graph-dependent quantities. Lastly, we introduce `TaS-FG` (Track and Stop for Feedback Graphs), an asymptotically optimal algorithm, and demonstrate its efficiency across different graph configurations.

1 INTRODUCTION

Online learning in a stochastic scenario is a sequential decision making problem in which, at each round, the learner chooses one action (arm) out of a finite set of actions, and observes some feedback depending on the setting (Lattimore and Szepesvári, 2020). In the bandit setting the learner observes a random reward distributed according to the distribution of the corresponding action (Robbins, 1952), while in the full information setting the random rewards of all the arms are observed (Littlestone and Warmuth, 1994).

Feedback graphs (Mannor and Shamir, 2011) extend these two settings by defining what type of feedback is available to the learner, effectively bridging scenarios between full-information and pure bandit feedback. The feedback is specified by a directed graph G, with the actions as its vertices, and the edges defining what feedback is revealed to the learner when an action is chosen.

Feedback graphs for online learning have been extensively investigated in the context of regret minimization, a framework used to minimize the hindsight loss relative to optimal strategies (Lattimore and Szepesvári, 2020). Under this framework, the problem has been studied in the case where the learner is *self-aware* or not self-aware.

The self-aware scenario is a specific case where all the vertices have self-loops, and therefore the learner can observe the reward of the chosen action (Mannor and Shamir, 2011; Alon et al., 2017; Arora et al., 2019; Lykouris et al., 2020; Rouyer et al., 2022; Marinov et al., 2022; Kocák and Carpentier, 2023). On the other hand, there are some problems where the learner is not self-aware, such as in the apple tasting problem (Helmbold et al., 2000). Another closely-related problem where the learner is not self-aware is the *revealing action* problem (Cesa-Bianchi and Lugosi, 2006), in which there exists a special action that allows the learner to observe full feedback, while other actions have no feedback. This more general setting has also been studied under the regret minimization framework for both the stochastic and adversarial regimes (Alon et al., 2015; Cohen et al., 2016; Chen et al., 2021; Kong et al., 2022; Eldowa et al., 2023; Zhang et al., 2024a,b).

Feedback graphs have also been analyzed under the *informed* and the *uninformed* settings (Alon et al., 2015). In the former case the graph is revealed to the learner prior to each decision (Arora et al., 2019; Marinov et al., 2022; Zhang et al., 2024b), while in the latter (and harder) case the graph is unknown at decision time (Alon et al., 2017; Zhang et al., 2024a).

Several recent works have considered bandits with stochastic feedback graphs. For instance, (Cortes et al., 2020) and (Esposito et al., 2022) study regret minim-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

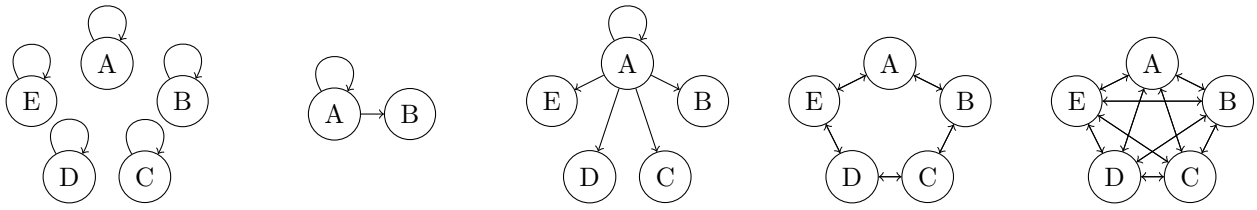

Figure 1: Examples of feedback graphs. From left to right: (1) bandit feedback; (2) apple tasting; (3) revealing action; (4) ring; (5) loopless clique.

ization under stochastic feedback graphs. (Dai et al., 2024) study how stochastic feedback graphs drive user impacts in online platforms.

Prior work, to the best of our knowledge, has mostly focused on regret minimization, whereas the pure exploration problem with fixed confidence (Paulson, 1964; Bechhofer et al., 1968; Bubeck et al., 2011; Kaufmann et al., 2016), where the goal is to determine the best action with a given confidence level δ, has not been thoroughly investigated in the feedback graph setting. In (Du et al., 2021) the authors investigate the pure-exploration problem in combinatorial bandits with partial-linear feedback. However, while their approach is quite general and can be applied to informed settings, it does not exploit the structure of the graph to drive exploration. Chen et al. (2024) study deterministic graphs where the arms are partitioned into T different groups, such that the pull of an arm results in an observation of all the arms in its group. In this setting, they prove a $(\epsilon, 0.05)$-PAC lower bound of order $O\left(\sum_{i=1}^{T} \log(m_i+1)/\epsilon^2\right)$, where m_i is the size of the i-group. Nonetheless, their work does not address the more general pure exploration problem with stochastic feedback graphs.

To address this gap in the literature, we study the pure exploration problem for general stochastic graphs, which may or may not be self-aware. In our model, each edge in the graph has an associated probability of providing feedback, introducing uncertainty about whether feedback is observed from a particular edge. Additionally, we study both the informed scenario, where the graph structure is known and the learner knows the set of edges that were activated, and the uninformed scenarios, where the graph is unknown and the learner does not know which edge has been activated [1].

Using tools from Best Arm Identification (BAI) (Garivier and Kaufmann, 2016), we derive an instance-specific lower bound on the sample complexity of learning the best action with fixed confidence $\delta \in (0,1)$, and present unidentifiability results for Bernoulli rewards in the uninformed case. Additionally, our findings reveal how the sample complexity scales with key graph-dependent quantities, such as the independence number and the number of self-loops in a graph. Lastly, we introduce `TaS-FG` (Track and Stop for Feedback Graphs), an asymptotically optimal algorithm, and demonstrate its efficiency across different graph configurations[2].

2 PROBLEM SETTING

We now briefly explain the graph structure, and some graph-specific quantities, and then explain the BAI setting with feedback graphs.

2.1 Graphs

We indicate by G a generic directed graph with vertices $V = [K]$, where $[K] = \{1, \dots, K\}$ denotes the set of the first K integers. We denote by $G \in [0,1]^{K \times K}$ the matrix of weights of the edges, where $G_{u,v}$ indicates the weight of the edge $(u,v) \in V^2$. Therefore, $E = \{(u,v) \in V^2 : G_{u,v} > 0\}$ is the set of edges in G.

For every $v \in V$ let $N_{in}(v) = \{v' \in V : (v',v) \in E\}$ be the *in-neighborhood* of v in G. Similarly, we let $N_{out}(v) = \{v' \in V : (v,v') \in E\}$ be the *out-neighborhood* of v in G. If a vertex v has a self-loop, that is $(v,v) \in E$, then $v \in N_{in}(v)$ and $v \in N_{out}(v)$. We now define the concept of observability.

Definition 1 (Graph observability)**.** *In a directed graph G with vertices V we say that $v \in V$ is observable if $N_{in}(v) \neq \emptyset$. A vertex $v \in V$ is strongly observable if $\{v\} \subseteq N_{in}(v)$, or $V \backslash \{v\} \subseteq N_{in}(v)$ or both conditions hold. A vertex is weakly observable if it is observable but not strongly. We also let $W(G), SO(G)$ be, respectively, the set of weakly observable and strongly observable vertices. Lastly, a graph G is observable (resp. strongly observable) if all its vertices are observable (resp. strongly observable). A graph is weakly*

[1] We note that in the adversarial literature, the terms *informed* and *uninformed* are used in a slightly different way– typically, they distinguish between observing the feedback graph before versus after making a decision, respectively.

[2] Code repository: https://github.com/rssalessio/Pure-Exploration-with-Feedback-Graphs

observable if it is observable but not strongly.

We also define the notion of *domination* and of *set independence.*

Definition 2 (Domination and set independence)**.** *In a directed graph G with vertices V we say that $D \subseteq V$ dominates $W \subseteq V$ (and we write $W \ll D$) if for any $w \in W$ there exists $d \in D$ such that $w \in N_{out}(d)$. We also say that $I \subseteq V$ is an independent set if it is a collection of vertices with no edges connecting any pair of them.*

Lastly, we define some graph-dependent quantities.

Definition 3 (Graph-dependent quantities)**.** *In a directed graph G with vertices V we let:*

- $\delta(G) = |D(G)|$ *be the weak domination number of G, which is the size of the smallest set $D(G) \in \arg\min_{D \subseteq V : W(G) \ll D} |D|$ that dominates the set of weakly observable vertices $W(G)$.*
- $\alpha(G) = |I(G)|$ *be the independence number of G, which is the largest possible number of vertices in an independent set of G. Formally, $I(G) \in \mathcal{I}(G) \coloneqq \arg\max_{I \subseteq V : \forall (u,v) \in I^2, u \neq v, G_{u,v}=0} |I|$.*
- $\sigma(G) = |L(G)|$ *be the number of vertices with a self-loop, where $L(G) = \{v \in V : \{v\} \subseteq N_{in}(v)\}$ is the set of vertices with a self-loop.*

Example 2.1. *In Fig. 1 we have that: (1) the bandit feedback is strongly observable, with $\alpha(G) = 5$; (2) apple tasting is also strongly observable, with $\alpha(G) = 1$; (3) revealing action is weakly observable, with $\delta(G) = 1, \alpha(G) = 4$; (4) ring graph, which is weakly observable, with $\delta(G) = 3, \alpha(G) = 2$; (5) loopless clique is strongly observable, with $\alpha(G) = 1$. In addition to the graphs in Fig. 1, we also have that: (6) the union of bandit feedback and revealing action is called* loopy star*, a strongly observable graph with $\alpha(G) = 1$; (7) the full feedback graph is strongly observable with $\alpha(G) = 1$.*

2.2 Best Arm Identification with Feedback Graphs

We consider models $\nu = \{G, (\nu_u)_{u \in V}\}$ that consist of a graph $G \in [0,1]^{K \times K}$, with $|V| = K$ vertices, where each vertex $u \in V$ is characterised by ν_u, a probability distribution with average value μ_u.

To investigate the problem, without loss of generality, we analyse the particular case of distributions that belong to a canonical exponential family with one parameter (Efron, 2022):

$$\frac{\mathrm{d}\nu_u}{\mathrm{d}\rho}(x) = \exp(\theta_u x - b(\theta_u)) =: f_u(x), \qquad (1)$$

where f_u is the associated density, $\theta_u \in \Theta \subset \mathbb{R}$ is the canonical parameter (which satisfies $\eta_u = \theta(\mu_u)$ for some mapping $\theta : \mathbb{R} \to \Theta$), ρ is some dominating measure and $b : \Theta \to \mathbb{R}$ is a convex, twice-differentiable function. The mean of a distribution is denoted by $\dot{b}(\theta) \coloneqq \frac{\mathrm{d}b}{\mathrm{d}\theta}(\theta)$, which satisfies $\dot{b}(\theta_u) = \mu_u$. This class of distribution includes the Binomial distribution with n samples, the Poisson distribution, Gaussians with known variance, and others (see also (Cappé et al., 2013; Efron, 2022) for more details). This assumption has been previously used in pure exploration problems, see for example (Garivier and Kaufmann, 2016; Degenne and Koolen, 2019; Degenne et al., 2019).

In the following, with some abuse of notation, we interchangeably write $\nu = \{G, (\mu_u)_{u \in V}\}$, in the sense that a model is identified with the graph G and the mean value of the arms $(\mu_u)_{u \in V}$.

Settings. At each time-step $t = 1, 2, \dots,$ the agent chooses a vertex $V_t \in V$, and observes a collection of independent draws $Z_t \coloneqq \{Z_{t,u}\}_{u \in V}$, where $Z_{t,u} = Y_{t,(V_t,u)} R_{t,u}$, with $Y_{t,(V_t,u)}$ being distributed as $\mathrm{Ber}(G_{V_t,u})$ and the reward $R_{t,u}$ is drawn from ν_u [3]. We focus on the *uninformed* and *informed* settings.

Definition 4 (Uninformed setting)**.** *In the uninformed setting the learner does not know the graph nor which edge is activated at each time-step t. In other words, at time t after choosing $V_t = v$ the learner does not know $E_t \coloneqq \{u \in V : Y_{t,(V_t,u)} = 1\}$.*

A simpler case is the informed setting.

Definition 5 (Informed setting)**.** *In the informed setting the learner at each time-step t knows the graph or which edge was activated after choosing $V_t = v$, i.e., the set E_t is revealed to the learner.*

We denote by $\nu_{v,u}$ the product distribution of $Z_{t,u}$ when $V_t = v$. For simplicity, we also indicate by $Z_{v,u}$ an i.i.d. sample from $\nu_{v,u}$. Lastly, we indicate by $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) the probability law under ν of the observed rewards.

The goal of the learner is to identify the reward associated to the *best vertex* $a^\star(\nu) = \arg\max_{u \in V} \mu_u$, also known as the *best action* or *best arm* (in the following we also write $a^\star$ whenever it's clear from the context). To ensure that the problem is well-defined, for simplicity, we require that the graph is fully observable.

Assumption 1. *We assume ν to be observable.*

We remark that the extension to a non-fully observable graph is straightforward, as long as there are 2 observable vertices (otherwise the problem is ill-defined).

[3] When both random variables follow Bernoulli distributions, one can expect unidentifiability issues if the graph is unknown.

Observe that we do not require to have self-loops in the model compared to some of the previous settings studied in the literature (Kocák and Carpentier, 2023; Mannor and Shamir, 2011).

Algorithm and objective. We consider a broad class of algorithms for the learner that consist of:

1. *a sampling rule*, which determines, based on past observations, which vertex is chosen at time t; that is, V_t is $\mathcal{F}_{t-1}$-measurable, with $\mathcal{F}_t = \sigma(V_1, Z_1, V_2, \dots, V_t, Z_t)$ in the uninformed case, while in the informed one $\mathcal{F}_t = \sigma(V_1, E_1, Z_1, \dots, V_t, E_t, Z_t)$.
2. *a stopping rule* τ that stops the algorithm when sufficient evidence has been gathered to identify the optimal vertex. It is a stopping time with respect to $(\mathcal{F}_t)_t$ satisfying $\mathbb{P}_\nu(\tau < \infty) = 1$.
3. *a recommendation rule* $\hat{a}_\tau \in V$ that returns the estimated optimal vertex, and $\hat{a}_\tau$ is a $\mathcal{F}_\tau$-measurable random variable.

We focus on the fixed-confidence setting, with a risk parameter[4] $\delta \in (0,1)$, which entails devising an probably-correct estimator of the best vertex.

Definition 6 (δ-PC Algorithm)**.** *We say that an algorithm* `Alg` *is δ-PC (Probably Correct) if, for any model ν satisfying Assumption 1, we have $\mathbb{P}_\nu(\tau < \infty, \hat{a}_\tau \neq a^\star(\nu)) < \delta$.*

The goal in this setting is to obtain a δ-PC algorithm that requires, on average, the minimum number of draws $\mathbb{E}_\nu[\tau]$. Therefore we study the minimum achievable sample complexity $\mathbb{E}_\nu[\tau]$ by any δ-PC algorithm.

Notation. In the following, we denote by $\mathrm{KL}(P,Q)$ the KL-divergence between two distributions P and Q, and by $\mathrm{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y))$ the Bernoulli KL-divergence between two Bernoulli distributions of parameters x and y respectively. For distributions P, Q belonging to the canonical exponential family with one parameter, with canonical parameters θ, θ' respectively, we have $\mathrm{KL}(P,Q) = b(\theta') - b(\theta) - \dot{b}(\theta)(\theta' - \theta)$. We also define a generalized version of the Jensen-Shannon divergence as $I_\alpha(P,Q) = \alpha\mathrm{KL}(P, \alpha P + (1-\alpha)Q) + (1-\alpha)\mathrm{KL}(Q, \alpha P + (1-\alpha)Q)$ with $\alpha \in [0,1]$. The sub-optimality gap in a vertex $u \in V$ is defined as $\Delta_u := \mu_{a^\star} - \mu_u$, and the minimum gap as $\Delta_{\min} = \min_{a \neq a^\star} \Delta_a$.

Finally, we let $N_v(t)$ be the number of times a vertex $v \in V$ has been chosen up to time-step t by `Alg` (to not be confused with the in/out-neighborhoods $N_{in}(v)$ and $N_{out}(v)$), thus $N_v(t) = N_v(t-1) + \mathbf{1}_{\{V_t = v\}}$, with $N_v(0) = 0$. We also indicate by $N_{v,u}(t)$ the number of times edge (v,u) was activated after choosing v up to time-step t, thus $N_{v,u}(t) = N_{v,u}(t-1) + \mathbf{1}_{\{V_t = v, Y_{t,(v,u)} = 1\}}$ with $N_{v,u}(0) = 0$. Similarly, we denote by $M_u(t)$ the number of times we observed a reward from vertex u up to time-step t. Hence, one can write $M_u(t) = \sum_{v \in V} N_{v,u}(t)$.

[4] Note that δ is the confidence parameter, while $\delta(G)$ is the weak domination parameter.

3 SAMPLE COMPLEXITY LOWER BOUNDS

The recipe to derive instance-specific sample complexity lower bounds is based on a *change of measure* argument. This argument allows to derive an instance-dependent quantity $T^\star(\nu)$, also known as *characteristic time* (and its inverse $(T^\star(\nu))^{-1}$ is the *information rate*), that permits to lower bound the sample complexity of an algorithm. Change of measure arguments have a long history (Wald, 1947; Lorden, 1971; Lai, 1981; Lai and Robbins, 1985), and have been applied to find lower bounds for regret minimization (Combes and Proutiere, 2014; Garivier et al., 2019) and best-arm identification (Garivier and Kaufmann, 2016). We use this technique to derive the sample complexity lower bound in both the uninformed and informed settings.

3.1 Lower Bound in the Uninformed Setting

The uninformed case, while seemingly daunting, admits a separation in behavior between the class of *continuous* and *discrete* rewards.

We find that for Bernoulli rewards the best vertex is *unidentifiable*, i.e., we cannot reject the null hypothesis that a certain vertex is optimal, no matter how much data is gathered. This result comes from the fact that an agent cannot discern between low-probability edges linked to high-reward vertices, and vice-versa.

For the continuous case, however, we find a different behavior. The intuition is that an agent can conclude, with almost sure certainty, that an edge is not activated if zero reward is observed from that edge.

3.1.1 Lower bound for continuous rewards in the uninformed setting

Recall the definition of the generalized Jensen-Shannon divergence $I_\alpha(P,Q) := \alpha\mathrm{KL}(P, \alpha P + (1-\alpha)Q) + (1-\alpha)\mathrm{KL}(Q, \alpha P + (1-\alpha)Q)$ for two distributions P, Q and $\alpha \in [0,1]$.

For continuous rewards, we obtain the following instance-specific sample complexity lower bound, which is proved in Appendix C.1.

Theorem 1. *For any δ-PC algorithm and any model ν with reward distributions $\{\nu_u\}_{u\in V}$ with continuous support, satisfying Assumption 1, we have that*

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu)\mathrm{kl}(\delta, 1-\delta), \tag{2}$$

where

$$(T^\star(\nu))^{-1} = \sup_{\omega\in\Delta(V)} \min_{u\neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u)$$
$$s.t.\ m_u = \sum_{v\in N_{in}(u)} \omega_v G_{v,u} \quad \forall u \in V.$$

In the following we also denote by $\omega^\star = \arg\inf_{\omega\in\Delta(V)} T(\omega;\nu)$ the optimal solution, where

$$T(\omega;\nu)^{-1} = \min_{u\neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u), \quad \text{s.t. } m = G^\top\omega, \tag{3}$$

where $m = \begin{bmatrix} m_1 & \dots & m_K \end{bmatrix}^\top$, similarly ω, are in vector form.

Discussion and scaling. The characteristic time in Theorem 1 displays some similarities to the characteristic time found in classical BAI (Garivier and Kaufmann, 2016). First, as one would expect, the amount of evidence $(m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u)$ does not depend directly on the vertex selection rate ω, but on the observation rate m, which depends on the edge activation probabilities.

Secondly, to gain a better intuition of the above quantities, we can focus on the Gaussian case where $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$, with $\lambda > 0$. For this particular choice, $T^\star(\nu)$ is the solution to the following convex problem:

$$T^\star(\nu) = \inf_{\omega\in\Delta(V)} \max_{u\neq a^\star} \left(m_u^{-1} + m_{a^\star}^{-1}\right) \frac{2\lambda^2}{\Delta_u^2} \text{ s.t. } m = G^\top\omega. \tag{4}$$

This expression allows us to gain a better understanding of the scaling of $T^\star(\nu)$, as shown in the next two propositions (which are proved in Appendix C.3).

Proposition 1. *Consider an observable model $\nu = (\{\nu_u\}_u, G)$ with Gaussian rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$. If $\delta(G) + \sigma(G) > 0$, we can upper bound $T^\star$ as*

$$T^\star(\nu) \leq \frac{4\left(\delta(G) + \sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor\right)\lambda^2}{\min_{u\neq a^\star} \min(\bar{G}_u, \bar{G}_{a^\star})\Delta_u^2}, \tag{5}$$

where $\bar{G}_u$ for any $u \in V$ is defined as

$$\bar{G}_u := \max\left(\max_{v\in D(G)} G_{v,u}, \min_{v\in L(G): G_{v,u}>0} G_{v,u}\right).$$

Instead, for the loopless clique, we obtain the following result.

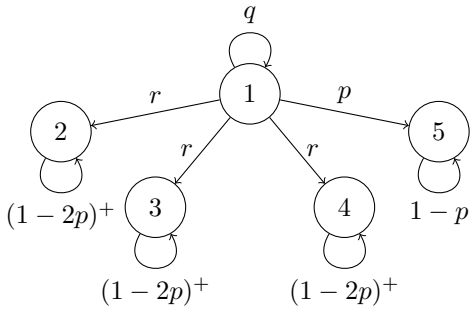

Figure 2: Loopy star graph. To each edge is associated an activation probability (obs. that $(x)^+ = \max(x, 0)$).

Proposition 2. *For an observable $\nu = (\{\nu_u\}_u, G)$ with Gaussian rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$ satisfying $\delta(G) + \sigma(G) = 0$ (i.e., the loopless clique), we have*

$$T^\star(\nu) \leq \frac{4\bar{G}\lambda^2}{\Delta_{\min}^2}, \tag{6}$$

where $\bar{G} := \min_{v,w:v\neq w} \max_{u\neq a^\star} \frac{1}{G_{v,w}(u)} + \frac{1}{G_{v,w}(a^\star)}$ and $G_{v,w}(u) := G_{v,u} + G_{w,u}$.

For both results we see how the sample complexity does not scale with K, but depends on the structural properties of G. For example, Proposition 1 exhibits a scaling in $\sigma(G)$ for bandit feedback, and $\sigma(G)/2$ for a complete graph (i.e., there is an edge between any pair of vertices) with self-loops. Nevertheless, we find it challenging to enhance this scaling without using additional graph-specific parameters.

Example 3.1 (The loopy star)**.** *We study the graph in Fig. 2 with Gaussian rewards, where $\lambda = 1$, $\mu_5 = 1$ and $\mu_u = 0.5, u \in \{1, \dots, 4\}$. Notably, this graph is the union of a bandit feedback graph and revealing action graph. This example is relevant in adversarial regret minimization: removing any self-loop changes the minimax regret from $\tilde{\Theta}(\sqrt{\alpha(G)T})$ to $\tilde{\Theta}(T^{2/3})$ (Alon et al., 2015). The graph depends on the parameters (p, q, r), where the number of self-loops decreases as p increases. In Fig. 3 we show the characteristic time $T^\star(\nu)$ (in solid lines) of the loopy star graph for different values of (p, q) with $r = 1/4$. In particular, there is no sharp transition as in the regret minimization setting when removing self-loops: since exploitation is unnecessary, selecting a less promising node is acceptable as long as it yields useful information. Additionally, the only visible change happens at $p \approx 0.1$ (and $q = 1$ is fixed), where it is no longer convenient for the algorithm to sample the vertices $\{2, 3, 4\}$, and focuses only on vertices 1 and 5. In the figure we also plot (in dashed lines) $\|G^\top\omega^\star\|_2$, an indication of the observation frequency of the vertices. We see how this quantity is directly correlated with the characteristic time $T^\star(\nu)$.*

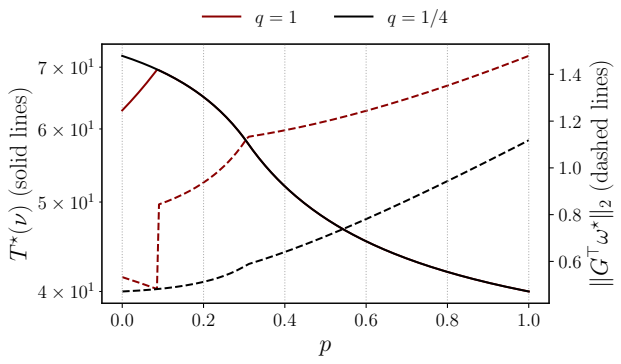

Figure 3: Loopy star example with $r = 1/4$. The solid lines depict $T^\star(\nu)$ for $q = 1$ and $q = 1/4$ for different values of p. Similarly, on the right axis, the dashed lines show $\|G^\top \omega^\star\|_2$, which indicates the amount of information gathered per time-step.

3.1.2 A heuristic solution

In general, it is difficult to guess what the optimal solution $\omega^\star$ may be. Garivier and Kaufmann (2016) show that $\omega_u \propto 1/\Delta_u^2$ is almost optimal for Gaussian rewards (with $\Delta_{a^\star} = \Delta_{\min}$) in multi-armed bandit problems. Is it the same also for feedback graphs?

Taking inspiration from Garivier and Kaufmann (2016), we propose that $m_u \propto 1/\Delta_u^2$, where $m = G^\top \omega$. Define then the vector $\Delta^{-2} := (1/\Delta_u^2)_{u \in V}$. From a geometrical perspective, we can maximize the similarity $m^\top \Delta^{-2}$, or rather, $\omega^\top G \Delta^{-2}$. In the classical Euclidean space this is achieved by $\omega \propto G\Delta^{-2}$. To obtain a distribution, we project $G\Delta^{-2}$ to the nearest distribution ω_{heur} in the KL sense, and obtain $\omega_{\text{heur}} := G\Delta^{-2}/\|G\Delta^{-2}\|_1$ (see also Appendix C.3.3).

Such allocation ω_{heur} makes intuitive sense: the probability of selecting a vertex u is proportional to $\sum_{v \in V} G_{uv} \Delta_v^{-2}$, thus assigning higher preference to vertices that permit the learner to sample actions with small sub-optimality gaps.

For this heuristic allocation ω_{heur} we can provide the following upper bound on its scaling (we refer the reader to Appendix C.3.3 for a proof).

Proposition 3. *For an observable model $\nu = (\{\nu_u\}_u, G)$ with Gaussian random rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$ we can upper bound $T(\omega_{\text{heur}}; \nu)$ as*

$$T^\star(\nu) \le T(\omega_{\text{heur}}; \nu) \le 4\lambda^2 \frac{\|G\Delta^{-2}\|_1}{\sigma_{\min}(G)^2},$$

where $\sigma_{\min}(G)$ is the minimum singular value of G.

Observing the following upper bound $||G\Delta^{-2}||_1 \le K^{3/2}\sigma_{\max}(G)/\Delta_{\min}^2$ where $\sigma_{\max}(G)$ denotes the maximum singular value of G, we conclude that in the worst case scenario $T(\omega_{\text{heur}}; \nu) \le O\left(\frac{K^{3/2}\sigma_{\max}(G)}{\Delta_{\min}^2 \sigma_{\min}(G)^2}\right)$. Observe that this scaling can be improved in K for sparse solutions that efficiently use the graph structure. We discuss in Appendix C.3.3 such an approach that scales according to the minimum number of vertices that dominate the graph.

Another simple solution is the sparse allocation $\omega_u = \mathbf{1}_{\{u \in \mathcal{G}\}}/|\mathcal{G}|$, where $\mathcal{G} = \arg\max_v (G\Delta^{-2})_v$. This is an efficient allocation, since it scales as $O\left(\frac{|\mathcal{G}|}{\Delta_{\min}^2 \max_{u \in \mathcal{G}} \min_v G_{u,v}}\right)$. However, such solution is admissible only if $\mathcal{G}$ dominates the graph, which ultimately depends on the structure of G.

3.1.3 The curious case of Bernoulli rewards

We conclude the study of sample complexity lower bounds in the uninformed setting by examining the case of Bernoulli rewards, finding that it is generally impossible to estimate the best vertex in this case. If the learner does not know whether the graph activations are deterministic and the rewards follow a Bernoulli distribution, a zero outcome could arise either because the reward of the chosen arm is truly zero or because the edge responsible for providing feedback was not activated. Without knowing which edge was activated, the learner cannot discern between the randomness of the reward and that of the edge. This intuition is formalized in the following result, which shows that the information rate is zero (the proof is given in Appendix C.1.4).

Proposition 4. *Under Assumption 1, if $(\nu_u)_{u \in V}$ are Bernoulli distributions with parameters $(\mu_u)_{u \in V}$, then $a^\star$ is unidentifiable, in the sense that $(T^\star(\nu))^{-1} = 0$.*

3.2 Lower Bound in the Informed Setting

When the set of activated edges E_t is revealed to the learner at each time-step t, we can show that the lower bound in Theorem 1 also holds for Bernoulli rewards. Furthermore, the result also applies to the case where the graph probabilities G are known a-priori.

Theorem 2. *Consider a δ-PC algorithm and a model ν satisfying Assumption 1. If the set E_t is revealed to the learner for each $t \ge 1$ (or G is known), then*

$$\mathbb{E}_\nu[\tau] \ge T^\star(\nu) \mathrm{kl}(\delta, 1-\delta), \tag{7}$$

where $T^\star(\nu)$ is as in Theorem 1.

See the proof in Appendix C.2. Henceforth, similar remarks from the uninformed case also apply here.

4 TaS-FG ALGORITHM

In this section we propose TaS-FG (Track and Stop for Feedback Graphs), an algorithm inspired by Track

and Stop (TaS, Garivier and Kaufmann (2016)) that is asymptotically optimal as $\delta \to 0$ with respect to $T^\star(\nu)$ for both the *informed* and *uninformed* cases. The algorithm consists of: (1) the model estimation procedure and recommender rule; (2) the sampling rule, dictating which vertex to select at each time-step; (3) the stopping rule, defining when enough evidence has been collected to identify the best vertex with sufficient confidence, and therefore to stop the algorithm.

4.1 Estimation Procedure and Recommender Rule

The algorithm maintains a maximum likelihood estimate $\hat{\nu}(t) = (\hat{G}(t), \hat{\mu}(t))$ of the model. Using these estimates we define the estimated optimal vertex at time t as $\hat{a}_t = \arg\max_a \hat{\mu}_a(t)$, and the estimated sub-optimality gap in $u \neq \hat{a}_t$ as $\hat{\Delta}_u(t) = \hat{\mu}_{\hat{a}_t}(t) - \hat{\mu}_u(t)$, and $\hat{\Delta}_{\hat{a}_t}(t) = \hat{\Delta}_{\min}(t) \coloneqq \min_{u \neq \hat{a}_t} \hat{\Delta}_u(t)$. The recommender rule at the stopping time τ is defined as $\hat{a}_\tau = \arg\max_{a \in V} \hat{\mu}_a(\tau)$. Observe that we differentiate between the informed and the uninformed cases.

Informed case. In the informed case, the model parameters can simply be estimated as $\hat{\mu}_u(t) = \frac{1}{M_u(t)} \sum_{n=1}^t R_{n,u}$, with $M_u(t) = \sum_{v \in V} N_{v,u}(t)$ and $N_{v,u}(t) = N_{v,u}(t-1) + \mathbf{1}_{\{V_t = v, u \in E_t\}}$. If the graph is unknown, the estimator of G is $\hat{G}_{v,u}(t) = \frac{N_{v,u}(t)}{N_v(t)}$.

Uninformed case. In the uninformed case, we focus on continuous rewards. We employ the fact that observing zero reward has measure zero, and therefore one can define $N_{v,u}(t) = N_{v,u}(t-1) + \mathbf{1}_{\{V_t = v, Z_t \neq 0\}}$ to obtain an unbiased estimator of the number of times edge (v,u) was activated. Henceforth, one can define the estimator $\hat{\mu}(t), \hat{G}_{v,u}(t)$ as in the observable case.

4.2 Sampling Rule

Interestingly, to design an algorithm with minimal sample complexity, we can look at the solution $\omega^\star = \arg\inf_{\omega \in \Delta(V)} T(\omega; \nu)$.

The solution $\omega^\star$ provides the best proportion of draws, that is, an algorithm selecting a vertex $u \in V$ with probability $\omega_u^\star$ matches the lower bound in Theorem 1 and is therefore optimal with respect to $T^\star(\nu)$. Therefore, an idea is to ensure that N_t/t tracks $\omega^\star$, where N_t is the visitation vector $N(t) \coloneqq \begin{bmatrix} N_1(t) & \dots & N_K(t) \end{bmatrix}^\top$.

However, the instance ν is initially unknown. Common algorithms in the best arm identification literature (Garivier and Kaufmann, 2016; Kaufmann et al., 2016), or in best policy identification problems (Al Marjani et al., 2021; Russo and Vannella, 2024), track an estimated optimal allocation $\omega^\star(t) = \arg\inf_{\omega \in \Delta(V)} T(\omega; \hat{\nu}(t))$ using the current estimate of the model $\hat{\nu}(t) = (\hat{G}(t), \{\hat{\mu}_u(t)\}_u)$ (see line 4 in Algorithm 1). This technique is also known as *certainty-equivalence* principle (Jedra and Proutiere, 2020).

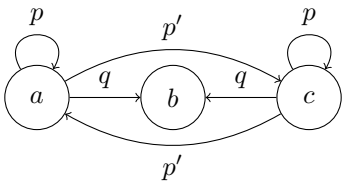

Figure 4: Example of symmetric feedback graph where the solution set $C^\star(\nu) = \arg\inf_{w \in \Delta(V)} T(w; \nu)$ is not unique for $\mu_a = \mu_c$ and $a^\star(\mu) = b$.

To apply this principle, the learner needs to at-least make sure to use a sampling rule guaranteeing that $\hat{\nu}(t) \to \nu$. Common techniques employ some form of *forced exploration* to ensure convergence, for example by introducing uniform noise in the action selection process. In general, there are two families of tracking procedures: the C-tracking and D-tracking rules introduced in (Garivier and Kaufmann, 2016):

- *C-tracking*: compute $\tilde{\omega}^\star(t)$, which is the ℓ_∞ projection of $\omega^\star(t)$ on $\Delta_t(V) = \{\omega \in \Delta(V) : \forall u \in V, \omega_u \geq \epsilon_t\}$ for some $\epsilon_t > 0$. Then the action selection is $u_t = \arg\min_{u \in V} N_t(u) - \sum_{n=1}^t \tilde{\omega}_u^\star(t)$.
- *D-tracking*: if there is a vertex u with $N_u(t) \leq \sqrt{t} - K/2$ then choose $u_t = u$. Otherwise, choose the vertex $u_t = \arg\min_{u \in V} N_u(t) - t\omega_u^\star(t)$. Compared to C-tracking, this rule enjoys deterministic guarantees on the visitation frequency of order $\sqrt{t}$.

Uniqueness of the solution $\omega^\star$. These approaches guarantee that N_t/t approximately track $\omega^\star(t)$, which ultimately should converge to $\omega^\star$. Unfortunately, for the problem studied in this manuscript, the optimal solution $\omega^\star$ *may not be unique*, i.e., the set $C^\star(\nu) = \arg\inf_{\omega \in \Delta(V)} T(\omega; \nu)$ may admit more than one optimal solution. This may be an issue if we want N_t/t to track $\omega^\star$. An example of graph with non-unique solution is the following one.

Example 4.1 (Multiple optimal allocations)**.** *Consider the example in Fig. 4 with Gaussian rewards $(\mathcal{N}(\mu_u, 1))_{u \in \{a,b,c\}}$. This is an example of symmetric graph where vertex b is optimal, and the solution set $C^\star(\nu) = \arg\inf_{\omega \in \Delta(V)} T(\omega; \nu)$ is actually a convex set. In fact, for any $x \in [0,1]$ the solution $\omega(x) = \begin{bmatrix} x & 0 & 1-x \end{bmatrix}$ is optimal. As an example, for $p = p' = 1/2, q = 1$ and $\mu_a = \mu_c = 0, \mu_b = 1$ we obtain $T^\star(\nu) = T(\omega; \nu) = 6$ for any $\omega \in \{\omega(x) : x \in [0,1]\}$.*

In literature, some strategies have been proposed to deal with the aforementioned issue. Russo and Proutiere (2023) regularize $T^\star(\omega;\nu)$ to make it strongly convex in ω, and obtain a sample complexity upper bound that is asymptotically optimal in δ and the regularization parameter. In (Jedra and Proutiere, 2020; Degenne and Koolen, 2019) they use an approach inspired by Berge (1963)'s Maximum theorem and the fact that $C^\star(\nu)$ is a convex set. Using this fact, they show that with an appropriate tracking rule it is possible to ensure $\lim_{t\to\infty} \inf_{\omega\in C^\star(\nu)} \|\omega - N(t)/t\|_\infty = 0$.

Based on this result, Jedra and Proutiere (2020) propose an averaged D-tracking sampling rule for linear multi-armed bandit problem, while Degenne and Koolen (2019) opt for a C-tracking procedure claiming that D-tracking may fail to converge in classical bandit problems. Nonetheless, we can guarantee convergence using an averaged variant of D-tracking similarly to (Jedra and Proutiere, 2020).

Proposition 5. *Let* $S_t = \{u \in V : N_u(t) < \sqrt{t} - K/2\}$. *The averaged D-tracking rule, defined as*

$$V_t \in \begin{cases} \arg\min_{u\in S_t} N_u(t) & S_t \neq \emptyset \\ \arg\min_{u\in V} N_u(t) - \sum_{n=1}^t \omega_u^\star(n) & \text{otherwise} \end{cases}, \tag{8}$$

ensures that $\lim_{t\to\infty} \inf_{\omega\in C^\star(\nu)} \|N(t)/t - \omega\|_\infty \to 0$ *a.s.*

The proof for the above proposition can be found in Appendix D.1.

4.3 Stopping Rule and Sample Complexity Guarantees

The stopping rule determines when enough evidence has been collected to determine the optimal action with a prescribed confidence level. The problem of determining when to stop can be framed as a statistical hypothesis testing problem (Chernoff, 1959), where we are testing between K different hypotheses $(\mathcal{H}_u : (\mu_u > \max_{v\neq u} \mu_v))_{u\in V}$.

We consider the following statistic in line 3 of Algorithm 1

$$L(t) = tT(N(t)/t; \hat\nu(t))^{-1}, \tag{9}$$

which is a Generalized Likelihood Ratio Test (GLRT), similarly as in (Garivier and Kaufmann, 2016). Comparing with Theorem 1, one needs to stop as soon as $L(t) \geq \mathrm{kl}(\delta, 1-\delta) \sim \ln(1/\delta)$. However, to account for the random fluctuations, a more natural threshold is $\beta(t,\delta) = \ln((1+\ln(t))/\delta)$. Several thresholds have been proposed in literature (Magureanu et al., 2014; Garivier and Kaufmann, 2016). We employ the following threshold

$$\beta(t,\delta) := 2\mathcal{C}_{\exp}\left(\frac{\ln\left(\frac{K-1}{\delta}\right)}{2}\right) + 6\ln(1+\ln(t)), \tag{10}$$

where $\mathcal{C}_{\exp}(x) \approx x + 4\ln(1+x+\sqrt{2x})$ for $x \geq 5$. A precise definition of $\mathcal{C}_{\exp}(x)$ can be found in (Kaufmann and Koolen, 2021, Theorem 7), or in Appendix D.2. For this threshold, we obtain the following guarantee, proved in Appendix D.2.

Algorithm 1 TaS-FG (TaS for Feedback Graphs)

1: **Input:** confidence δ.
2: Set $t \leftarrow 1$
3: **while** $L(t) < \beta(\delta, t)$ **do**
4: Compute $\omega^\star(t) = \arg\inf_{\omega\in\Delta(V)} T(\omega; \hat\nu(t))$.
5: Select V_t according to the D-tracking rule in Proposition 5 and observe Z_t.
6: Update statistics $N(t), M(t), \hat\mu(t), \hat G(t)$ as explained in Section 4.1 and set $t \leftarrow t+1$.
7: **end while**
8: **Return** $\hat a_\tau = \arg\max_{a\in V} \hat\mu_a(\tau)$

Proposition 6. *Using the threshold function defined in Eq. (10) guarantees that* TaS-FG *is* δ*-PC, i.e.*

$$\mathbb{P}_\nu(\tau < \infty, \hat a_\tau \neq a^\star(\mu)) \leq \delta.$$

Additionally, we have the following sample complexity optimality guarantees, proved in Appendix D.3.

Theorem 3. *For all* $\delta \in (0, 1/2)$, TaS-FG *(1) terminates a.s.* $\mathbb{P}_\nu(\tau < \infty) = 1$*; (2) is a.s. asymptotically optimal* $\mathbb{P}_\nu\left(\limsup_{\delta\to 0} \frac{\tau}{\ln(1/\delta)} \leq T^\star(\nu)\right) = 1$*; (3) is optimal in expectation* $\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\ln(1/\delta)} \leq T^\star(\nu)$.

Heuristic TaS-FG**.** Now consider TaS-FG with the heuristic solution $\omega_{\mathrm{heur}}(t) := \frac{\hat G(t)\hat\Delta^{-2}(t)}{\|\hat G(t)\hat\Delta^{-2}(t)\|_1}$ (where $\hat G(t)$ is the estimated graph at time t, and $\hat\Delta^{-2}(t) = (1/\hat\Delta_u^2(t))_{u\in V}$). That is, we let $\omega^\star(t) = \omega_{\mathrm{heur}}(t)$ in line 4 of Algorithm 1. Then we obtain the following guarantees, proved in Appendix D.3.

Corollary 1. TaS-FG *with* $\omega^\star(t) = \omega_{\mathrm{heur}}(t)$ *is* δ*-PC, and guarantees that (1) the algorithm stops a.s.; (2)* $\mathbb{P}_\nu\left(\limsup_{\delta\to 0} \frac{\tau}{\ln(1/\delta)} \leq T(\omega_{\mathrm{heur}};\nu)\right) = 1$ *and (3)* $\limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\ln(1/\delta)} \leq T(\omega_{\mathrm{heur}};\nu)$.

5 NUMERICAL RESULTS

In this section, we present the numerical results of TaS-FG and other algorithms on various graph configurations, with varying graph sizes. Due to lack of

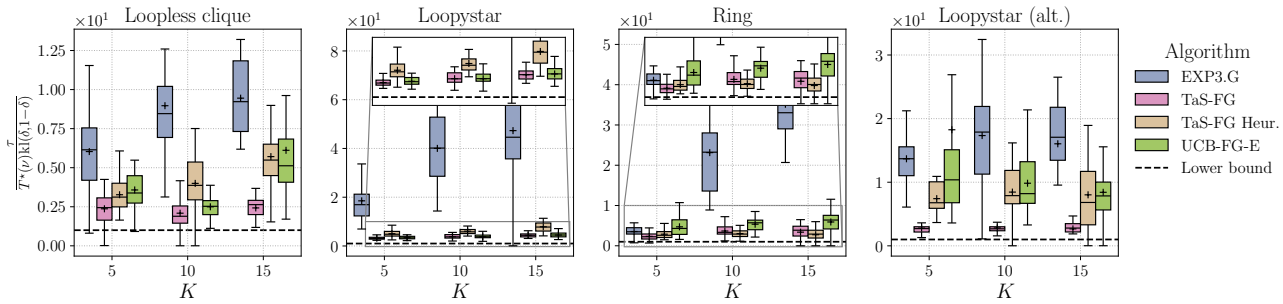

Figure 5: Box plots of the normalized sample complexity $\frac{\tau}{T^\star(\nu)\mathrm{kl}(\delta,1-\delta)}$ for $\delta = e^{-7}$ over 100 seeds. Boxes indicate the interquartile range, while the median and mean values are, respectively, the solid line and the + sign in black.

space, we only present results for $\delta \approx 10^{-3}$, and we refer the reader to Appendix A for more numerical results, including evaluation on different values of δ.

Algorithms. We compare TaS-FG to heuristic TaS-FG (i.e., TaS-FG with the heuristic solution proposed in Section 3.1.2) and to EXP3.G (Alon et al., 2015), an algorithm for regret minimization in the adversarial case (which is similar to most algorithms designed to handle feedback graphs). We also compare with two variants of UCB: (1) UCB-FG-E, which acts greedily with respect to the upper confidence bound of $\left(\hat{G}(t)\hat{\mu}(t)\right)$; and (2) UCB-FG-V, which selects $\arg\max_v \hat{G}^{\mathrm{ucb}}_{v,\hat{a}^{\mathrm{ucb}}(t)}(t)$, where $\hat{G}^{\mathrm{ucb}}(t), \hat{a}^{\mathrm{ucb}}(t)$ are, respectively, the UCB estimates of G and $a^\star$ at time t. In our experiments, we used the same stopping rule for all algorithms, as the guarantees of Proposition 6 hold for any sampling rule. However, we observed that the rate $\beta(t,\delta)$ given in Eq. (10) is relatively conservative. Therefore, we opted for a less conservative threshold that still achieves satisfactory error rates, that is $\beta(t,\delta) = \ln(1/\delta) + 3\ln(1+2\ln(t))$.

Results. The main sample complexity results are depicted in Fig. 5 (results are omitted for UCB-FG-V due to its much larger sample complexity; we report those in Appendix A). We evaluated these algorithms on several graphs: **(1)** the loopystar (Fig. 2) with $(p,q,r) = (1/5, 1/4, 1/4)$ and **(2)** an alternative version (denoted by 'Loopystar (alt.)' in Fig. 5) where $\mu_1 = 1$ is optimal and $\mu_u = 0.5$ for all other vertices, with $(p,q,r) = (0, 1/4, 1/[8(K-1)])$; **(3)** the ring and **(4)** loopless clique graphs (see, respectively, Fig. 7 and Fig. 8 in Appendix A for more details). Overall, we observe the good performance of TaS-FG and heuristic TaS-FG, while EXP3.G is not as performant. Notably, UCB-FG-E also seems to achieve good sample complexity, comparable to heuristic TaS-FG, although it may not perform as well when $(G\mu)_{a^\star}$ is small in comparison to that of other vertices.

6 CONCLUSIONS

In this paper we characterized the sample complexity of identifying the best action in an online learning problem with a feedback graph. In this setting, the graph dictates what feedback the learner observes, with the additional complexity that the graph may be unknown and have stochastic activations (a.k.a. *uninformed setting*). For such a setting, we showed that for Bernoulli rewards the best action remains unidentifiable. In contrast, for continuous rewards, we derived an instance-specific lower bound applicable to all δ-PC algorithms, a result that also extends to the informed setting where the graph or activations are known. Building on this lower bound, we introduced TaS-FG, an algorithm achieving asymptotically optimal sample complexity in both settings, for which we demonstrated its efficiency numerically on several graph configurations.

Acknowledgments

The authors are pleased to acknowledge that the computational work reported on in this paper was performed on the Shared Computing Cluster, which is administered by Boston University's Research Computing Services.

References

Al Marjani, A., Garivier, A., and Proutiere, A. (2021). Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. (2017). Nonstochastic

multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826.

Arora, R., Marinov, T. V., and Mohri, M. (2019). Bandits with feedback graphs and switching costs. *Advances in Neural Information Processing Systems*, 32.

Bechhofer, R. E., Kiefer, J., and Sobel, M. (1968). Sequential identification and ranking procedures. *The university of Chicago Press*.

Berge, C. (1963). Topological spaces. *Proceedings of the Edinburgh Mathematical Society*, 13:339 – 339.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chen, H., He, Y., and Zhang, C. (2024). On interpolating experts and multi-armed bandits. In *International Conference on Machine Learning*, pages 6776–6802. PMLR.

Chen, H., Li, S., Zhang, C., et al. (2021). Understanding bandits with graph feedback. *Advances in Neural Information Processing Systems*, 34:24659–24669.

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

Cohen, A., Hazan, T., and Koren, T. (2016). Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pages 811–819. PMLR.

Combes, R. and Proutiere, A. (2014). Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529. PMLR.

Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., and Zhang, N. (2020). Online learning with dependent stochastic feedback graphs. In *International Conference on Machine Learning*, pages 2154–2163. PMLR.

Dai, J., Flanigan, B., Jagadeesan, M., Haghtalab, N., and Podimata, C. (2024). Can probabilistic feedback drive user impacts in online platforms? In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR.

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

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

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Du, Y., Kuroki, Y., and Chen, W. (2021). Combinatorial pure exploration with full-bandit or partial linear feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7262–7270.

Efron, B. (2022). *Exponential families in theory and practice*. Cambridge University Press.

Eldowa, K., Esposito, E., Cesari, T., and Cesa-Bianchi, N. (2023). On the minimax regret for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 36.

Esposito, E., Fusco, F., van der Hoeven, D., and Cesa-Bianchi, N. (2022). Learning on the edge: Online learning with stochastic feedback graphs. *Advances in Neural Information Processing Systems*, 35:34776–34788.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR.

Garivier, A., Ménard, P., and Stoltz, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399.

Gurobi Optimization, LLC (2024). Gurobi Optimizer Reference Manual.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*.

Hartung, J., Knapp, G., and Sinha, B. K. (2011). *Statistical meta-analysis with applications*. John Wiley & Sons.

Helmbold, D. P., Littlestone, N., and Long, P. M. (2000). Apple tasting. *Information and Computation*, 161(2):85–139.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*.

Jedra, Y. and Proutiere, A. (2020). Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017.

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

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

Kocák, T. and Carpentier, A. (2023). Online learning with feedback graphs: The true shape of regret. In *International Conference on Machine Learning*, pages 17260–17282. PMLR.

Kong, F., Zhou, Y., and Li, S. (2022). Simultaneously learning stochastic and adversarial bandits with general graph feedback. In *International Conference on Machine Learning*, pages 11473–11482. PMLR.

Lai, T. L. (1981). Asymptotic optimality of invariant sequential probability ratio tests. *The Annals of Statistics*, pages 318–333.

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Information and computation*, 108(2):212–261.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The annals of mathematical statistics*, pages 1897–1908.

Lykouris, T., Tardos, E., and Wali, D. (2020). Feedback graph regret bounds for thompson sampling and ucb. In *Algorithmic Learning Theory*, pages 592–614. PMLR.

Magureanu, S., Combes, R., and Proutiere, A. (2014). Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR.

Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24.

Marinov, T. V., Mohri, M., and Zimmert, J. (2022). Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. *Advances in Neural Information Processing Systems*, 35:24947–24959.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, pages 174–180.

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

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. (2022). A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 35:35035–35048.

Russo, A. and Proutiere, A. (2023). On the sample complexity of representation learning in multi-task bandits with global and local structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9658–9667.

Russo, A. and Vannella, F. (2024). Multi-reward best policy identification. *Advances in Neural Information Processing Systems*, 37:105583–105662.

Van Rossum, G. and Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., Ram, Y., Yarkoni, T., Williams, M. L., Evans, C., Fitzgerald, C., Brian, Fonnesbeck, C., Lee, A., and Qalieh, A. (2017). mwaskom/seaborn: v0.8.1.

Zhang, M., Zhang, Y., Luo, H., and Mineiro, P. (2024a). Efficient contextual bandits with uninformed feedback graphs. In *International Conference on Machine Learning*, pages 60336–60358. PMLR.

Zhang, M., Zhang, Y., Vrousgou, O., Luo, H., and Mineiro, P. (2024b). Practical contextual bandits with feedback graphs. *Advances in Neural Information Processing Systems*, 36.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes], we present a description of the setting, as well as a description of the assumptions (see also Section 2.2 and Assumption 1).
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes], we provide a sample complexity analysis of the algorithm in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes], in the appended `README.md` file we included a description of all the needed dependecies as well as the anonymized source code.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes], we include in our results the full set of assumptions.
   (b) Complete proofs of all theoretical results. [Yes], we provide in the appendix a complete proofs of all the theoretical results presented.
   (c) Clear explanations of any assumptions. [Yes], we provide in Section 2.2 an explanation of all the assumptions used.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes], in the appended `README.md` file we describe how to reproduce the main experimental results.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes], in Appendix A we describe the training details.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes], we provide a measure of variability in our results, and provide its definition .
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes], in Appendix A we provide a description of the computing infrastructure used.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes], for all libraries that we used we cited those libraries appropriately (see also Appendix A).
   (b) The license information of the assets, if applicable. [Yes], we provide license information of the assets used, as well as the license information of our code in the appended material.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes], we provide new code to solve the optimization problems, and run the numerical results presented in this manuscript.
   (d) Information about consent from data providers/curators. [Not Applicable], we use assets for which consent is not needed.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable], we not use sensible content.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable], we did not use crowdsourcing nor conducted research with human subjects.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable], we did not use crowdsourcing nor conducted research with human subjects.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable], we did not use crowdsourcing nor conducted research with human subjects.

Alessio Russo[1], Yichen Song[1], Aldo Pacchiano[1,2]

Pure Exploration with Feedback Graphs (Supplementary Material)

Contents

A Additional Numerical Results

In this section of the appendix we describe the graphs used in the numerical results; the details of the algorithms used and exhibits additional numerical results.

A.1 Graphs Details

Here we briefly describe the graphs used in the numerical results. Also refer to the code for more details.

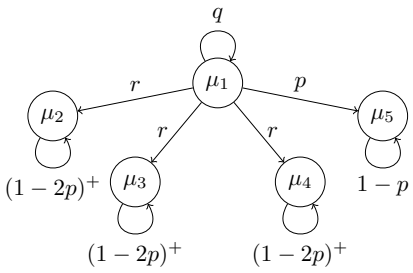

Figure 6: Loopy star graph.

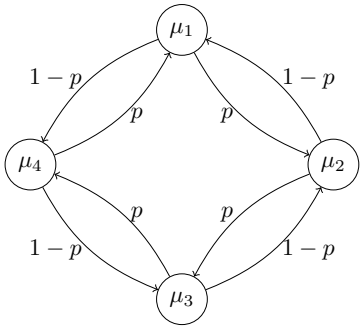

Figure 7: Ring graph.

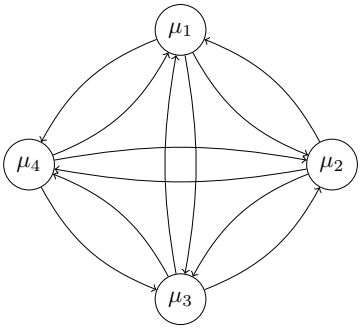

Figure 8: Loopless clique graph.

A.1.1 Loopy star graph

The loopy star graph (Fig. 6) is the only graph with self-loops in our experiments. However, it depends on several parameter (p, q, r) that affect the underlying topology. The rewards are Gaussianly distributed, with variance 1. The best arm has average reward 1, while sub-optimal arms have average reward 0.5.

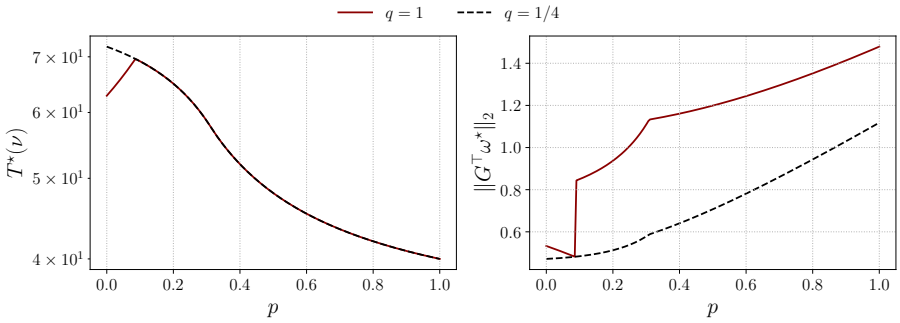

Figure 9: On the left: characteristic time of the loopy star graph with $K = 5, r = 0.5$ for different values of p, q (the best arm is $\nu = 5$). On the right: plot of $\|m^\star\|$ as a function of $\omega^\star$.

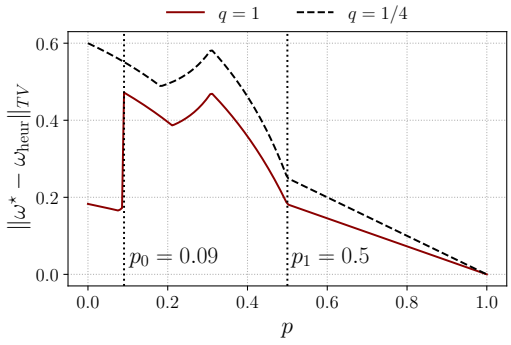

Figure 10: Difference between $\omega^\star$ and ω_{heur} in the loopy star graph (same setting as in Fig. 9).

We simulated two settings:

1. Main setting: we set $p = 1/5, q = 1/4, r = 1/5$. Hence, self-loops may bring more information. The best arm is $v = 5$. In Fig. 9 and Fig. 10 we depict $T^\star(\nu)$, $\|G^\top \omega^\star\|$ and the difference $\|\omega^\star - \omega_{\text{heur}}\|$ for this setting. Notably, for increasing values of p the approximate solution ω_{heur} converges to the optimal solution.

2. Alternative setting: we set $p = 0, q = 1/4, r = \frac{1-2q}{4(K-1)}$. Hence, it is not worth for the agent to choose $v = 1$, and the optimal arm is $v = 1$. In Fig. 11 and Fig. 12 we depict $T^\star(\nu)$, $\|G^\top \omega^\star\|$ and the difference $\|\omega^\star - \omega_{\text{heur}}\|$ for this setting. Also in this case for increasing values of p the approximate solution ω_{heur} converges to the optimal solution.

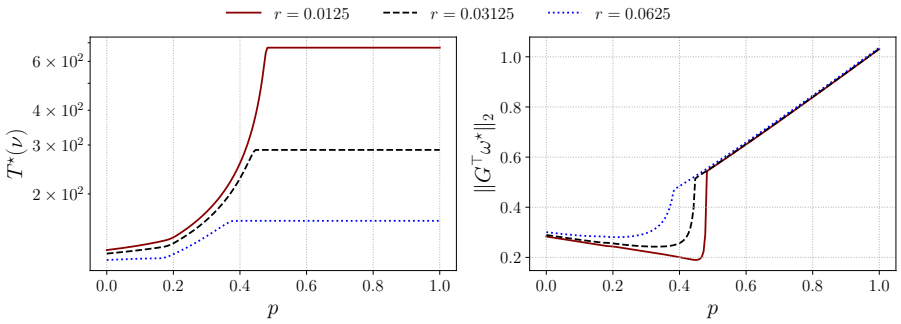

Figure 11: On the left: characteristic time of the loopy star graph in the alternative setting with $K = 5$ and $r = \alpha \frac{1-2q}{4(K-1)}, \alpha \in \{0.1, 0.25, 0.5\}$ (the best arm is $v = 1$). On the right: plot of $\|m^\star\|$ as a function of $\omega^\star$.

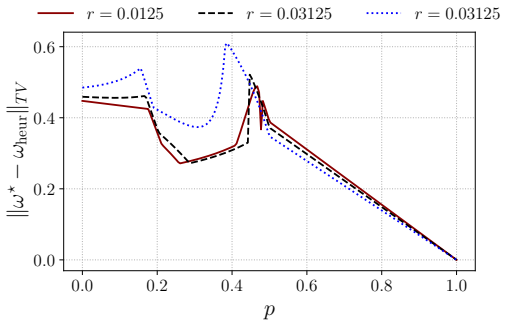

Figure 12: Difference between $\omega^\star$ and ω_{heur} in the loopy star graph (same setting as in Fig. 11).

A.1.2 Ring graph

In this graph (Fig. 7) each node is connected to two adjacent nodes. For each node u, the feedback from the node on the right v_r (clockwise direction) is "seen" with probability p, i.e., $G_{u,v_r} = p$; on the other hand, the feedback from the node on the "left" v_l (anti-clockwise direction) is "seen" with probability $1 - p$, i.e., $G_{u,v_l} = 1 - p$. All other edges have 0 probability. We used a value of $p = 0.3$ and rewards are Gaussian (with variance 1), with mean values linearly distributed in $[0, 1]$ across the K arms.

In Fig. 13 and Fig. 14 we depict $T^\star(\nu)$, $\|G^\top \omega^\star\|$ and the difference $\|\omega^\star - \omega_{\text{heur}}\|$ for this setting. In this case the approximate solution ω_{heur} converges to the optimal solution for $p \to 0$ or $p \to 1$.

A.1.3 Loopless clique graph

This graph (Fig. 8) is fully connected without any self-loops (the probabilities in the figure are omitted to avoid cluttering). Assuming the vertices $u \in V$ are numbered according to the natural numbers $V = \{1, 2, \ldots, K\}$, the

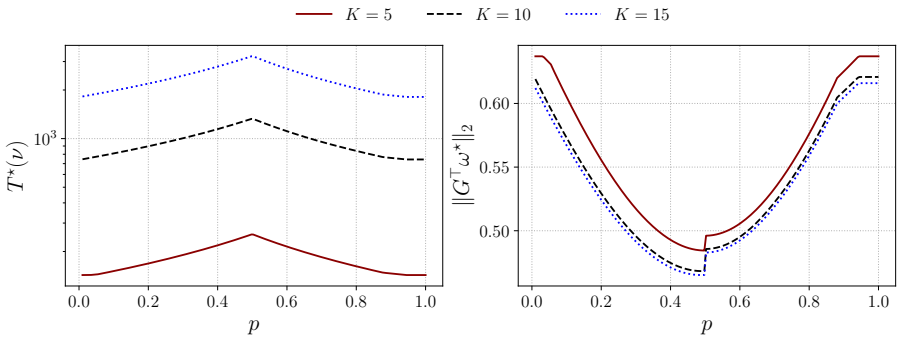

Figure 13: On the left: characteristic time of the ring graph for different values of K, p. On the right: plot of $\|m^\star\|$ as a function of $\omega^\star$.

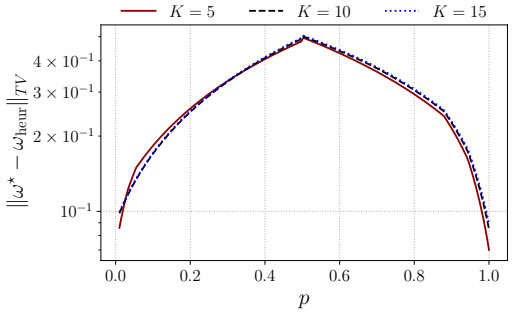

Figure 14: Difference between $\omega^\star$ and ω_{heur} in the ring graph (same setting as in Fig. 13).

edge probabilities are

$$G_{u,v} = \begin{cases} 0 & v = u, \\ p/u & \forall v \neq u \wedge v \notin 2\mathbb{N}, \\ 1 - (p/u) & \text{otherwise.} \end{cases}$$

where $2\mathbb{N}$ is the set of even numbers. We see that in this type of graph it may be better to choose a vertex according to its index depending on what type of feedback the learner seeks. We used a value of $p = 0.5$ and rewards are Gaussian (with variance 1), with mean values linearly distributed in $[0, 1]$ across the K arms.

In Fig. 15 and Fig. 16 we depict $T^\star(\nu)$, $\|G^\top \omega^\star\|$ and the difference $\|\omega^\star - \omega_{\text{heur}}\|$ for this setting. In this case the approximate solution ω_{heur} does not seem to converge to $\omega^\star$ for any value of p.

A.2 Numerical Results

In this section we present additional details on the numerical results.

A.2.1 Algorithms and results

In addition to `TaS-FG` and heuristic `TaS-FG` we also used EXP3.G (Alon et al., 2015), an algorithm for regret minimization in the adversarial case. We also compare with two variant of UCB: (1) UCB-FG-E, which acts greedily with respect to the upper confidence bound of $\left(\hat{G}(t)\hat{\mu}(t)\right)$; (2) UCB-FG-V, which selects $\arg\max_v \hat{G}^{\text{ucb}}_{v,\hat{a}^{\text{ucb}}(t)}(t)$, where $\hat{G}^{\text{ucb}}(t), \hat{a}^{\text{ucb}}(t)$ are, respectively, the UCB estimates of G and $a^\star$ at time t. For all algorithms, the graph estimator $\hat{G}(t)$ was initialized in an optimistic way, i.e., $\hat{G}_{u,v}(1) = 1$ for all $u, v \in V$.

EXP3.G algorithm. EXP3.G Alon et al. (2015) initializes two vectors $p, q \in \mathbb{R}^K$ uniformly, so that $p_i = q_i = 1/K$ for $i = 1, \dots, K$. At every time-step, an action V_t is drawn from $V_t \sim p_t$, where $p_t \leftarrow (1 - \eta)q_t + \eta\mathcal{U}$ with $\eta \in (0, 1)$ being an exploration facotr and $\mathcal{U}$ is the uniform distribution over $\{1, \dots, K\}$.

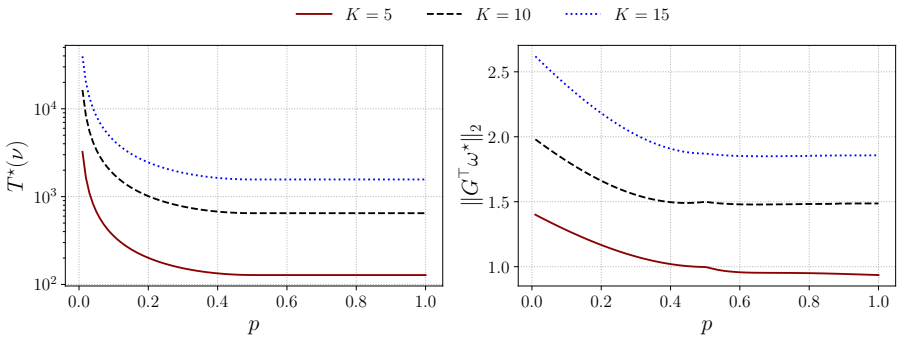

Figure 15: On the left: characteristic time of the loopless graph for different values of K, p. On the right: plot of $\|m^\star\|$ as a function of $\omega^\star$.

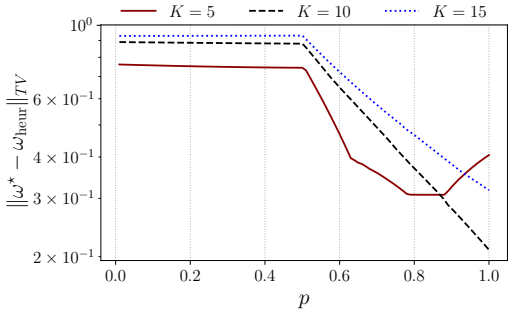

Figure 16: Difference between $\omega^\star$ and ω_{heur} in the loopless graph (same setting as in Fig. 15).

After observing the feedback, the algorithm sets

$$\hat{q}_t \leftarrow q_t \exp(-\eta x_t) \text{ and } q_t \leftarrow \hat{q}_t / \sum_u \hat{q}_{t,u},$$

where $x_{t,u} = -Z_{t,u} / \sum_{v \in N_{in}(u)} p_{t,v}$. In the experiments, we let $\eta = 3/10$.

UCB-FG-E algorithm. This method is a variant of UCB that acts greedily with respect to the upper confidence bound of $\left(\hat{G}(t)\hat{\mu}(t)\right)$. In practice, we let $\hat{\mu}_u^{\text{ucb}}(t) = \hat{\mu}_u(t) + \sqrt{\frac{2\ln(1+t)}{M_u(t)}}$, $\hat{G}_{u,v}^{\text{ucb}}(t) = \hat{G}_{u,v}(t) + \sqrt{\frac{\ln(1+t)}{2N_u(t)}}$ and select the action to take according to $V_t = \arg\max_u \hat{G}^{\text{ucb}}(t)\hat{\mu}^{\text{ucb}}(t)$.

UCB-FG-V algorithm. This method is a variant of UCB that selects $V_t = \arg\max_v \hat{G}^{\text{ucb}}_{v,\hat{a}^{\text{ucb}}(t)}(t)$, where $\hat{G}^{\text{ucb}}(t), \hat{a}^{\text{ucb}}(t)$ are, respectively, the UCB estimates of G and $a^\star$ at time t (note that $a^{\text{ucb}}(t) = \arg\max_u \hat{\mu}_u^{\text{ucb}}(t)$).

In Fig. 17 we depict the sample complexity of the algorithms for different values of K, δ. Note that the sample complexity τ is not normalized, and results were computed over 100 seeds.

A.2.2 Libraries and computational resources

Libraries used in the experiments. We set up our experiments using Python 3.10.12 (Van Rossum and Drake Jr, 1995) (For more information, please refer to the following link `http://www.python.org`), and made use of the following libraries: NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), Seaborn (Waskom et al., 2017), Pandas (McKinney et al., 2010), Matplotlib (Hunter, 2007), CVXPY (Diamond and Boyd, 2016). As numerical optimizer we used Gurobi 10.0.1 (Gurobi Optimization, LLC, 2024).

New code is published under the MIT license. To run the code, please, read the attached `README.md` file for instructions.

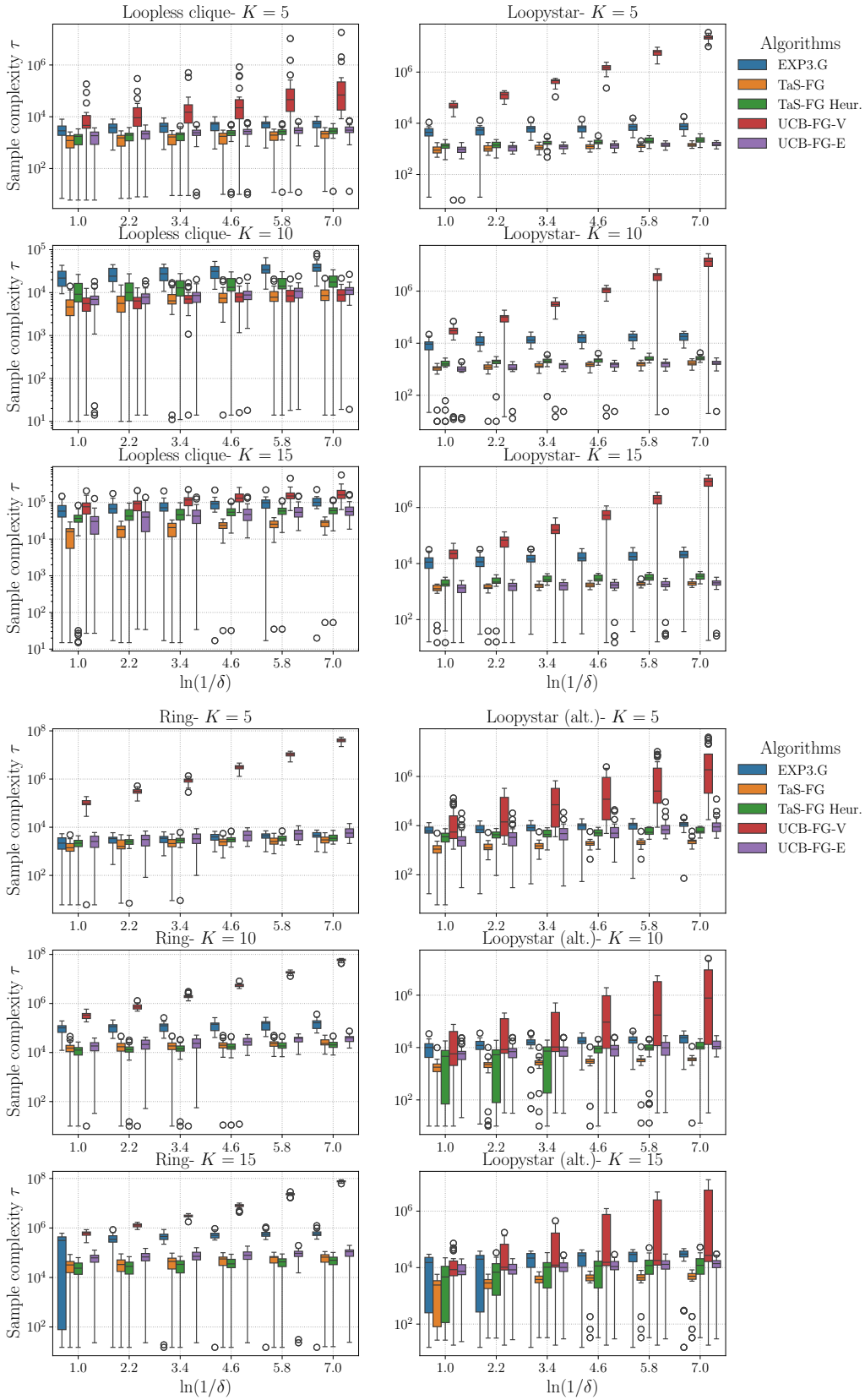

Figure 17: Sample complexity results, from top-left (clockwise): loopless clique, loopystar, ring and alternative loopystar graphs. Box plots are not normalized, and were computed over 100 seeds. Boxes indicate the interquartile range, while the median and mean values are, respectively, the solid line and the + sign in black.

Computational resources. Experiments were run on a shared cluster node featuring a Linux OS with 2 fourteen-core 2.6 GHz Intel Gold 6132 and 384GB of ram. The total computation time per core to design the experiments, debug the code and obtain the final results was roughly of 82 hours/core. To obtain the final results over 100 seeds we estimated that 10/hours/core are sufficient.

B Feedback Graph Properties

In the following we list some properties for feedback graphs.

Lemma 1. *For a strongly observable graph G we have that $|SO(G)| \geq \max(\sigma(G), \alpha(G))$.*

Proof. First, note that $\sigma(G) \leq |SO(G)|$, since all vertices with a self-loop are strongly observable.

To prove that $|SO(G)| \geq \alpha(G)$, by contradiction, assume that $\alpha(G) > |SO(G)|$. First, for any $I \in \mathcal{I}, u \in I$, u is strongly observable. Since $|I| = \alpha(G)$, then at-least $|SO(G)| \geq \alpha(G)$, which is a contradiction. □

Lemma 2. *Assume $|\mathcal{I}(G)| > 1$ and let $I_1, I_2 \in \mathcal{I}(G)$, with $I_1 \neq I_2$. Then, for any $u \in I_1$ we have that there exists $v \in I_2 \setminus I_1$ satisfying $G_{uv} > 0$ or $G_{vu} > 0$.*

Proof. Let $I_1, I_2 \in \mathcal{I}(G)$ satisfying $I_1 \neq I_2 \Rightarrow \exists v \in I_2 \setminus I_1$. Let $u \in I_1$. By contradiction, for all $v \in I_2 \setminus I_1$ we have that $G_{u,v} = 0$ and $G_{v,u} = 0$. In that case, we can construct a new a new set $\tilde{I} = I_1 \cup \{v\}$ such that $G_{u,v} = 0$ and $G_{v,u} = 0$ for any $u, v \in \tilde{I}$, implying that $\tilde{I} \in \mathcal{I}(G)$. But $|I_1| < |\tilde{I}|$, which contradicts the fact that $I_1 \in \mathcal{I}(G)$. □

Lemma 3. *For a strongly observable graph if $\alpha(G) > 1$ then $\forall I \in \mathcal{I}(G), \forall u \in I$ we have $\{u\} \in N_{in}(u)$, that is, all vertices in I have self-loops. As a corollary, we have that $\sigma(G) \geq \alpha(G)$ if $\alpha(G) > 1$.*

Proof. By contradiction, assume there exists $I_0 \in \mathcal{I}(G)$ with $u \in I_0$ such that $\{u\} \notin N_{in}(u)$. Since u is strongly observable, it means that $V \setminus \{u\} \in N_{in}(u)$.

Now, consider the case $\alpha(G) > 1$. If that is the case, then let $v \in I_0$. By strong observability of u, we have $\{v\} \in N_{in}(u)$, which contradicts the fact that I_0 is an independent set.

The latter statement is a consequence of the fact that all vertices in every I have self-loops. □

Corollary 2. *Consider a strongly observable graph with $\alpha(G) > 1$. Then, $\forall I \in \mathcal{I}$ there exists $I_0 \in \mathcal{I}$ such that $I \ll I_0$.*

Proof. By contradiction assume that $\exists I \in \mathcal{I}(G)$ such that $\forall I_0 \in \mathcal{I}(G)$ there exists $u(I_0) \in I$ such that $N_{in}(u(I_0)) \cap I_0 = \emptyset$, where $u : \mathcal{I}(G) \to V$. However, since the graph is strongly observable, we have that either (1) $\{u(I_0)\} \in N_{in}(u(I_0))$, or (2) $V \setminus \{u(I_0)\} \subset N_{in}(u(I_0))$ or (3) both.

Consider the case $\alpha(G) > 1$. By Lemma 3 then each vertex in I has a self-loop. Hence taking $I_0 = I$ leads to $N_{in}(u(I)) \cap I \ni \{u(I)\}$, which is a contradiction. □

Lemma 4. *Consider a strongly observable graph. Then $|SO(G)| = \sigma(G) = \alpha(G)$ only for bandit feedback graphs. As a corollary, for non-bandit feedback graphs we have $|SO(G)| > \alpha(G)$.*

Proof. The first part of the lemma is easy to prove as $\alpha(G)$ is maximal when all the vertices have only self-loops, thus $\alpha(G) = |SO(G)| = K$.

To prove the second part, note that it always holds true that $K = |SO(G)| \geq \alpha(G)$ by Lemma 1. However, equality is reached only for bandit feedback graphs. Therefore, for other feedback graphs it holds that $|SO(G)| > \alpha(G)$. □

Lemma 5. *Consider a graph (V, G), with $\alpha(G) < \sigma(G)$. For any subset $W \subset L(G)$ of size $|W| = \alpha(G) + 1$, at most $\alpha(G)$ vertices are needed to dominate W.*

Proof. The proof is simple: first, by Lemma 6, we have $\alpha(W) \leq \alpha(G)$. Therefore W is not an independent set, and there must exists $v, u \in W$ such that $u \in N_{out}(v)$. Therefore $W \setminus \{u\}$ dominates W. □

Lemma 6. *For any set G satisfying $\alpha(G) = k$, we must have $\max(1, k + |V| - |G|) \leq \alpha(V) \leq k$ for any subset V of G.*

Proof. The right hand-side is trivial since any subset $V \subset G$ can have at most k independent vertices. The left hand-side follows from the fact that removing an element from G can at most reduce the number of independent vertices by 1. □

B.1 Domination Number of the Set of Strongly Observable Vertices

We now provide one of the main results that shows an upper bound on the number of vertices with self-loops needed to dominate the set of strongly observable vertices.

In the proofs we denote by $(V|_A, G|_A)$ the restriction of a graph (V, G) to a set of vertices $A \subseteq V$. Practically, we have that $V|_A = A$, and $G|_A \in \mathbb{R}^{|A|\times|A|}$ with $(G|_A)_{u,v} = G_{u,v}$ for $u, v \in A$

We begin with the following preliminary lemma that studies the case $\alpha(G) = 1$.

Lemma 7. *Let $E(G) = SO(G) \setminus L(G)$ be the set of strongly observable vertices that do not have a self-loop. Assume that $\alpha(G) = 1$ and $|SO(G)| > 0$. Then, at most $\sigma(G) - \left\lfloor \frac{\sigma(G)}{2} \right\rfloor$ vertices in $L(G)$ are needed to dominate the set of strongly observable vertices $SO(G)$.*

Proof. The idea of the proof is to find the least number of vertices in $L(G)$ that dominates $SO(G)$, and we prove this by induction.

The reason why we can do that, is that since any $v \in E(G)$ lacks a self-loop, strong observability forces v to have edges from every other vertex. Hence, any vertex in $L(G)$ has an out-edge to $v \in E(G)$, and thus dominates $E(G)$. Therefore we only need to find the domination number of $L(G)$.

We begin by considering the case $\alpha(G) = 1$, with $\sigma(G) = 1, \sigma(G) = 2, \sigma(G) = 3$ and then the general case $\sigma(G) > 1$.

First, note that $|SO(G)| \geq |L(G)| = \sigma(G)$. If $v \in E(G)$, then by definition $v \notin L(G)$, and so v must have in-edges from every other vertex, that is $N_{in}(v) = V \setminus \{v\}$.

Case $\sigma(G) = 1$. There is exactly one vertex v with a self-loop. This single vertex trivially dominates itself and also dominates any vertex in $E(G)$. Hence the domination number of $SO(G)$ is 1.

Case $\sigma(G) = 2$. The two self-loop vertices in $L(G) = \{u, v\}$ must have a directed edge between them in some direction (or both) since $\alpha(G) = 1$. Therefore, it must either be $G_{u,v} > 0$ or $G_{v,u} > 0$ for $L(G) = \{u, v\}$. Hence, there exists one vertex that dominates $SO(G)$.

Case $\sigma(G) = 3$. We can partition $L(G)$ into two sets $L_1(G), L_2(G)$, each of size 2, such that $L_1(G) \cup L_2(G) = L(G)$ and $|L_1(G)| = |L_2(G)|$. Since $\alpha(L_1(G)) = \alpha(L_2(G)) = 1$, we can repeatedly apply the logic from the case $\sigma(G) = 2$ to obtain that at most 2 vertices are required to dominate $SO(G)$.

Case $\sigma(G) \geq 4$. For any $\sigma(G)$, form pairs of self-loop vertices. If $\sigma(G)$ is even, we get $\sigma(G)/2$ disjoint pairs; each pair needs just one dominator (since $\alpha(G) = 1$ forces an edge in each pair). If $\sigma(G)$ is odd, then we can construct $(\sigma(G) - 1)/2$ pairs plus a singleton, which requires one more dominator. Hence $(\sigma(G) - 1)/2 + 1 = \sigma(G) - \frac{\sigma(G)-1}{2} = \sigma(G) - \lfloor \sigma(G)/2 \rfloor$ suffice.

□

Theorem 4. *Let $E(G) = SO(G) \setminus L(G)$ be the set of strongly observable vertices that do not have a self-loop. Assume that $|SO(G)| > 0$. Then:*

- *(Case 1) If $SO(G) \setminus E(G) = \emptyset$, then $\min(|SO(G)|, 2)$ vertices are needed to dominate the set of strongly observable vertices $SO(G)$.*
- *(Case 2) If $SO(G) \setminus E(G) \neq \emptyset$, then at most $\sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor$ vertices in $L(G)$ are needed to dominate the set of strongly observable vertices $SO(G)$.*

Proof. Recall that $L(G) = \{v \in V : \{v\} \in N_{in}(v)\}$ is the set of vertices that do have a self-loop. It follows that $L(G) \cup E(G) = SO(G)$. Moreover, if $v \in E(G)$, then by definition $v \notin L(G)$, and so v must have in-edges from every other vertex, that is $N_{in}(v) = V \setminus \{v\}$.

Case 1. In the first case, since $E(G) = SO(G)$, if $|SO(G)| = 1$ we only need one vertex to dominate $SO(G)$. Instead, if $|SO(G)| > 1$, it is possible to dominate $SO(G)$ with two vertices (just pick two vertices from $E(G)$).

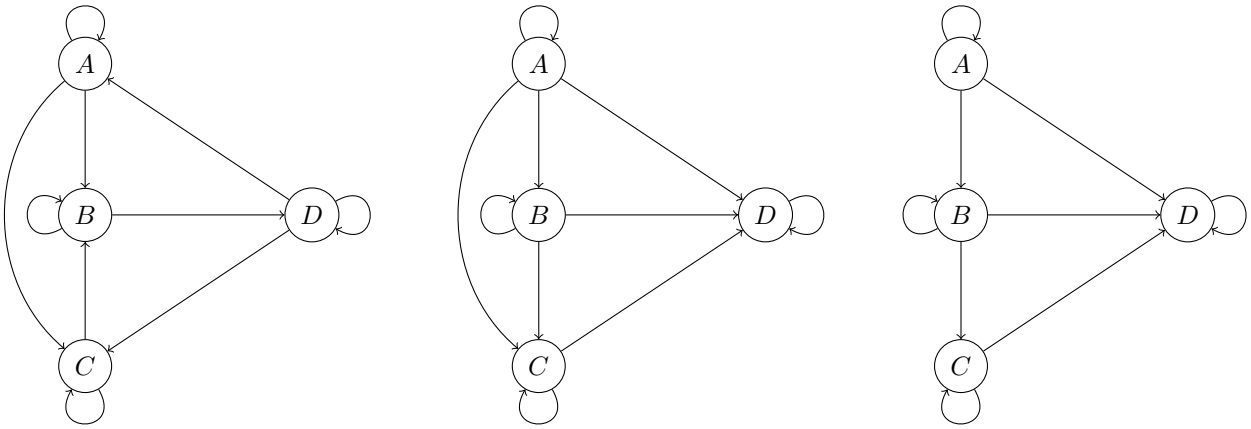

Figure 18: Example of strongly observable graphs and their domination number. **On the left**: a graph with $\sigma(G) = 4$ and $\alpha(G) = 1$. The smallest sets of vertices that dominate this graph are $\{A, B\}, \{B, D\}$. The maximally independent sets are $\mathcal{I} = \{\{A\}, \{B\}, \{C\}, \{D\}\}$. Note that $\sigma(G) - \lfloor \sigma(G)/(\alpha(G)+1) \rfloor = 4 - \lfloor 4/2 \rfloor = 2$. **In the middle**: a graph with $\alpha(G) = 1, \sigma(G) = 4$. The smallest set of vertices that dominate the graph is $\{A\}$. The maximally independent sets are $\mathcal{I} = \{\{A\}, \{B\}, \{C\}, \{D\}\}$. We have $\sigma(G) - \lfloor \sigma(G)/(\alpha(G)+1) \rfloor = 2$. **On the right**: a graph with $\alpha(G) = 2, \sigma(G) = 4$. The smallest sets of vertices that dominate the graph are $\{A, B\}, \{A, C\}$. The maximally independent sets are $\mathcal{I} = \{\{A, C\}\}$. We have $\sigma(G) - \lfloor \sigma(G)/(\alpha(G)+1) \rfloor = 4 - \lfloor 4/3 \rceil = 3$.

Case 2. For the second case, since any $v \in E(G)$ lacks a self-loop, strong observability forces v to have edges from every other vertex. Hence, any vertex in $L(G)$ has an out-edge to $v \in E(G)$, and thus dominates $E(G)$. Therefore we only need to find the domination number of $L(G)$.

1. The case for $\alpha(G) = 1$ was proved in Lemma 7.

2. Now, assume $1 < \alpha(G)$ and $\alpha(G) \geq \sigma(G)$. Note that we can always dominate $L(G)$ with $\sigma(G)$ elements, and since $\sigma(G)/(\alpha(G)+1) < 1$, we have that the statement of case 1 holds.

3. In the last case we have $1 < \alpha(G) < \sigma(G)$.

   Because $\alpha(G) < \sigma(G)$, then $L(G)$ is not an independent set–there must be edges among its vertices. Hence, by Lemma 5, in any subset of $\alpha(G)+1$ vertices from $L(G)$, it is possible to dominate that subset with at most $\alpha(G)$ vertices. Therefore, consider the two cases below:

   - If $\sigma(G)$ is divisible by $\alpha(G)+1$, then divide $L(G)$ into $m = \lfloor \sigma(G)/(\alpha(G)+1) \rfloor$ sets $L_i(G)$, each of size $\alpha(G)+1$, so that $\cup_i L_i(G) = L(G)$.
     In each set $L_i(G)$ at most $\alpha(G)$ vertices are needed to dominate it.
   - Alternatively, if $\sigma(G)$ is *not* divisible by $\alpha(G)+1$, then there is a leftover block of size $r < \alpha(G)+1$, with
     $$r + m(\alpha(G)+1) = \sigma(G).$$
     Note that at most r vertices are needed to dominate this set.

   Hence, in either case, the dominating-set size of $L(G)$ is at most $m\alpha(G) + r$, which is

   $$m\alpha(G) + r = m\alpha(G) + \sigma(G) - m(\alpha(G)+1) = \sigma(G) - m = \sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor.$$

□

C Sample Complexity Lower Bounds

The sample complexity analysis delves on the required minimum amount of *evidence* needed to discern between different hypotheses (e.g., vertex v is optimal vs vertex v is not optimal). The evidence is quantified by the log-likelihood ratio of the observations under the true model and a *confusing model*. This confusing model, is usually, the model that is *statistically* closer to the true model, while admitting a different optimal vertex.

To state the lower bounds, we first define the concept of *absolute continuity* between two models. For any two models $\nu = \{G, (\nu_u)_{u\in V}\}, \nu' = \{G', (\nu'_u)_{u\in V}\}$ with the same number of vertices we say that ν is absolutely continuous w.r.t ν', that is $\nu \ll \nu'$, if for all $(v,u) \in V^2$ we have $\nu_{v,u} \ll \nu'_{v,u}$

Given this definition of absolute continuity, we can define the set of confusing models as follows

$$\mathrm{Alt}(\nu) \coloneqq \{\nu' : a^\star(\nu) \neq a^\star(\nu'), \nu \ll \nu'\},$$

which is the set of models for which $a^\star(\nu)$ is not optimal. We also denote by $\mathrm{Alt}_u(\nu) = \{\nu' : \mu'_u > \mu'_{a^\star}\}$ the set of models where $u \neq a^\star(\nu)$ may be the optimal vertex in ν'.

C.1 The Uninformed Setting

In this section we prove Theorem 1 and Proposition 4. We start by stating a general expression of the lower bound.

C.1.1 General lower bound expression

In Theorem 5 we state a general expression for $T^\star(\nu)$ and then show in the next sections the proofs of Theorem 1 and Proposition 4.

Theorem 5. *For any δ-PC algorithm and any model ν satisfying Assumption 1, we have that*

$$\mathbb{E}_\nu[\tau] \geq T^\star(\nu)\mathrm{kl}(\delta, 1-\delta), \tag{11}$$

where $(T^\star(\nu))^{-1}$ is equivalent to the following two expressions

$$(T^\star(\nu))^{-1} = \begin{cases} \sup_{\omega\in\Delta(V)} \inf_{\nu'\in\mathrm{Alt}(\nu)} \sum_{v\in V} \omega_v \sum_{u\in N_{out}(v)} \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}), \\ \sup_{\omega\in\Delta(V)} \inf_{\nu'\in\mathrm{Alt}(\nu)} \sum_{u\in V} \sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}). \end{cases} \tag{12}$$

We refer to the first expression as "pull-based", partitioning the log-likelihood by which vertex was chosen. We refer to the second expression as "observation-based", partitioning by which vertex was observed.

Proof. Consider two bandit models $\nu = \{G, (\nu_u)_u\}, \nu' = \{G', (\nu'_u)_u\}$ with the same number of vertices and unique optimal vertex in both models, such that $\nu \ll \nu'$. For each v there exists a measure λ_v such that ν_v and ν'_v have, respectively, densities f_v and f'_v. Similarly, for $\nu_{v,u}$ and $\nu'_{v,u}$ there exists a measure $f_{v,u}$ and $f'_{v,u}$ respectively.

First expression (pull-based). Hence, consider the log-likelihood ratio between ν and ν' of the data observed up to time t, and consider writing it in terms of the out-neighborhood of the vertices selected by the algorithm:

$$\begin{aligned} L_t &= \ln \frac{\mathrm{d}\mathbb{P}_\nu(V_1, Z_1, \ldots, V_t, Z_t)}{\mathrm{d}\mathbb{P}_{\nu'}(V_1, Z_1, \ldots, V_t, Z_t)}, \\ &= \sum_{n=1}^{t} \sum_{u\in N_{out}(V_n)} \ln\left(\frac{f_{v,u}(Z_{n,u})}{f'_{v,u}(Z_{n,u})}\right), \\ &= \sum_{n=1}^{t} \sum_{v\in V} \sum_{u\in N_{out}(v)} \mathbf{1}_{\{V_n=v\}} \ln\left(\frac{f_{v,u}(Z_{n,u})}{f'_{v,u}(Z_{n,u})}\right), \\ &= \sum_{v\in V} \sum_{u\in N_{out}(v)} \sum_{s=1}^{N_v(t)} \ln\left(\frac{f_{v,u}(W_{s,(v,u)})}{f'_{v,u}(W_{s,(v,u)})}\right), \end{aligned}$$

where $(W_{s,(v,u)})_s$ is an i.i.d. sequence of samples observed from $\nu_{v,u}$. Hence, if we take the expectation with respect to ν, by Wald's lemma, we have that

$$\begin{aligned}\mathbb{E}_\nu[L_t] &= \sum_{v\in V}\sum_{u\in N_{out}(v)} \mathbb{E}_\nu[N_v(t)]\mathrm{KL}(\nu_{v,u},\nu'_{v,u}),\\ &= \sum_{v\in V}\mathbb{E}_\nu[N_v(t)]\sum_{u\in N_{out}(v)}\mathrm{KL}(\nu_{v,u},\nu'_{v,u}).\end{aligned}$$

Therefore, applying (Kaufmann et al., 2016, Lemma 1) at $t=\tau$, we find that for any δ-PC algorithm we have that

$$\sum_{v\in V}\mathbb{E}_\nu[N_v(\tau)]\sum_{u\in N_{out}(v)}\mathrm{KL}(\nu_{v,u},\nu'_{v,u}) \geq \mathrm{kl}(\delta,1-\delta).$$

Consider the set of confusing models $\mathrm{Alt}(\nu)=\{\nu'=(\mu',G') : a^\star(\mu)\neq a^\star(\mu'), \nu\ll\nu'\}$, and define the selection rate of a vertex v as $\omega_v = \mathbb{E}_\nu[N_v(\tau)]/\mathbb{E}_\nu[\tau]$. Then, by minimizing over the set of confusing models, and then optimizing $\omega=(\omega_v)_{v\in V}$ over the simplex $\Delta(V)$, we obtain

$$\mathbb{E}_\nu[\tau]\underbrace{\sup_{\omega\in\Delta(V)}\inf_{\nu'\in\mathrm{Alt}(\nu)}\sum_{v\in V}\omega_v\sum_{u\in N_{out}(v)}\mathrm{KL}(\nu_{v,u},\nu'_{v,u})}_{=:(T^\star(\nu))^{-1}} \geq \mathrm{kl}(\delta,1-\delta).$$

and therefore $\mathbb{E}_\nu[\tau]\geq T^\star(\nu)\mathrm{kl}(\delta,1-\delta)$.

Second expression (obsrvation-based). The second version of $T^\star(\nu)$ comes from considering the in-neighborhood of v for each vertex:

$$\begin{aligned}L_t &= \sum_{n=1}^t\sum_{u\in V}\sum_{v\in N_{in}(u)}\mathbf{1}_{\{V_n=v\}}\ln\left(\frac{f_{v,u}(Z_{n,u})}{f'_{v,u}(Z_{n,u})}\right),\\ &= \sum_{u\in V}\sum_{v\in N_{in}(u)}\sum_{s=1}^{N_v(t)}\ln\left(\frac{f_{v,u}(W_{s,(v,u)})}{f'_{v,u}(W_{s,(v,u)})}\right)\end{aligned}$$

Hence

$$\mathbb{E}_\nu[L_t] = \sum_{u\in V}\sum_{v\in N_{in}(u)}\mathbb{E}_\nu[N_v(t)]\mathrm{KL}(\nu_{v,u},\nu'_{v,u}),$$

from which we can immediately conclude the proof by following the same steps as for the previous expression. □

C.1.2 Continuous vs discrete rewards

Before proceeding further in our analysis, we rewrite the KL-divergence in terms of the associated Radon-Nykodim derivatives. Note that for a product random variable $Z=XY$ with $Z\sim\nu_{X,Y}$, we have that $\mathbb{P}_Z(A)=\int_A f_Z(z)\mathrm{d}\mu(z)$ with respect to some dominating measure $\mu(z)$. We consider some cases:

- **Continuous case:** for Y distributed as a Bernoulli of parameter p, and X as a continuous r.v. with density $f_X(x)$ we have that $\mathbb{P}_Z(A)=(1-p)\mathbf{1}_{\{0\in A\}}+p\int_A f_X(z)\mathrm{d}\lambda(z)$ where λ is the Lebesgue measure Let $\mu(A)=\delta_0(A)+\lambda(A)$ be the dominating measure. To find $f_Z(z)$ we can apply the Radon-Nykodim derivative in $z=0$, which tells us that
$$\mathbb{P}_Z(0)=1-p=\int_{\{0\}}f_Z(z)\mathrm{d}\mu(z)=f_Z(0).$$
On the other hand for $z\neq 0$ we have
$$\mathbb{P}_Z(A)=p\int_A f_X(z)\mathrm{d}\lambda(z)=\int_A f_Z(z)\mathrm{d}\mu(z)=\int_A f_Z(z)\mathrm{d}\lambda(z)\Rightarrow f_Z(z)=pf_X(z)\text{ a.e.}$$

Therefore

$$f_Z(z) = (1-p)\mathbf{1}_{\{z=0\}} + pf_X(z)\mathbf{1}_{\{z\neq 0\}}.$$

In words, the "continuous" part has no contribution to the overall probability mass when $z = 0$, since the Lebesgue measure of $\{X = 0\}$ is 0, while for $z \neq 0$ the main contribution comes from the continuous part. In the setting studied in this paper, the intuition is that when we observe 0, then almost surely we know its due to the edge not being activated.

- **Discrete case:** for Y distributed as a Bernoulli of parameter p, and X as a categorical r.v. over $\{0, \dots, N\}$ with probabilities $\{q_0, \dots, q_N\}$ we have that $\mu(A) = \sum_{i=0}^{N} \delta_i(A)$, and

$$f_Z(z) = (1-p)\mathbf{1}_{\{z=0\}} + p\left[\sum_{i=0}^{N} q_i \mathbf{1}_{\{z=i\}}\right],$$

hence $\mathbb{P}_Z(Z = 0) = 1 - p + pq_0$ and $\mathbb{P}(Z = i) = pq_i$ for $i \in \{1, \dots, N\}$.

C.1.3 The continuous case: proof of Theorem 1

We now consider the continuous case. From the second expression (observation-based) in Theorem 5 we derive the result of Theorem 1.

Proof of Theorem 1. We continue from the result of Theorem 5. Note that $\mathrm{Alt}(\nu) = \{\nu' = (G', \{\nu'_u\}_u) \mid \exists v_0 \neq a^\star : \mu'_{v_0} > \mu'_{a^\star}\}$, where $a^\star = a^\star(\mu)$. Hence, letting $\mathrm{Alt}_v(\nu) = \{\nu' \mid \mu'_v > \mu'_{a^\star}\}$, we have $\mathrm{Alt}(\nu) = \cup_{v\neq a^\star}\mathrm{Alt}_v(\nu)$. Therefore, due to the properties of the KL divergence we obtain

$$\begin{aligned}
&\inf_{\nu'\in\mathrm{Alt}(\nu)} \sum_{u\in V}\sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&\quad = \min_{u\neq a^\star} \inf_{\nu'\in\mathrm{Alt}_u(\nu):\mu'_v>\mu'_{a^\star}} \sum_{u\in V}\sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}).
\end{aligned}$$

Following the discussion in Appendix C.1.2, we can write

$$\begin{aligned}
\mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) &= \mathbb{E}_{Z\sim\nu_{v,u}}\left[\ln \frac{\mathrm{d}\mathbb{P}_{\nu_{v,u}}(Z)/\mathrm{d}\mu(Z)}{\mathrm{d}\mathbb{P}'_{\nu_{v,u}}(Z)/\mathrm{d}\mu(Z)}\right], \\
&= \mathbb{E}_{Z\sim\nu_{v,u}}\left[\ln \frac{f_{v,u}(Z)}{f'_{v,u}(Z)}\right], \\
&= (1-G_{v,u})\ln\frac{1-G_{v,u}}{1-G'_{v,u}} + G_{v,u}\mathbb{E}_{Z\sim\nu_u}\left[\ln\frac{G_{v,u}f_u(Z)}{G'_{v,u}f'_u(Z)}\right], \\
&= \mathrm{kl}(G_{v,u}, G'_{v,u}) + G_{v,u}\mathbb{E}_{Z\sim\nu_u}\left[\ln\frac{f_u(Z)}{f'_u(Z)}\right], \\
&= \mathrm{kl}(G_{v,u}, G'_{v,u}) + G_{v,u}\mathrm{KL}(\nu_u, \nu'_u).
\end{aligned}$$

Therefore, noting that the constraint involves only the pair $(\nu'_u, \nu'_{a^\star})$ through their parameters $(\mu'_u, \mu'_{a^\star})$, we conclude that

$$\begin{aligned}
&\inf_{\nu'\in\mathrm{Alt}(\nu)} \sum_{u\in V}\sum_{v\in N_{in}(u)} \omega_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) \\
&\quad = \min_{u\neq a^\star} \inf_{\nu'\in\mathrm{Alt}_u(\nu):\mu'_v>\mu'_{a^\star}} \sum_{u\in V}\sum_{v\in N_{in}(u)} \omega_v \left(\mathrm{kl}(G_{v,u}, G'_{v,u}) + G_{v,u}\mathrm{KL}(\nu_u, \nu'_u).\right), \\
&\quad = \min_{u\neq a^\star} \inf_{\nu':\mu'_u\geq\mu'_{a^\star}} \sum_{v\in N_{in}(u)} \omega_v G_{v,u}\mathrm{KL}(\nu_u, \nu'_u) + \sum_{w\in N_{in}(a^\star)} \omega_w G_{w,a^\star}\mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&\quad = \min_{u\neq a^\star} \inf_{\nu':\mu'_u\geq\mu'_{a^\star}} m_u\mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star}\mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}),
\end{aligned}$$

where $m_u \coloneqq \sum_{v\in N_{in}(u)} \omega_v G_{v,u}$ and $m_{a^\star} = \sum_{\omega\in N_{in}(a^\star)} \omega_w G_{w,a^\star}$.

Therefore, by optimizing over ν' as in (Garivier and Kaufmann, 2016, Lemma 3) we obtain

$$(T^\star(\nu))^{-1} = \sup_{\omega\in\Delta(V)} \min_{u\neq a^\star} (m_u + m_{a^\star}) I_{\frac{m_{a^\star}}{m_u+m_{a^\star}}}(\nu_{a^\star}, \nu_u) \text{ s.t. } m_u = \sum_{v\in N_{in}(u)} \omega_v G_{v,u}.$$

□

C.1.4 The discrete case: proof of Proposition 4

We find that if $(\nu_u)_{u\in V}$ are Bernoulli distributed, then we obtain that it is not possible, in general, to estimate the best vertex. The reason is simple: without knowing which edge was activated, the learner does not know how to discern between the randomness of the reward and the randomness of the edge.

Proof of Proposition 4. Let $n \in \mathbb{R}_+^K$ and consider the optimization problem

$$\inf_{\nu'\in \mathrm{Alt}(\nu)} \sum_{u\in V} \sum_{v\in N_{in}(u)} n_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}).$$

If $\nu_{v,u}$ is a Bernoulli distribution of parameter $q_{v,u} \coloneqq G_{v,u}\mu_u$ (sim. for $\nu'_{v,u}$ with $q'_{v,u} \coloneqq G'_{v,u}\mu'_u$), then we can write

$$\inf_{\nu'\in \mathrm{Alt}(\nu)} \sum_{u\in V} \sum_{v\in N_{in}(u)} n_v \mathrm{KL}(\nu_{v,u}, \nu'_{v,u}) = \inf_{\nu'\in \mathrm{Alt}(\nu)} \sum_{u\in V} \sum_{v\in N_{in}(u)} n_v \mathrm{kl}(q_{v,u}, q'_{v,u}).$$

Using the fact that $\nu' \in \mathrm{Alt}(\nu)$, and that ν' is observable, it must imply that there exists $(v,u) \in V^2$ such that $\mu'_u > \mu'_{a^\star}$ and $G'_{v,u} > 0$. Similarly, from the observability of ν $\exists w \in V$ s.t. $G_{w,a^\star} > 0$. Hence, by absolute continuity, we have $G'_{w,a^\star} > 0$, otherwise the event $\mathcal{E} = \{Z_{w,a^\star} = 1\}$ would satisfy $\mathbb{P}_\nu(\mathcal{E}) > 0$ and $\mathbb{P}_{\nu'}(\mathcal{E}) = 0$.

Then, similarly as in previous results, we have that

$$\begin{aligned}
&\inf_{\nu'\in \mathrm{Alt}(\nu)} \sum_{u\in V} \sum_{v\in N_{in}(u)} n_v \mathrm{kl}(q_{v,u}, q'_{v,u}), \\
&\quad = \min_{u\neq a^\star} \min_{v\in N_{in}(u), w\in N_{in}(a^\star)} \inf_{\mu', G': \mu'_u \geq \mu'_{a^\star}, G'_{v,u}, G'_{w,a^\star} \geq 0} n_v \mathrm{kl}(q_{v,u}, q'_{v,u}) + n_w \mathrm{kl}(q_{w,a^\star}, q'_{w,a^\star}).
\end{aligned}$$

Therefore, for $u \neq a^\star$, $v \in N_{in}(u), w \in N_{in}(a^\star)$, we are interested in the following non-convex problem

$$\begin{aligned}
\min_{\mu', G'>0} \quad & n_v \mathrm{kl}(q_{v,u}, q'_{v,u}) + n_w \mathrm{kl}(q_{w,a^\star}, q'_{w,a^\star}) \\
\text{s.t.} \quad & q'_{v,u} = G'_{v,u}\mu'_u \qquad \forall (v,u) \in V^2, \\
& \mu'_u \geq \mu'_{a^\star}.
\end{aligned}$$

However, the solution, as one may expect, is 0. Simply note one can take $q'_{v,u} = q_{v,u}$ by choosing $\mu'_u = \mu_{a^\star}$ and $G'_{v,u} = q_{v,u}/\mu_{a^\star}$ (which is ≤ 1). Similarly, we have $q'_{w,a^\star} = q_{w,a^\star}$ by choosing $\mu'_{a^\star} = \mu_{a^\star}$ and $G'_{w,a^\star} = q_{w,a^\star}/\mu_{a^\star} = G_{w,a^\star}$. Henceforth, in the Bernoulli case we have that $(T^\star(\nu))^{-1} = 0$. □

C.2 The Informed Setting

We now give a proof of Theorem 2, which follows closely the one in the uninformed case.

Proof of Theorem 2. First note that knowing the set of edges that was activated is equivalent to knowing the value of $(Y_{t,(V_t,u)})_u$. Then, denote the density associated to $Y_{t,(V_t,u)}$ by $g_{v,u}$ under ν (sim. $g'_{v,u}$ for ν'). We can

write the log-likelihood ratio as

$$\begin{aligned}
L_t &= \sum_{n=1}^{t} \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbf{1}_{\{V_n = v\}} \ln\left(\frac{\mathrm{d}\mathbb{P}_\nu(Z_{n,u}, Y_{n,(v,u)} V_n)}{\mathrm{d}\mathbb{P}_{\nu'}(Z_{n,u}, Y_{n,(v,u)}, V_n)}\right), \\
&= \sum_{n=1}^{t} \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbf{1}_{\{V_n = v\}} \left[\mathbf{1}_{\{Y_{n,(v,u)}=1\}} \ln\left(\frac{f_u(Z_{n,u}) g_{v,u}(1)}{f'_u(Z_{n,u}) g'_{v,u}(1)}\right)\right. \\
&\qquad \left. + \mathbf{1}_{\{Y_{n,(v,u)}=0\}} \ln\left(\frac{g_{v,u}(0)}{g'_{v,u}(0)}\right)\right], \\
&= \sum_{u \in V} \sum_{v \in N_{in}(u)} \sum_{n=1}^{N_v(t)} \left[\mathbf{1}_{\{Y_{n,(v,u)}=1\}} \ln\left(\frac{f_u(W_{n,u}) g_{v,u}(1)}{f'_u(W_{n,u}) g'_{v,u}(1)}\right) + \mathbf{1}_{\{Y_{n,(v,u)}=0\}} \ln\left(\frac{g_{v,u}(0)}{g'_{v,u}(0)}\right)\right].
\end{aligned}$$

where $(W_{n,u})_n$ is a sequence of i.i.d. random variables distributed according to ν_u. Hence

$$\begin{aligned}
\mathbb{E}_\nu[L_\tau] = \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbb{E}_\nu \Bigg[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_v(\tau) \geq n\}} \Big[\mathbf{1}_{\{Y_{n,(v,u)}=1\}} \ln\left(\frac{f_u(W_{n,u})}{f'_u(W_{n,u})}\right) \\
\mathbf{1}_{\{Y_{n,(v,u)}=1\}} \ln\left(\frac{g_{v,u}(1)}{g'_{v,u}(1)}\right) + \mathbf{1}_{\{Y_{n,(v,u)}=0\}} \ln\left(\frac{g_{v,u}(0)}{g'_{v,u}(0)}\right) \Big] \Bigg].
\end{aligned}$$

Note that $\{N_v(\tau) \geq n\} = \{N_v(\tau) \leq n-1\} \in \mathcal{F}_{n-1}$, therefore $W_{n,u}$ and $Y_{n,(v,u)}$ are independent of that event. Hence

$$\begin{aligned}
\mathbb{E}_\nu[L_\tau] &= \sum_{u \in V} \sum_{v \in N_{in}(u)} \sum_{n=1}^{\infty} \mathbb{P}_\nu(N_v(\tau) \geq n) \left[G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \mathrm{kl}(G_{v,u}, G'_{v,u})\right], \\
&= \sum_{u \in V} \sum_{v \in N_{in}(u)} \mathbb{E}_\nu[N_v(\tau)] \left[G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \mathrm{kl}(G_{v,u}, G'_{v,u})\right].
\end{aligned}$$

Therefore, applying (Kaufmann et al., 2016, Lemma 1) at $t = \tau$, $\mathbb{E}_\nu[L_\tau] \geq \mathrm{kl}(\delta, 1-\delta)$.

Consider the set of confusing models $\mathrm{Alt}(\nu) = \{\nu' = (\mu', G') : a^\star(\mu) \neq a^\star(\mu'), \nu \ll \nu'\}$, and define the selection rate of a vertex v as $\omega_v = \mathbb{E}_\nu[N_v(\tau)]/\mathbb{E}_\nu[\tau]$. Then, by minimizing over the set of confusing models, and then optimizing over $\omega = (\omega_v)_{v \in V}$ over the simplex $\Delta(V)$, we obtain

$$\underbrace{\mathbb{E}_\nu[\tau] \sup_{w \in \Delta(V)} \inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \left[G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \mathrm{kl}(G_{v,u}, G'_{v,u})\right]}_{=:(T^\star(\nu))^{-1}} \geq \mathrm{kl}(\delta, 1-\delta).$$

Hence, consider now the expression $\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} w_v \left[G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \mathrm{kl}(G_{v,u}, G'_{v,u})\right]$ and observe that it simplifies as in the proof of Theorem 1:

$$\begin{aligned}
&\inf_{\nu' \in \mathrm{Alt}(\nu)} \sum_{u \in V} \sum_{v \in N_{in}(u)} \omega_v \left[G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \mathrm{kl}(G_{v,u}, G'_{v,u})\right] \\
&\quad = \min_{u \neq a^\star} \inf_{G', \nu'_u, \nu'_{a^\star} : \mu'_u \geq \mu'_{a^\star}} \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \mathrm{KL}(\nu_u, \nu'_u) + \sum_{v \in N_{in}(a^\star)} \omega_v G_{v,a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}), \\
&\quad = \min_{u \neq a^\star} \inf_{G', \nu'_u, \nu'_{a^\star} : \mu'_u \geq \mu'_{a^\star}} m_u \mathrm{KL}(\nu_u, \nu'_u) + m_{a^\star} \mathrm{KL}(\nu_{a^\star}, \nu'_{a^\star}),
\end{aligned}$$

as in the proof of Theorem 1. For the known graph case, note that the set of confusing models becomes $\mathrm{Alt}'(\nu) := \{\nu' = (\mu', G) : a^\star(\mu) \neq a^\star(\mu'), \nu_u \ll \nu'_u \ \forall u \in V\}$, from which one can conclude the same result. □

C.3 Scaling Properties

To gain a better intuition of the characteristic time in Theorem 1, we can focus on the Gaussian case where $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$. For this case we have that $\mathrm{KL}(\nu_u, \nu_v) = (\mu_u - \mu_v)^2/(2\lambda^2)$, and $I_\alpha(\nu_u, \nu_v) = \dfrac{\alpha(1-\alpha)(\mu_u - \mu_v)^2}{2\lambda^2}$. Therefore $T^\star(\nu)$ can be computed by solving the following convex problem

$$\begin{aligned} T^\star(\nu) = \inf_{m \in \mathbb{R}_+^K, \omega \in \Delta(V)} \max_{u \neq a^\star} \left(m_u^{-1} + m_{a^\star}^{-1}\right) \frac{2\lambda^2}{\Delta_u^2} \\ \text{s.t. } m_u = \sum_{v \in N_{in}(u)} \omega_v G_{v,u} \quad \forall u \in V. \end{aligned} \tag{13}$$

C.3.1 General case

We restate here Proposition 1 and provide a proof.

Proposition 7. *For an observable model $\nu = (\{\nu_u\}_u, G)$ with a graph G satisfying $\delta(G) + \sigma(G) > 0$ and Gaussian random rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$ we can upper bound $T^\star$ as*

$$T^\star(\nu) \leq \frac{4\left[\delta(G) + \sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor\right] \lambda^2}{\min_{u \neq a^\star} \min(\bar{G}_u, \bar{G}_{a^\star}) \Delta_u^2},$$

where $\bar{G}_u := \max\left(\max_{v \in D(G)} G_{v,u}, \min_{v \in L(G): G_{v,u} > 0} G_{v,u}\right)$ (sim. $\bar{G}_{a^\star}$).

Proof. Since $\sigma(G) + \delta(G) > 0$, the graph has at least one of the following: (i) a positive weak-domination number $\delta(G)$ or (ii) at least one strongly observable vertex with a self-loop $\sigma(G)$ (or both properties hold at the same time).

Consider the expression of $T^\star(\nu)$ in Eq. (13). Note that for each weakly observable vertex u there exists $v \in N_{in}(u) \cap D(G)$, where $D(G)$ is the smallest set dominating the set of weakly observable vertices (see Definition 3).

Observe the following two properties:

1. For any strongly observable vertex by Theorem 4 we need at-most $\sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor$ vertices in $L(G)$ to dominate $|SO(G)|$ if $\sigma(G) > 0$.
2. If $\sigma(G) = 0$, then by definition all strongly observable vertices lack self-loops, and $\delta(G)$ must be positive; thus $D(G)$ dominates the entire graph.

Therefore we need at-most $\kappa = \delta(G) + \sigma(G) - \left\lfloor \frac{\sigma(G)}{\alpha(G)+1} \right\rfloor$ vertices to dominate the graph. Using this information, we allocate probability mass uniformly across these vertices.

1. For any vertex $u \in W(G)$ let $\omega_v = 1/\kappa$ for all $v \in D(G)$. This fact allows us to lower bound m_u as

$$\begin{aligned} m_u &= \sum_{v \in N_{in}(u) \setminus D(G)} \omega_v G_{v,u} + \sum_{w \in D(G) \cap N_{in}(u)} \omega_w G_{w,u}, \\ &\geq \sum_{w \in D(G) \cap N_{in}(u)} \omega_w G_{w,u}, \\ &\geq \sum_{w \in D(G) \cap N_{in}(u)} \frac{1}{\kappa} G_{w,u}, \\ &\geq \max_{w \in D(G) \cap N_{in}(u)} \frac{1}{\kappa} G_{w,u}, \\ &\geq \max_{w \in D(G)} \frac{1}{\kappa} G_{w,u}. \end{aligned}$$

2. For any vertex $u \in L(G)$ with a self-loop we can lower bound m_u as

$$m_u \geq \min_{v \in L(G): G_{v,u} > 0} \frac{1}{\kappa} G_{v,u}.$$

Therefore, for any $u \in V$ we have

$$\frac{1}{\kappa m_u} \leq \begin{cases} \dfrac{1}{\min_{v \in L(G): G_{v,u} > 0} G_{v,u}} & \text{if } u \in L(G), \\ \dfrac{1}{\max_{v \in D(G)} G_{v,u}} & \text{otherwise,} \end{cases}$$

Then, let $\bar{G}_u \coloneqq \max\left(\max_{v \in D(G)} G_{v,u}, \min_{v \in L(G): G_{v,u} > 0} G_{v,u}\right)$. In other words, $\bar{G}_u$ captures the relevant edge-activation probability from either a dominating vertex or a self-loop vertex. We conclude that

$$\begin{aligned} T^\star(\nu) &\leq \max_{u \neq a^\star} \kappa \left(\bar{G}_u^{-1} + \bar{G}_{a^\star}^{-1}\right) \frac{2\lambda^2}{\Delta_u^2}, \\ &\leq \frac{4\kappa\lambda^2}{\min_{u \neq a^\star} \min(\bar{G}_u, \bar{G}_{a^\star}) \Delta_u^2}, \end{aligned}$$

where in the last expression we used $a + b \leq 2\max(a, b)$. □

C.3.2 The loopless clique

We now consider the scaling for the case $\delta(G) + \sigma(G) = 0$, which corresponds to the loopless clique. We restate here Proposition 2 and provide a proof.

Proposition 8. *For an observable model $\nu = (\{\nu_u\}_u, G)$ with $\delta(G) + \sigma(G) = 0$, and Gaussian random rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$, we can upper bound $T^\star$ as*

$$T^\star(\nu) \leq \frac{4\bar{G}\lambda^2}{\Delta_{\min}^2},$$

where

$$\bar{G} \coloneqq \min_{(v,w) \in V^2: v \neq w} \max_{u \neq a^\star} \frac{1}{G_{v,w}(u)} + \frac{1}{G_{v,w}(a^\star)},$$

and $G_{v,w}(u) \coloneqq G_{v,u} + G_{w,u}$.

Proof. The proof uses the idea that only 2 vertices are needed to dominate the graph. Hence, we find two vertices (v, w) that allow to bound $\max_{u \neq a^*}(m_u^{-1} + m_{a^*}^{-1})$ in the expression of $T^\star$.

First, $\delta(G) = 0$ implies that there are no weakly observable vertices. Since we also have $\sigma(G) = 0$ then $V = E(G)$ (where $E(G)$ is the set of strongly observable vertices without self-loops). Hence, by definition, the graph is the loopless clique. By Theorem 4 we need 2 vertices to dominate the graph. Define the following set

$$\mathcal{A} \coloneqq \operatorname*{arg\,min}_{(v,w) \in V^2: v \neq w} \max_{u \neq a^\star} G_{v,w}(u)^{-1} + G_{v,w}(a^\star)^{-1}, \text{ where } G_{v,w}(u) \coloneqq G_{v,u} + G_{w,u},$$

and denote by $\bar{G} \coloneqq \min_{(v,w) \in V^2: v \neq w} \max_{u \neq a^\star} G_{v_0,w_0}(u)^{-1} + G_{v_0,w_0}(a^\star)^{-1}$ the optimal value for any $(v_0, w_0) \in \mathcal{A}$. Then, let $\omega_{v_0} = \omega_{w_0} = 1/2$ for a generic pair $(v_0, w_0) \in \mathcal{A}$. For any $u \in V$ we have m_u

$$\begin{aligned} m_u &= \sum_{v \in V} w_v G_{v,u} \\ &\geq (G_{v_0,u} + G_{w_0,u})/2. \end{aligned}$$

Note that by Lemma 4, $m_u > 0$ for any $u \in V$. Therefore,

$$m_u^{-1} \leq 2 G_{v_0,w_0}(u)^{-1}.$$

Since the bound hold for any $(v_0, w_0) \in \mathcal{A}$, we conclude that

$$\begin{aligned} T^*(\nu) &\leq \max_{u \neq a^*} (m_u^{-1} + m_{a^*}^{-1}) \frac{2\lambda^2}{\Delta_u^2}, \\ &\leq \max_{u \neq a^*} \left(G_{v_0,w_0}(u)^{-1} + G_{v_0,w_0}(a^\star)^{-1} \right) \frac{4\lambda^2}{\Delta_{\min}^2}, \\ &= \frac{4\bar{G}\lambda^2}{\Delta_{\min}^2}. \end{aligned}$$

□

C.3.3 Heuristic solution: scaling and spectral properties

In general we find it hard to characterize the scaling of $T^\star(\nu)$ in terms of the spectral properties of G without a more adequate analysis. Furthermore, we also wonder if it is possible to find a closed-form solution that can be easily used.

Intuition suggests that, by exploiting the underlying topology of the graph, a good solution ω should be sparse (which, in turns, helps to minimize the sample complexity). However, it may be hard to find a simple sparse solution.

To that aim, we can gain some intuition from the BAI problem for the classical multi-armed bandit setup. From the analysis in Garivier and Kaufmann (2016, Appendix A.4), an approximately optimal solution in the Gaussian case is given by $\omega_u^\star \propto 1/\Delta_u^2$, with $\Delta_{a^\star} = \Delta_{\min}$.

Hence, we also propose that $m_u \propto 1/\Delta_u^2$, with $m = G^\top \omega$. At this point we could try to minimize the MSE loss between m and the vector $\Delta^{-2} \coloneqq (1/\Delta_u^2)_{u \in V}$, subject to $\|w\|_1 = 1$. However, in this problem we are more interested in the directional alignment between m and Δ^{-2}, rather than in their magnitude, since the magnitude of ω is constrained [5]. Therefore, one may be interested in maximizing teh similarity $m^\top \Delta^{-2}$, or rather

$$\max_{\omega : \|\omega\|_2 \leq \alpha} \omega^\top G \Delta^{-2}$$

for some constraint α. Obviously the solution is $\omega = \alpha G\Delta^{-2}/\|G\Delta^{-2}\|_2$ in the classical Euclidean space with the ℓ_2 norm. However, such solution is not a distribution. To that aim, we project $G\Delta^{-2}$ on the closest distribution in the KL sense, defined as $\mathrm{Proj}_{\mathrm{KL}}(x) \coloneqq \min_{p \in \mathcal{P}} \mathrm{KL}(p, x)$ for any x such that $x_u \geq 0$ for every u.

Lemma 8. *The projection of $G\Delta^{-2}$ in the KL sense is given by $\omega_{\mathrm{heur}} = G\Delta^{-2}/\|G\Delta^{-2}\|_1$.*

Proof. For some vector x satisfying $x \geq 0$, write the Lagrangian of $\min_{p \in \mathcal{P}} \mathrm{KL}(p, x)$:

$$\mathcal{L}(p, \lambda) = \sum_u p_u \ln\left(\frac{p_u}{x_u}\right) + \lambda \left(1 - \sum_u p_u\right).$$

Then, we check the first order condition $\partial \mathcal{L}/\partial p_u = 0 \Rightarrow \ln\left(\frac{p_u}{x_u}\right) + 1 - \lambda = 0$, implying that $p_u = x_u e^{\lambda - 1}$. Using the fact that $\sum_u p_u = 1$ we also obtain $e^{1-\lambda} = \sum_u x_u = \|x\|_1$. Therefore $\lambda = 1 - \ln(\|x\|_1)$, from which we conclude that $p_u = x_u/\|x\|_1$. □

Such allocation $w_{\mathrm{heur}} \coloneqq \frac{G\Delta^{-2}}{\|G\Delta^{-2}\|_1}$ makes intuitively sense: a vertex u will be chosen with probability proportional to $\sum_{v \in V} G_{uv}\Delta_v^{-2}$, thus assigning higher preference to vertices that permit the learner to sample arms with small sub-optimality gaps.

For this heuristic allocation w_{heur} we can provide the following upper bound on its scaling.

[5] Note that also the MSE problem $\arg\min_\beta \|y - A\beta\|_2^2 = \arg\min_\beta -2y^\top A\beta + \|A\beta\|_2^2$ also tries to solve a problem of alignment through the term $2y^\top A\beta$, while the second term $\|A\beta\|_2^2$ can be considered a form of regularization.

Lemma 9. *For a model $\nu = (\{\nu_u\}_u, G)$ with an observable graph G and Gaussian random rewards $\nu_u = \mathcal{N}(\mu_u, \lambda^2)$ we can upper bound $T(\omega_{\text{heur}}; \nu)$ as*

$$T^\star(\nu) \leq T(\omega_{\text{heur}}; \nu) \leq \frac{\|G\Delta^{-2}\|_1}{\sigma_{\min}(G)^2} \cdot 4\lambda^2 \leq \frac{K \min\left(\sqrt{K}\sigma_{\max}(G), \sum_i \sigma_i(G)\right)}{\Delta_{\min}^2 \sigma_{\min}(G)^2} \cdot 4\lambda^2.$$

where $\sigma_{\min}(G), \sigma_{\max}(G), \sigma_i(G)$ are, respectively the minimum singular value of G, the maximum singular value of G and the i-th singular value of G (to not be confused with $\sigma(G)$!).

Proof. Again, we prove this corollary by lower bound each m_u. Note that we have $m = G^\top G \omega_{\text{heur}}$. Denote by G_u the u-th row of G, then

$$\begin{aligned} m_u &= \frac{\sum_{v \in V} (G^T G)_{u,v} \Delta_v^{-2}}{\|G\Delta^{-2}\|_1} \\ &\geq \frac{\|G_u\|_2^2 \Delta_u^{-2}}{\|G\Delta^{-2}\|_1} \\ &\geq \frac{\|G_u\|_2^2 \Delta_u^{-2}}{\|G\Delta^{-2}\|_1}. \end{aligned}$$

And therefore, using that $\|G_u\|_2^2 \geq \sigma_{\min}(G)^2$, we observe that

$$\begin{aligned} T(\omega_{\text{heur}}; \nu) &\leq \max_{u \neq a^\star} \left(\frac{\Delta_u^2}{\|G_u\|_2^2} + \frac{\Delta_{\min}^2}{\|G_{a^\star}\|_2^2} \right) \frac{2\|G\Delta^{-2}\|_1 \lambda^2}{\Delta_u^2}, \\ &\leq \max_{u \neq a^\star} \left(\Delta_u^2 + \Delta_{\min}^2 \right) \frac{2\|G\Delta^{-2}\|_1 \lambda^2}{\sigma_{\min}(G)^2 \Delta_u^2}, \\ &\leq \frac{\|G\Delta^{-2}\|_1}{\sigma_{\min}(G)^2} \cdot 4\lambda^2. \end{aligned}$$

Lastly, denoting by G_u the u-th row of G, we obtain that $\|G\Delta^{-2}\|_1 \leq \sum_u |G_u^\top \Delta^{-2}| \leq \|\Delta^{-2}\|_2 \sum_u \|G_u\|_2 \leq \frac{\sqrt{K}}{\Delta_{\min}^2} \sum_u \|G_u\|_2$. We conclude by noting that $\sum_u \|G_u\|_2 \leq K\sigma_{\max}(G)$, and thus $\|G\Delta^{-2}\|_1 \leq \frac{K^{3/2}\sigma_{\max}(G)}{\Delta_{\min}^2}$

Additionally, we also note that $\|G\Delta^{-2}\|_1 \leq \sum_u |G_u^\top \Delta^{-2}| \leq \|\Delta^{-2}\|_\infty \sum_u \|G_u\|_1$ by Holder's inequality. Now, let $\|\cdot\|_*$ denote the Schatten-1 norm. Using that $\|\text{vec}(G)\|_1 \leq K\|G\|_* = K\sum_i \sigma_i(G)$ we have $\|G\Delta^{-2}\|_1 \leq \frac{K\sum_i \sigma_i(G)}{\Delta_{\min}^2}$. □

An alternative approach that is sparse. We note that the above analysis does not take fully advantage of the graph structure. An alternative approach that yields a better scaling is to instead consider the similarity problem

$$\max_{\omega : \|\omega\|_1 = 1} \omega^\top G\Delta^{-2}.$$

The optimal solution then is simply $\omega_u = \mathbf{1}_{\{u \in \mathcal{G}\}}/|\mathcal{G}|$, where $\mathcal{G} = \arg\max_v (G\Delta^{-2})_v$. This is an efficient allocation, since it scales as $O\left(\frac{|\mathcal{G}|}{\Delta_{\min}^2 \max_{u \in \mathcal{G}} \min_v G_{u,v}}\right)$. However, such solution is admissible only if it guarantees that $V \ll \mathcal{G}$, i.e., this set of vertices $\mathcal{G}$ dominates the graph.

Alternatively, one can choose the top k vertices $\mathcal{U} = \{u_1, u_2, \ldots, u_k\}$ ordered according to $(G\Delta^{-2})_{u_1} \geq \cdots \geq (G\Delta^{-2})_{u_k} \geq \ldots (G\Delta^{-2})_{u_K}$ satisfying $V \ll \{u_1, \ldots, u_k\}$ (i.e., these vertices dominate the graph). Then, one can simply let $\omega_u = \mathbf{1}_{\{u \in \mathcal{U}\}}/|\mathcal{U}|$. Since these are also the vertices that maximize the average information collected from the graph, we believe this solution to be sample efficient. A simple analysis, shows that the worst case scaling in this scenario is $O\left(\frac{|\mathcal{U}|}{\Delta_{\min}^2 \max_{u \in \mathcal{U}} \min_v G_{u,v}}\right)$.

D Analysis of TaS-FG

In this section we provide an analysis of TaS-FG for an observable model in the uninformed case (with continuous rewards) or in the informed case. In Appendix D.1 we analyse the sampling rule. In Appendix D.2 we analyse the stopping rule. Lastly, in Appendix D.3, we analyse the sample complexity.

D.1 Sampling Rule

The proof of the tracking proposition Proposition 5 is inspired by D-tracking Garivier and Kaufmann (2016); Jedra and Proutiere (2020). In Degenne and Koolen (2019) they show that classical D-tracking Garivier and Kaufmann (2016) may fail to converge when $C^\star(\nu)$ is a convex set of possible optimal allocations. However, using a modified version it is possible to prove the convergence. We take inspiration from Jedra and Proutiere (2020), where they applied a modified D-tracking to the linear bandit case and showed convergence of $w^\star(t)$.

The intuition behind the proof is that tracking the average of the converging sequence $(\omega^\star(t))_t$, which is a convex combination, converges to a stable point in the convex set $C^\star(\nu)$. The proof makes use of the following result, from Berge (1963).

Theorem 6 (Maximum theorem Berge (1963))**.** *Let $C^\star(\nu) = \arg\inf_{\omega\in\Delta(V)} T(\omega;\nu)$. Then $T^\star(\nu) \coloneqq T(\omega^\star;\nu), \omega^\star \in C^\star(\nu)$, is continuous at ν (in the sense of $(G,\{\mu_u\}_u) \in [0,1]^{K\times K}\times[0,1]^K$), $C^\star(\nu)$ is convex, compact and non-empty. Furthermore, we have that for any open neighborhood $\mathcal{V}$ of $C^\star(\nu)$, there exists an open neighborhood $\mathcal{U}$ of ν, such that for all $\nu'\in\mathcal{U}$ we have $C^\star(\nu')\subseteq\mathcal{V}$.*

Here we state, and prove, a more general version, of Proposition 5 (which follows by taking $\alpha_{t,n} = 1/t$ in the next proposition).

Proposition 9. *Let $S_t = \{u\in V : N_u(t) < \sqrt{t} - K/2\}$. The D-tracking rule, defined as*

$$V_t \in \begin{cases} \arg\min_{u\in S_t} N_u(t) & S_t \neq \emptyset \\ \arg\min_{u\in V} N_u(t) - t\sum_{n=1}^t \alpha_{t,n}\omega_u^\star(n) & \text{otherwise} \end{cases}, \tag{14}$$

where, for every $t\geq 1$, the sequence $\alpha_t = (\alpha_{t,n})_{n=1}^t$ satisfies: (1) $\alpha_{t,n}\in[0,1]$; (2) for every fixed $n\in\{1,\dots,t\}$ we have $\alpha_{t,n} = o(1)$ in t ; (3) for all t, $\sum_{n=1}^t \alpha_{t,n} = 1$.

Such tracking rule ensures that for all $\epsilon>0$ there exists $t(\epsilon)$ such that for all $t\geq t(\epsilon)$ we have

$$\|N(t)/t - \bar{v}(t)\|_\infty \leq 5(K-1)\epsilon,$$

where $\bar{v}(t) \coloneqq \arg\inf_{\omega\in C^\star(\nu)} \|\omega - \sum_{n=1}^t \alpha_n\omega^\star(n)\|_\infty$, and $\lim_{t\to\infty}\inf_{\omega\in C^\star(\nu)}\|N(t)/t-\omega\|_\infty \to 0$ almost surely.

Proof. Define the projection of $x\in\mathbb{R}^n$ onto C as $\mathrm{Proj}_C(x) \coloneqq \arg\inf_{\omega\in C}\|x-w\|_\infty$, which is guaranteed to exists if C is convex and compact.

Let $n(t) \coloneqq \mathrm{Proj}_{C^\star(\nu)}(N_t/t)$. The proof lies showing that $\|N_t/t - n(t)\|_\infty \to 0$ as $t\to\infty$. To that aim, we first need to show that $\inf_{w\in C^\star(\nu)}\|\bar{\omega}^\star(t) - w\|_\infty \to 0$, where $\bar{\omega}^\star(t) \coloneqq \sum_{n=1}^t \alpha_n\omega^\star(n)$ is a convex combination of the estimated optimal allocations up to time t.

Begin by defining the following quantities

$$\bar{v}(t) \coloneqq \mathrm{Proj}_{C^\star(\nu)}(\bar{\omega}^\star(t)) \text{ and } v(t) \coloneqq \mathrm{Proj}_{C^\star(\nu)}(\omega^\star(t)),$$

which are, respectively, the projection onto $C^\star(\nu)$ of the average estimated allocation and the projection of the last estimated allocation.

By the forced exploration step, we have that $N_u(t)\to\infty$ for every $u\in V$. Since the model is observable, we can invoke the law of large number and guarantee that $\mathbb{P}_\nu(\lim_{t\to\infty}\hat{\nu}(t) = \nu) = 1$ in the sense that $(\hat{G}(t),\hat{\mu}(t))\to(G,\mu)$ almost surely. Then, by continuity of the problem (see Theorem 6) we have that $\forall\epsilon>0\exists t_0(\epsilon) : \sup_{\omega\in C^\star(\hat{\nu}(t))}\|\omega - \mathrm{Proj}_{C^\star(\nu)}(\omega)\|_\infty \leq \epsilon$ for all $t\geq t_0(\epsilon)$.

Henceforth, for $t \geq t_0(\epsilon)$ we have $\|\omega^\star(t) - v(t)\|_\infty \leq \sup_{w \in C^\star(\hat{\nu}(t))} \|\omega - \mathrm{Proj}_{C^\star(\nu)}(\omega)\|_\infty \leq \epsilon$, thus we derive

$$\begin{aligned}\left\|\sum_{n=1}^{t} \alpha_{t,n} v(n) - \bar{\omega}^\star(n)\right\|_\infty &\leq \sum_{n=1}^{t_0(\epsilon)} \alpha_{t,n} \|v(n) - \omega^\star(n)\|_\infty + \sum_{n=t_0(\epsilon)+1}^{t} \alpha_{t,n} \|v(n) - \omega^\star(n)\|_\infty, \\ &\leq t_0(\epsilon) \bar{\alpha}_{t,t_0(\epsilon)} + \epsilon,\end{aligned}$$

where $\bar{\alpha}_{t,t_0(\epsilon)} = \max_{1 \leq n \leq t_0(\epsilon)} \alpha_{t,n}$ and we used the fact that $\sum_n \alpha_{t,n} = 1$. Hence, for any $x \in C^\star(\nu)$ note that $\|\bar{\omega}^\star(t) - \bar{v}(t)\|_\infty \leq \|\bar{\omega}^\star(t) - x\|_\infty$. For a fixed t, one can choose $x = \sum_{n=1}^{t} \alpha_{t,n} v(n)$ since every $v(n) \in C^\star(\nu)$, and also a convex combination belongs to $C^\star(\nu)$ by convexity. Henceforth

$$\|\bar{\omega}^\star(t) - \bar{v}(t)\|_\infty \leq \left\|\bar{\omega}^\star(t) - \sum_{n=1}^{t} \alpha_{t,n} v(n)\right\|_\infty \leq t_0(\epsilon) \bar{\alpha}_{t,t_0(\epsilon)} + \epsilon.$$

Since for every fixed n we have $\bar{\alpha}_{t,n} = o(1)$ in t, there exists $t_1(\epsilon)$ such that for $t \geq t_1(\epsilon)$ we have $\bar{\alpha}_{t,t_0(\epsilon)} \leq \epsilon / t_0(\epsilon)$, which implies that $\|\bar{\omega}^\star(t) - \bar{v}(t)\|_\infty \leq 2\epsilon$ for $t \geq \max(t_0(\epsilon), t_1(\epsilon))$.

Now we prove that $\|N_t/t - n(t)\|_\infty \to 0$ as $t \to \infty$. Define $t_2(\epsilon) \coloneqq \max(t_0(\epsilon), t_1(\epsilon))$, and observe that $\|N_t/t - n(t)\|_\infty \leq \|N_t/t - \bar{v}(t)\|_\infty$.

Therefore we are interested in bounding the quantity $\|N_t/t - \bar{v}^\star(t)\|_\infty$, and we use similar arguments as in (Garivier and Kaufmann, 2016, Lemma 17). Let $t \geq t_2(\epsilon)$. Define $E_u(t) = N_u(t) - t\bar{v}_u(t)$ for every $u \in V$ and note that $\sum_u E_u(t) = 0 \Rightarrow \min_u E_u(t) \leq 0$. We want to show that $\sup_u |E_u(t)/t|$ is bounded.

We begin by showing the following

$$\{U_{t+1} = u\} \subseteq \mathcal{E}_1(t) \cup \mathcal{E}_2(t) \subseteq \{E_u(t) \leq 2t\epsilon\},$$

where $\mathcal{E}_1(t) = \{u = \arg\min_{u \in V} N_u(t) - t\bar{\omega}^\star_u(t)\}$ and $\mathcal{E}_2(t) = \{N_u(t) \leq g(t)\}$ with $g(t) = \max(0, (\sqrt{t} - K/2)) - 1$.

For the first part, if $\{U_{t+1} = u\} \subseteq \mathcal{E}_1(t)$, since $t \geq t_2(\epsilon)$ we have

$$\begin{aligned}E_u(t) &= N_u(t) - t\bar{v}_u(t) \pm t\bar{\omega}^\star_u(t), \\ &\leq N_u(t) - t\bar{\omega}^\star_u(t) + 2t\epsilon, \\ &= \min_v N_v(t) - t\bar{\omega}^\star_v(t) + 2t\epsilon, \\ &\leq \min_v E_v(t) + 4t\epsilon, \\ &\leq 4t\epsilon.\end{aligned}$$

where in the second equality we used the fact that $\{U_{t+1} = u\} \subset \mathcal{E}_1(t)$ and in the last inequality that $\min_v E_v(t) \leq 0$.

For the second part, as shown in (Garivier and Kaufmann, 2016, Lemma 17), there exists $t_3(\epsilon)$ such that for $t \geq t_3(\epsilon)$ then $g(t) \leq 4t\epsilon$ and $1/t \leq \epsilon$. Then if $\{U_{t+1} = u\} \subseteq \mathcal{E}_2(t)$ we have that

$$E_u(t) \leq g(t) - t\bar{v}_u(t) \leq 4t\epsilon.$$

Therefore, as in (Garivier and Kaufmann, 2016, Lemma 17), one can conclude that for $t \geq t' \coloneqq \max(t_2(\epsilon), t_3(\epsilon))$

$$E_u(t) \leq \max(E_u(t'), 4t\epsilon + 1).$$

Using that $\sum_u E_u(t) = 0$, and that for all $t \geq t', E_u(t') \leq t'$ and $1/t \leq \epsilon$ we have that

$$\sup_i |E_u(t)/t| \leq (K-1)\max(t'/t, 4\epsilon + 1/t) \leq (K-1)\max(5\epsilon, t'/t).$$

Hence, there exists $t'' \geq t'$ such that for all $t \geq t''$ we have $\sup_i |E_u(t)/t| \leq 5(K-1)\epsilon$. Letting $\epsilon \to 0$ concludes the proof. □

Hence, Proposition 5 follows by choosing $\alpha_{t,n} = 1/t$ in the previous proposition. Another possible choice is the exponential smoothing factor $\alpha_{t,n} = \kappa_t \lambda^{t-n}$ with $\lambda \in (0,1)$ and $\kappa_t = \frac{1-\lambda}{1-\lambda^t}$. In the next subsection we investigate which choice of $\alpha_{t,n}$ is better.

D.1.1 Is the average of the allocations the best convex combination?

A natural question that arises is which factor $\alpha_{t,n}$ to use. Why is $\alpha_{t,n} = 1/t$ a good choice? This question is related to the fluctuations of the underlying process, and to the stability of the exploration process. We try to give an answer by looking at the variance of the resulting allocation $\bar{w}^\star(t)$.

The i.i.d. case. We begin by considering the i.i.d. case, which supports the fact that a simple average is a good approach to minimize variance.

Lemma 10. *Consider an i.i.d. sequence of Gaussian random variables $\{X_n\}_n$ with 0 mean and variance σ^2. Let $\bar{X}_t(w) = \sum_{n=1}^t w_n X_n$, with $\{w_n\} \in \Delta(\{1,\dots,t\}) = \Delta([t])$. Then*

$$\min_{w\in\Delta([t])} \mathrm{Var}(X_t(w)) = \frac{\sigma^2}{t},$$

which is achieved for $w_i = 1/t$, for all $i \in [t]$.

Proof. Note that $\mathrm{Var}(X_t(w)) = \sum_{n=1}^t w_n^2\sigma^2$. Introduce the Lagrangian $\mathcal{L}(w,\lambda) = \sum_{n=1}^t w_n^2\sigma^2 + \lambda(1 - \sum_{n=1}^t w_n)$. Checking the first order condition yields $d\mathcal{L}/dw_n = 2w_n\sigma^2 = \lambda$, hence $w_n = \lambda/(2\sigma^2)$. Since $\sum_n w_n = 1$ we must have $t\lambda/(2\sigma^2) = 1 \Rightarrow \lambda = 2\sigma^2/t$. Therefore $w_n = 1/t$. We conclude that $\min_w \mathrm{Var}(X_t(w)) = \sigma^2/t$. □

More in general, the weighting should be inversely proportional to the variance of the underlying random variable, according to the inverse variance weighting principle Hartung et al. (2011). That is, one can use the same approach as in the previous lemma to easily derive that in case $X_n \sim \mathcal{N}(0,\sigma_n^2)$, then the optimal weighting is $w_n = \frac{1}{\sigma_n^2}\left(\sum_{k=1}^t \frac{1}{\sigma_k^2}\right)^{-1}$.

A more complex case. While the i.i.d. case seems to indicate that taking a simple average is a good approach to minimize the variance, it may not always be the case. In fact, we may expect the random variable $w^\star(t)$ to have smaller fluctuations as t grows larger. We try to give a more complete picture by also looking at a more complex case.

Consider a process $X_n = X_{n-1} + \xi_n$, where ξ_n is an i.i.d. zero-mean process with variance $\mathbb{E}[\xi_n^2] \le C/n^{1+\alpha}$ for some $\alpha, C > 0$ and $n \ge 2$. And let $X_1 = \xi_1$, with $\mathbb{E}[\xi_1] \le \sigma^2$.

We define S_t^{avg} to be be the simple average

$$S_t^{avg} = \frac{1}{t}(X_1 + X_2 + \cdots + X_t)$$

Similarly, we define the exponentially smoothed average with $\kappa_t = (1-\lambda)/(1-\lambda^t)$:

$$S_t^{exp} = \kappa_t \sum_{n=1}^t \lambda^{t-n} X_n.$$

Then, we obtain the following result on the variance of the two averages.

Lemma 11. *Consider the simple average S_t^{avg} and the exponentially smoothed average S_t^{exp} with factor $\lambda \in (0,1)$. Then $\mathrm{Var}(S_t^{avg}) \le \mathrm{Var}(S_t^{exp})$.*

Proof. Let S_t^{avg} be the simple average, and note the following rewriting:

$$S_t^{avg} = \frac{1}{t}(X_1 + X_2 + \cdots + X_t) = X_1 + \frac{(t-1)}{t}\xi_2 + \cdots + \xi_t = \sum_{n=1}^t \frac{t-n+1}{t}\xi_n.$$

Also rewrite the exponentially smoothed average as follows

$$\begin{aligned}
S_t^{exp} &= \kappa_t \sum_{n=1}^{t} \lambda^{t-n} X_n, \\
&= \kappa_t \sum_{n=1}^{t} \lambda^{t-n} \sum_{i=1}^{n} \xi_i, \\
&= \kappa_t \sum_{i=1}^{t} \xi_i \sum_{n=i}^{t} \lambda^{t-n}, \\
&= \kappa_t \sum_{i=1}^{t} \xi_i \sum_{n=0}^{t-i} \lambda^{n}, \\
&= \kappa_t \sum_{i=1}^{t} \xi_i \frac{1-\lambda^{t-i+1}}{1-\lambda}, \\
&= \frac{1}{1-\lambda^t} \sum_{i=1}^{t} (1-\lambda^{t-i+1}) \xi_i.
\end{aligned}$$

Due to the properties of X_n we have that $\mathbb{E}[\xi_n] = \mathbb{E}[\mathbb{E}[\xi_n|\mathcal{F}_{n-1}]] = 0$ and $\mathbb{E}[\xi_j \xi_n] = 0$. Therefore, we can write the variance of the averages as follows:

$$\begin{aligned}
\mathrm{Var}(S_t^{avg}) &\leq \sigma^2 + C \sum_{n=2}^{t} \frac{(t-n+1)^2}{t^2 n^{1+\alpha}}, \\
\mathrm{Var}(S_t^{exp}) &\leq \sigma^2 + \frac{C}{(1-\lambda^t)^2} \sum_{n=2}^{t} \frac{(1-\lambda^{t-n+1})^2}{n^{1+\alpha}}.
\end{aligned}$$

To show that $\mathrm{Var}(S_t^{avg}) \leq \mathrm{Var}(S_t^{exp})$ we can check if the following inequality holds for all $n \in \{2, \dots, t\}$:

$$\frac{t-n+1}{t} \leq \frac{1-\lambda^{t-n+1}}{1-\lambda^t}.$$

We can prove that the function $h(x) = \frac{1-\lambda^x}{x}$ is decreasing in $x \in [1, t]$. To that aim, compute the derivative $h'(x) = \frac{-\lambda^x x \ln(\lambda) - 1 + \lambda^x}{x^2}$. We are interested in checking if the numerator is negative. Then

$$-\lambda^x x \ln(\lambda) - 1 + \lambda^x \leq 0 \Rightarrow \lambda^x (1 - x \ln(\lambda)) \leq 1.$$

Rewrite as $e^{x \ln(\lambda)}(1 - x \ln(\lambda)) \leq 1$ and let $y = -x \ln(\lambda)$. Then

$$e^{-y}(1+y) \leq 1 \Rightarrow 1 + y \leq e^y,$$

which is always true for $y \geq 0$.

Hence, we have shown that $h(x) \geq h(t)$ for $x \leq t$. Letting $x = t - n + 1$, with $n = 2, \dots, t$, concludes the proof. □

Unfortunately considering an approach that minimizes the variance is rather difficult. However, numerical experiments seem to suggest that a simple empirical average is an effective approach.

D.2 Stopping Rule

The (fixed-confidence) Best Arm Identification problem in multi-armed bandit models can be seen as a hypothesis testing problem, where we are testing if $\mu \in \mathcal{H}_k$, where $\mathcal{H}_k = \{\mu' : \mu'_k > \max_{j \neq k} \mu'_j\}$. That is, we are testing if the optimal action in μ is k.

Denoting by $n_a(t)$ the number of times the outcome of a certain action a is observed, the generalized likelihood ratio statistics (GLR) for such problem can be written as

$$\inf_{\lambda \in \mathrm{Alt}(\hat{\mu}(t))} \sum_a n_a(t) \mathrm{KL}(\hat{\nu}_a(t), \lambda_a),$$

where $\nu_a(t)$ is the estimated distribution of rewards for arm a, which depends solely on $\hat{\mu}_a(t)$ due to the assumption that ν_a is a single-parameter exponential distribution, and $\mathrm{Alt}(\hat{\mu}(t)) = \{\lambda : \arg\max_a \lambda_a \neq \arg\max_a\}$ is the set of confusing model.

Taking inspiration from such approach, define the following GLR statistic

$$\Lambda(t) \coloneqq \min_{u \neq \hat{a}_t} \inf_{\lambda : \lambda_u \geq \lambda_{\hat{a}_t}} \sum_{v \in V} M_v(t) \mathrm{KL}(\hat{\nu}_v(t), \lambda_v),$$

where λ is an alternative reward parameter. Then, note that $\Lambda(t)$ can be conveniently rewritten as

$$\begin{aligned}\Lambda(t) &= \min_{u \neq \hat{a}_t} (M_u(t) + M_{\hat{a}_t}(t)) I_{\frac{M_{\hat{a}_t}(t)}{M_u(t)+M_{\hat{a}_t}(t)}}(\hat{\nu}_{\hat{a}_t}, \hat{\nu}_u(t)), \\ &= \min_{u \neq \hat{a}_t} M_{\hat{a}_t}(t) \mathrm{KL}(\hat{\nu}_{\hat{a}_t}(t), \hat{\nu}_{\hat{a}_t,u}) + M_u(t) \mathrm{KL}(\hat{\nu}_u(t), \hat{\nu}_{\hat{a}_t,u}),\end{aligned}$$

where $\hat{\nu}_{\hat{a}_t,u}$ is a distribution of rewards depending on the parameter $\hat{\mu}_{\hat{a}_t,u}(t)$ defined as

$$\hat{\mu}_{a,b}(t) = \frac{M_a(t)}{M_a(t) + M_b(t)} \hat{\mu}_a(t) + \frac{M_b(t)}{M_a(t) + M_b(t)} \hat{\mu}_b(t).$$

One can then show that $tT(N_t/t; \hat{\nu}(t))^{-1}$ is equivalent to $\Lambda(t)$. First, observe that

$$T(N_t/t; \hat{\nu}(t))^{-1} = \min_{u \neq \hat{a}_t} (m_u(t) + m_{\hat{a}_t}(t)) I_{\frac{m_{\hat{a}_t}(t)}{m_u(t)+m_{\hat{a}_t}(t)}}(\hat{\nu}_{\hat{a}_t}(t), \hat{\nu}_u(t)),$$

where $m_u(t) = \sum_v \hat{G}_{v,u}(t) \frac{N_v(t)}{t} = \sum_v \frac{N_{v,u}(t)}{N_v(t)} \frac{N_v(t)}{t} = M_u(t)/t$. Using this latter fact, and noting that $\frac{m_{\hat{a}_t}(t)}{m_u(t)+m_{\hat{a}_t}(t)} = \frac{M_{\hat{a}_t}(t)}{M_u(t)+M_{\hat{a}_t}(t)}$, we get

$$tT(N_t/t; \hat{\nu}(t))^{-1} = \min_{u \neq \hat{a}_t} (M_u(t) + M_{\hat{a}_t}(t)) I_{\frac{M_{\hat{a}_t}(t)}{M_u(t)+M_{\hat{a}_t}(t)}}(\hat{\nu}_{\hat{a}_t}(t), \hat{\nu}_u(t)) = \Lambda(t).$$

We can now provide the proof of the stopping rule.

Proof of Proposition 6. First note that the event $\{\hat{a}_\tau \neq a^\star(\mu)\} \subset \{\nu \in \mathrm{Alt}(\hat{\nu}(\tau))\}$. From the discussion above, using the notation $\Lambda(t) = tT(N_t/t; \hat{\nu}(t))^{-1}$, we observe that under $\{\nu \in \mathrm{Alt}(\hat{\nu}(\tau))\}$ then the following inequalities hold

$$\begin{aligned}\mathbb{P}_\nu(\tau < \infty, \hat{a}_\tau \neq a^\star(\mu)) &\leq \mathbb{P}_\nu(\exists t \in \mathbb{N} : \hat{a}_t \neq a^\star(\mu), tT(N_t/t; \hat{\nu}(t))^{-1} > \beta(t, \delta)), \\ &\leq \mathbb{P}_\nu\left(\exists t \in \mathbb{N} : \hat{a}_t \neq a^\star(\mu), \min_{u \neq \hat{a}_t} \inf_{\lambda : \lambda_u \geq \lambda_{\hat{a}_t}} \sum_{v \in V} M_v(t) \mathrm{KL}(\hat{\nu}_v(t), \lambda_v) > \beta(t, \delta)\right), \\ &\leq \mathbb{P}_\nu\left(\exists t \in \mathbb{N}, \exists u \neq \hat{a}_t : \hat{a}_t \neq a^\star(\mu), \inf_{\lambda : \lambda_u \geq \lambda_{\hat{a}_t}} \sum_{v \in V} M_v(t) \mathrm{KL}(\hat{\nu}_v(t), \lambda_v) > \beta(t, \delta)\right), \\ &\leq \mathbb{P}_\nu\left(\exists t \in \mathbb{N}, \exists u \neq \hat{a}_t : \sum_{v \in \{u, \hat{a}_t\}} M_v(t) \mathrm{KL}(\hat{\nu}_v(t), \nu_v) > \beta(t, \delta)\right).\end{aligned}$$

Now, from (Kaufmann and Koolen, 2021, Theorem 7), we know that

$$\mathbb{P}_\nu\left(\exists t \in \mathbb{N}, \exists u \neq \hat{a}_t : \sum_{v \in \{u, \hat{a}_t\}} M_v(t) \mathrm{KL}(\hat{\nu}_v(t), \nu_v) > 2\mathcal{C}_{\exp}\left(\frac{\ln\left(\frac{K-1}{\delta}\right)}{2}\right) + 3 \sum_{v \in \{\hat{a}_t, u\}} \ln(1 + \ln(M_v(t)))\right) \leq \delta.$$

where we applied (Kaufmann and Koolen, 2021, Theorem 7) over $K-1$ subsets $\{\underbrace{(u, \hat{a}_t)}_{\mathcal{S}_u}\}_{u \neq \hat{a}_t}$ of size 2 each, and

took a union bound. Finally, using Jensen's inequality we also have that

$$\begin{aligned}\ln(1+\ln(M_v(t)))+\ln(1+\ln(M_u(t))) &\le 2\ln\left(\frac{1+\ln(M_v(t))}{2}+\frac{1+\ln(M_u(t))}{2}\right),\\ &= 2\ln\left(1+\frac{\ln(M_v(t))+\ln(M_u(t))}{2}\right),\\ &\le 2\ln\left(1+\ln\left(\frac{M_v(t)+M_u(t)}{2}\right)\right).\end{aligned}$$

Note that $M_v(t)+M_u(t)$ cannot exceed $2t$ (just consider a full feedback graph where $G_{u,v}=1$ so that $M_v(t)=t$ for every v). Hence, this implies that the GLR statistics is δ-PC with the threshold

$$\beta(t,\delta)=2\mathcal{C}_{\exp}\left(\frac{\ln\left(\frac{K-1}{\delta}\right)}{2}\right)+6\ln(1+\ln(t)).$$

□

Definition of $\mathcal{C}_{\exp}(x)$. Last, but not least, we briefly explain the definition of $\mathcal{C}_{\exp}(x)$. We define $\mathcal{C}_{\exp}(x)$ as (Kaufmann and Koolen, 2021, Theorem 7) $\mathcal{C}_{\exp}(x) \coloneqq 2\tilde{h}_{3/2}\left(\frac{h^{-1}(1+x)+\ln(2\zeta(2))}{2}\right)$, where: $\zeta(s)=\sum_{n\ge1} n^{-s}$; $h(u)=u-\ln(u)$ for $u\ge1$; lastly, for for any $z\in[1,e]$ and $x\ge0$:

$$\tilde{h}_z(x)=\begin{cases}h^{-1}(x)e^{1/h^{-1}(x)} & \text{if } x\ge h(1/\ln z),\\ z(x-\ln\ln z) & \text{otherwise.}\end{cases}$$

D.3 Sample Complexity Analysis

The following sample complexity analysis follows the analysis of Garivier and Kaufmann (2016) while adopting necessary changes for our problem setup. We first prove part (1) and (2) of theorem 3, and prove part (3) of theorem 3 separately. In the end of this section, we provide the proof for corollary 1.

Proof of part (1) and (2) of theorem 3: Let $\mathcal{E}$ be the event:

$$\mathcal{E}=\left\{\inf_{w\in C^\star(\nu)}\left\|\frac{N(t)}{t}-\omega\right\|_\infty \xrightarrow{t\to\infty} 0, \hat{\nu}(t)\xrightarrow{t\to\infty}\nu\right\}.$$

By Proposition 5 and the law of large number, we have $\mathcal{E}$ holds with probability 1. On $\mathcal{E}$, with the continuity property of function $T(\omega,\nu)^{-1}$ at $(\omega^\star(\nu),\nu)$ and proposition 5, for every $\omega^\star(\nu)\in C^\star(\nu)$, we have for all $\epsilon>0$ there exists $t_0\in\mathbb{N}$ such that for all $t\ge t_0$:

$$T(N(t)/t;\hat{\nu}(t))^{-1}\ge\frac{1}{1+\epsilon}T^\star(\nu)^{-1}.$$

Therefore, for $t\ge t_0$:

$$L(t)=tT(N(t)/t;\hat{\nu}(t))^{-1}\ge\frac{t}{1+\epsilon}T^\star(\nu)^{-1}.$$

Hence,

$$\begin{aligned}\tau &= \inf\left\{t\in\mathbb{N}: L(t)\ge\beta(t,\delta)\right\},\\ &\le t_0\vee\inf\left\{t\in\mathbb{N}:\frac{t}{1+\epsilon}T^\star(\nu)^{-1}\ge\beta(t,\delta)\right\}.\end{aligned}$$

Recall that $\beta(t,\delta)\coloneqq 2\mathcal{C}_{\exp}\left(\frac{\ln\left(\frac{K-1}{\delta}\right)}{2}\right)+6\ln(1+\ln(t))$. Note that there exists a universal constant B such that $\beta(t,\delta)\le\ln(Bt/\delta)$. Hence,

$$\tau\le t_0\vee\inf\left\{t\in\mathbb{N}:\frac{t}{1+\epsilon}T^\star(\nu)^{-1}\ge\ln(Bt/\delta)\right\}.$$

Applying Lemma 18 of Garivier and Kaufmann (2016) by letting $\alpha = 1$:

$$\tau \leq t_0 \vee (1+\epsilon)T^\star(\nu)\left[\ln\left(\frac{Be(1+\epsilon)T^\star(\nu)}{\delta}\right) + \ln\ln\left(\frac{B(1+\epsilon)T^\star(\nu)}{\delta}\right)\right].$$

Thus, τ is finite with probability 1. And

$$\limsup_{\delta\to 0}\frac{\tau}{\ln(1/\delta)} \leq (1+\epsilon)T^*(\nu).$$

We conclude the proof of part (2) by letting $\epsilon \to 0$. □

Proof of part (3) of theorem 3: Let $T \in \mathbb{N}$, for $\epsilon > 0$, define $\mathcal{E}_T := \bigcap_{t=T^{\frac{1}{4}}}^{T}(\hat{\nu}(t) \in \mathcal{I}_\epsilon)$, where $\mathcal{I}_\epsilon := \{\nu' : \|\nu'-\nu\|_\infty \leq \epsilon\}$ and $\|\nu'-\nu\|_\infty := \max\{\|G'-G\|_\infty, \|\mu'-\mu\|_\infty\}$. Following the same argument as Lemma 19 of Garivier and Kaufmann (2016), one can show that there exist two constant B and C (that depend on ν and ϵ) such that $\mathbb{P}_\nu(\mathcal{E}_T^c) \leq BT\exp(-CT^{\frac{1}{8}})$. Denote

$$C_\epsilon^\star(\nu) = \inf_{\substack{\omega':\|\omega'-\mathrm{Proj}_{C^\star(\nu)}(\omega')\|_\infty \leq 5(K-1)\epsilon \\ \nu':\|\nu'-\nu\|_\infty\leq\epsilon}} T(\omega',\nu')^{-1}.$$

By proposition 9, for any ϵ, there exists $T(\epsilon)$ such that for any $T \geq T(\epsilon)$ and $t \geq \sqrt{T}$, $\|N(t)/t - \mathrm{Proj}_{C^\star(\nu)}(N(t)/t)\|_\infty \leq 5(K-1)\epsilon$. With this fact, on $\mathcal{E}_T$, for $T \geq T(\epsilon)$ and $t \geq \sqrt{T}$, one has:

$$L(t) = tT(N(t)/t;\hat{\nu}_t)^{-1} \geq tC_\epsilon^\star(\nu).$$

Therefore, let $T \geq T(\epsilon)$, on $\mathcal{E}_T$,

$$\begin{aligned}
\min\{\tau_\delta, T\} &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T}\mathbf{1}_{(\tau_\delta > t)}, \\
&= \sqrt{T} + \sum_{t=\sqrt{T}}^{T}\mathbf{1}_{(L(t)\leq\beta(t,\delta))}, \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T}\mathbf{1}_{(tC_\epsilon^\star(\nu)\leq\beta(t,\delta))}, \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T}\mathbf{1}_{(tC_\epsilon^\star(\nu)\leq\beta(T,\delta))}, \\
&\leq \sqrt{T} + \frac{\beta(T,\delta)}{C_\epsilon^\star(\nu)}.
\end{aligned}$$

Denote $T_0 = \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{\beta(T,\delta)}{C_\epsilon^\star(\nu)} \leq T\right\}$. Thus, one has for $T \geq \max\{T_0, T(\epsilon)\}$, $\mathcal{E}_T \subseteq (\tau_\delta \leq T)$. Hence,

$$\begin{aligned}
\mathbb{E}[\tau_\delta] &\leq \max\{T_0, T(\epsilon)\} + \sum_{T=\max\{T_0,T_\epsilon\}}^{\infty}\mathbb{P}(\tau_\delta > T), \\
&\leq T_0 + T(\epsilon) + \sum_{T=\max\{T_0,T_\epsilon\}}^{\infty} BT\exp(-CT^{\frac{1}{8}}).
\end{aligned}$$

We then upper bound T_0. By introducing a constant $C(\eta) = \inf\left\{T \in \mathbb{N} : T - \sqrt{T} \geq \frac{T}{1+\eta}\right\}$, one has

$$\begin{aligned}
T_0 &\leq C(\eta) + \inf\left\{T \in \mathbb{N} : \frac{\beta(T,\delta)}{C_\epsilon^\star(\nu)} \leq \frac{T}{1+\eta}\right\}, \\
&\leq C(\eta) + \inf\left\{T \in \mathbb{N} : \frac{C_\epsilon^\star(\nu)T}{1+\eta} \geq \ln(BT/\delta)\right\}.
\end{aligned}$$

Applying Lemma 18 of Garivier and Kaufmann (2016) again:

$$T_0(\delta) \leq C(\eta) + (1+\eta)C^\star_\epsilon(\nu)^{-1}\left[\ln\left(\frac{Be(1+\eta)}{C^\star_\epsilon(\nu)\delta}\right) + \ln\ln\left(\frac{B(1+\eta)}{C^\star_\epsilon(\nu)\delta}\right)\right].$$

Therefore,

$$\limsup_{\delta\to 0}\frac{\mathbb{E}[\tau_\delta]}{\ln(1/\delta)} \leq \frac{(1+\eta)}{C^\star_\epsilon(\nu)}.$$

From the continuity property of function $T(\omega,\nu)^{-1}$ at $(\omega^\star(\nu),\nu)$ for each $\omega^\star(\nu)$ in $C^\star(\nu)$, one has

$$\lim_{\epsilon\to 0} C^\star_\epsilon(\nu) = T^\star(\nu)^{-1},$$

Letting η go to 0:

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

□

Proof of Corollary 1. The proof relies on the following proposition, which is a direct application of Proposition 9.

Proposition 10. *Let $S_t = \{u \in V : N_u(t) < \sqrt{t} - K/2\}$. The D-tracking rule, defined as*

$$V_t \in \begin{cases} \arg\min_{u\in S_t} N_u(t) & S_t \neq \emptyset \\ \arg\min_{u\in V} N_u(t) - t\sum_{n=1}^t \alpha_{t,n}\omega_{\text{heur}}(n) & \text{otherwise} \end{cases}, \tag{15}$$

where, for every $t \geq 1$, the sequence $\alpha_t = (\alpha_{t,n})_{n=1}^t$ satisfies: (1) $\alpha_{t,n} \in [0,1]$; (2) for every fixed $n \in \{1,\dots,t\}$ we have $\alpha_{t,n} = o(1)$ in t ; (3) for all t, $\sum_{n=1}^t \alpha_{t,n} = 1$.

Such tracking rule ensures that for all $\epsilon > 0$ there exists $t(\epsilon)$ such that for all $t \geq t(\epsilon)$ we have

$$\|N(t)/t - \omega_{\text{heur}}\|_\infty \leq 5(K-1)\epsilon,$$

and $\lim_{t\to\infty}\|N(t)/t - \omega_{\text{heur}}\|_\infty \to 0$ almost surely.

Proof. Proposition 10 can be proved using the same analysis as in Proposition 9, except that one needs to replace $C^\star(\nu)$ with $\{\omega_{\text{heur}}\}$. We thus skip the full proof. □

Corollary 1 can then be proved using the same analysis as in Theorem 3 combined with Proposition 10. □

We also provide an almost-surely lower bound for this heuristic algorithm, as established in the following Proposition 11.

Proposition 11. *TaS-FG with $\omega^\star(t) = \omega_{\text{heur}}(t)$ satisfies that $\mathbb{P}_\nu\left(\liminf_{\delta\to 0}\frac{\tau}{\ln(1/\delta)} \geq T^\star(\nu)\right) = 1$.*

Proof. By the continuity property of function $T(\omega,\nu)^{-1}$ at $(\omega_{\text{heur}},\nu)$ and Proposition 10, for any $\epsilon > 0$, with probability 1, there exists t_1 such that for any $t \geq t_1$,

$$L(t) = tT(N(t)/t;\hat\nu(t))^{-1} \leq \frac{t}{1-\epsilon}T(\omega_{\text{heur}},\nu)^{-1} \leq \frac{t}{1-\epsilon}T^\star(\nu)^{-1}.$$

By the definition of the stopping time,

$$\begin{aligned}\tau &= \inf\{t\in\mathbb{N} : L(t) \geq \beta(t,\delta)\},\\ &\geq \inf\{t\in\mathbb{N} : L(t) \geq \ln(1/\delta)\}.\end{aligned}$$

When $\delta\to 0$, $\inf\{t\in\mathbb{N} : L(t) \geq \ln(1/\delta)\} > t_1$. Therefore,

$$\begin{aligned}\liminf_{\delta\to 0}\frac{\tau}{\ln(1/\delta)} &\geq \liminf_{\delta\to 0}\frac{\inf\left\{t\in\mathbb{N} : tT^\star(\nu)^{-1} \geq (1-\epsilon)\ln(1/\delta)\right\}}{\ln(1/\delta)},\\ &\geq T^\star(\nu)(1-\epsilon).\end{aligned}$$

Letting ϵ go to 0 concludes the proof. □