POLITECNICO DI TORINO
Repository ISTITUZIONALE

On the Minimax Regret for Online Learning with Feedback Graphs

Original
On the Minimax Regret for Online Learning with Feedback Graphs / Eldowa, K., Esposito, E., Cesari, T., Cesa-Bianchi, N.. - 36:(2023), pp. 46122-46133. (NeurIPS 2023, the Thirty-seventh Annual Conference on Neural Information Processing Systems New Orleans (USA) Dec 10 through Dec 16, 2023).

Availability:
This version is available at: 11583/2990221 since: 2024-07-02T10:33:22Z

Publisher:
Curran Associates

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

08 November 2026

On the Minimax Regret for Online Learning with Feedback Graphs

Khaled Eldowa[*]
Università degli Studi di Milano, Milan, Italy
khaled.eldowa@unimi.it

Emmanuel Esposito[*]
Università degli Studi di Milano, Milan, Italy
& Istituto Italiano di Tecnologia, Genoa, Italy
emmanuel@emmanuelesposito.it

Tommaso Cesari
University of Ottawa, Ottawa, Canada
tcesari@uottawa.ca

Nicolò Cesa-Bianchi
Università degli Studi di Milano, Milan, Italy
& Politecnico di Milano, Milan, Italy
nicolo.cesa-bianchi@unimi.it

Abstract

In this work, we improve on the upper and lower bounds for the regret of online learning with strongly observable undirected feedback graphs. The best known upper bound for this problem is $\mathcal{O}\big(\sqrt{\alpha T \ln K}\big)$, where K is the number of actions, α is the independence number of the graph, and T is the time horizon. The $\sqrt{\ln K}$ factor is known to be necessary when $\alpha = 1$ (the experts case). On the other hand, when $\alpha = K$ (the bandits case), the minimax rate is known to be $\Theta\big(\sqrt{KT}\big)$, and a lower bound $\Omega\big(\sqrt{\alpha T}\big)$ is known to hold for any α. Our improved upper bound $\mathcal{O}\big(\sqrt{\alpha T(1+\ln(K/\alpha))}\big)$ holds for any α and matches the lower bounds for bandits and experts, while interpolating intermediate cases. To prove this result, we use FTRL with q-Tsallis entropy for a carefully chosen value of $q \in [1/2, 1)$ that varies with α. The analysis of this algorithm requires a new bound on the variance term in the regret. We also show how to extend our techniques to time-varying graphs, without requiring prior knowledge of their independence numbers. Our upper bound is complemented by an improved $\Omega\big(\sqrt{\alpha T(\ln K)/(\ln \alpha)}\big)$ lower bound for all $\alpha > 1$, whose analysis relies on a novel reduction to multitask learning. This shows that a logarithmic factor is necessary as soon as $\alpha < K$.

1 Introduction

Feedback graphs [29] provide an elegant interpolation between two popular online learning models: multiarmed bandits and prediction with expert advice. When learning with an undirected feedback graph G over K actions, the online algorithm observes not only the loss of the action chosen in each round, but also the loss of the actions that are adjacent to it in the graph. Two important special cases of this setting are: prediction with expert advice (when G is a clique) and K-armed bandits (when G has no edges). When losses are generated adversarially, the regret in the feedback graph setting with strong observability has been shown to scale with the independence number α of G. Intuitively, denser graphs, which correspond to smaller independence numbers, provide more feedback to the learner, thus enabling a better control on regret. More specifically, the best known upper and lower bounds on the regret after T rounds are $\mathcal{O}\big(\sqrt{\alpha T \log K}\big)$ and $\Omega\big(\sqrt{\alpha T}\big)$ [3, 4]. It has been known for three decades that this upper bound is tight for $\alpha = 1$ (the experts case, [9, 10]). When $\alpha = K$

[*]Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

(the bandits case), the lower bound $\Omega\big(\sqrt{KT}\big)$—which has also been known for nearly three decades [7, 8]—was matched by a corresponding upper bound $\mathcal{O}\big(\sqrt{KT}\big)$ only in 2009 [5]. These results show that in feedback graphs, the logarithmic factor $\sqrt{\log K}$ is necessary (at least) for the $\alpha = 1$ case, while it must vanish from the minimax regret as α grows from 1 to K, but the current bounds fail to capture this fact. In this work, we prove new upper and lower regret bounds that for the first time account for this vanishing logarithmic factor.

To prove our new upper bound, we use the standard FTRL algorithm run with the q-Tsallis entropy regularizer (q-FTRL for short). It is well-known [1] that for $q = \frac{1}{2}$ this algorithm (run with appropriate loss estimates) achieves regret $\mathcal{O}\big(\sqrt{KT}\big)$ when $\alpha = K$ (bandits case), while for $q \to 1^-$ the same algorithm (without loss estimates) recovers the bound $\mathcal{O}\big(\sqrt{T \log K}\big)$ when $\alpha = 1$ (experts case). When G contains all self-loops, we show in Theorem 1 that, if q is chosen as a certain function $q(\alpha, K)$, then $q(\alpha, K)$-FTRL, run with standard importance-weighted loss estimates, achieves regret $\mathcal{O}\big(\sqrt{\alpha T(1 + \log(K/\alpha))}\big)$. This is a strict improvement over the previous bound, and matches the lower bounds for bandits and experts while interpolating the intermediate cases. This interpolation is reflected by our choice of q, which goes from $\frac{1}{2}$ to 1 as α ranges from 1 to K. The main technical hurdle in proving this result is an extension to arbitrary values of $q \in \big[\frac{1}{2}, 1\big)$ of a standard result—see, e.g., [29, Lemma 3]—that bounds in terms of α the variance term in the regret of q-FTRL. In Theorem 2, using a modified loss estimate, this result is extended to any strongly observable undirected graph [2], a class of feedback graphs in which some of the actions do not reveal their loss when played. In Theorem 3, we show via a doubling trick that our new upper bound can also be obtained (up to constant factors) without the need of knowing (or computing) α. As the resulting algorithm is oblivious to α, our analysis also applies to arbitrary sequences of graphs G_t, where K is constant but the independence number α_t of G_t can change over time, and the algorithm observes G_t only after choosing an action (the so-called uninformed case). In this setting, the analysis of the doubling trick is complicated by the non-trivial dependence of the regret on the sequence of α_t.

We also improve on the $\Omega\big(\sqrt{\alpha T}\big)$ lower bound by proving a new $\Omega\big(\sqrt{\alpha T \log_\alpha K}\big)$ lower bound for all $\alpha > 1$. This is the first result showing the necessity—outside the experts case—of a logarithmic factor in the minimax regret for all $\alpha < K$. Our proof uses a stochastic adversary generating both losses and feedback graphs via i.i.d. draws from a joint distribution. This sequence of losses and feedback graphs can be used to define a hard instance of the multi-task bandits problem, a variant of the combinatorial bandits framework [12]. We then prove our result by adapting known lower bounding techniques for multi-task bandits [6]. Note that for values of α bounded away from 2 and K, the logarithmic factor $\log_\alpha K$ in the lower bound is smaller than the corresponding factor $1 + \log(K/\alpha)$ in the upper bound. Closing this gap remains an open problem.

1.1 Additional related work

Several previous works have used the q-Tsallis regularizer with q tuned to specific values other than $\frac{1}{2}$ and 1. For example, in [35, Section 4], q is chosen as a function of K to prove a regret bound of $\mathcal{O}\big(\sqrt{\alpha T(\log K)^3}\big)$ for any strongly observable directed feedback graph, which shaves off a $\log T$ factor compared to previous works. This bound is worse than the corresponding bounds for undirected graphs because the directed setting is harder. Specific choices of q have been considered to improve the regret in settings of online learning with standard bandit feedback. For example, the choice $q = \frac{2}{3}$ was used in [31] to improve the analysis of regret in bandits with decoupled exploration and exploitation. Regret bounds for arbitrary choices of q are derived in [36, 23] for a best-of-both-worlds analysis of bandits, though $q = \frac{1}{2}$ remains the optimal choice. The $\frac{1}{2}$-Tsallis entropy and the Shannon entropy ($q = 1$) regularizers have been combined before in different ways to obtain best-of-both-worlds guarantees for the graph feedback problem [18, 22]. The idea of using values of $q \in (\frac{1}{2}, 1)$ for feedback graphs is quite natural and has been brought up before, e.g., in [32], but achieving an improved dependence on the graph structure by picking a suitable value of q has not been, to the best of our knowledge, successfully pursued before. On the other hand, an approach based on a similar use of the q-Tsallis regularizer has been employed by [26] for the problem of multiarmed bandits with sparse losses to achieve a $\mathcal{O}\big(\sqrt{sT \ln(K/s)}\big)$ regret bound, where s is the maximum number of nonzero losses at any round.

Our lower bound is reminiscent of the $\Omega\big(\sqrt{KT \log_K N}\big)$ lower bound proved in [33] for the problem of bandits with expert advice (with $N \geq K$ being the number of experts); see also [17] and [34]. In

that problem, at each time step, experts suggest distributions over actions to the learner, whose regret is computed against the best expert in hindsight. Although the two settings are different, the variant of the multitask bandit problem that our lower bound construction simulates is the same as the one used in the proof of [17, Theorem 7].

2 Problem Setting

For any integer $n \geq 1$, let $[n] = \{1, \dots, n\}$. We consider the following game played over T rounds between a learner with action set $V = [K]$ and the environment. At the beginning of the game, the environment secretly selects a sequence of losses $(\ell_t)_{t\in[T]}$, where $\ell_t \colon V \to [0,1]$,[2] and a sequence of undirected graphs $(G_t)_{t\in[T]}$ over the set of actions V, that is, $G_t = (V, E_t)$. At any time t, the learner selects an arm I_t (possibly at random), then pays loss $\ell_t(I_t)$ and observes the feedback graph G_t and all losses $\ell_t(i)$ of neighbouring actions $i \in N_{G_t}(I_t)$, where $N_{G_t}(i) = \{j \in V \,:\, (i,j) \in E_t\}$ (see Online Protocol 1). In this work, we only focus on strongly observable graphs [2]. An undirected graph G is strongly observable if for every $i \in V$, at least one of the following holds: $i \in N_G(i)$ or $i \in N_G(j)$ for all $j \neq i$.

The performance of the learner is measured by the regret

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t(i) \ .$$

where the expectation is over the learner's internal randomization.

Online Protocol 1 Online learning with feedback graphs

environment: (hidden) losses $\ell_t \colon V \to [0,1]$ and graphs $G_t = (V, E_t)$, for all $t = 1, \dots, T$
for $t = 1, \dots, T$ **do**
  The learner picks an action $I_t \in V$ (possibly at random)
  The learner incurs loss $\ell_t(I_t)$
  The learner observes losses $\big\{\big(i, \ell_t(i)\big) : i \in N_{G_t}(I_t)\big\}$ and graph G_t
end for

We denote by Δ_K the simplex $\big\{p \in [0,1]^K \,:\, \|p\|_1 = 1\big\}$. For any graph G, we define its independence number as the cardinality of the largest set of nodes such that no two nodes are neighbors, and denote it by $\alpha(G)$. For simplicity, we use N_t to denote the neighbourhood N_{G_t} in the graph G_t and we use α_t to denote the independence number $\alpha(G_t)$ of G_t at time t.

3 FTRL with Tsallis Entropy for Undirected Feedback Graphs

As a building block, in this section, we focus on the case when all the feedback graphs $G_1, \dots, G_T$ have the same independence number $\alpha_1 = \cdots = \alpha_T = \alpha$, whereas the general case is treated in the next section. For simplicity, we start with the assumption that all nodes have self-loops: $(i,i) \in E_t$ for all $i \in V$ and all t. We later lift this requirement and show that the regret guarantees that we provide can be extended to general strongly observable undirected feedback graphs, only at the cost of a constant multiplicative factor.

The algorithm we analyze is q-FTRL (described in Algorithm 1), which is an instance of the follow the regularized leader (FTRL) framework—see, e.g., [30, Chapter 7]—with the (negative) q-Tsallis entropy

$$\psi_q(x) = \frac{1}{1-q}\left(1 - \sum_{i\in V} x(i)^q\right) \qquad \forall x \in \Delta_K \ ,$$

as the regularizer, whose parameter $q \in (0,1)$ can be tuned according to our needs. Since we do not observe all the losses in a given round, the algorithm makes use of unbiased estimates for the losses. When all self-loops are present, we define the estimated losses in the following standard manner. Let

[2] For notational convenience, we will sometimes treat the loss functions $\ell_t \colon V \to [0,1]$ as vectors with components $\ell_t(1), \dots, \ell_t(K)$.

I_t be the action picked at round t, which is drawn from the distribution $p_t \in \Delta_K$ maintained by the algorithm, the loss estimate for an action $i \in V$ at round t is given by

$$\widehat{\ell}_t(i) = \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\{I_t \in N_t(i)\} \ , \tag{1}$$

where $P_t(i) = \mathbb{P}\big(I_t \in N_t(i)\big) = \sum_{j \in N_t(i)} p_t(j)$. This estimate is unbiased in the sense that $\mathbb{E}_t\big[\widehat{\ell}_t(i)\big] = \ell_t(i)$ for all $t \in [T]$ and all $i \in V$, where we denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \,|\, I_1, \ldots, I_{t-1}]$.

Algorithm 1 q-FTRL for undirected feedback graphs

input: $q \in (0,1)$, $\eta > 0$
initialization: $p_1(i) \leftarrow 1/K$ for all $i = 1, \ldots, K$
for $t = 1, \ldots, T$ **do**
  Select action $I_t \sim p_t$ and incur loss $\ell_t(I_t)$
  Observe losses $\big\{\big(i, \ell_t(i)\big) : i \in N_t(I_t)\big\}$ and graph G_t
  Construct a loss estimate $\widehat{\ell}_t$ for ℓ_t ▷ e.g., (1) or (6)
  Let $p_{t+1} \leftarrow \arg\min_{p \in \Delta_K} \eta \big\langle \sum_{s=1}^{t} \widehat{\ell}_s, p \big\rangle + \psi_q(p)$
end for

A key part of the standard regret analysis of q-FTRL (see, e.g., the proof of Lemma 3 in Appendix A) is handling the variance term, which, with the choice of estimator given in (1), takes the following form

$$B_t(q) = \sum_{i \in V} \frac{p_t(i)^{2-q}}{P_t(i)} \ . \tag{2}$$

By Hölder's inequality, this term can be immediately upper bounded by

$$B_t(q) \le \sum_{i \in V} p_t(i)^{1-q} \le \left(\sum_{i \in V} p_t(i)\right)^{1-q} \left(\sum_{i \in V} 1^{1/q}\right)^{q} = K^q \ ,$$

while previous results on the regret analysis of multiarmed bandits with graph feedback [29, 3] would give

$$B_t(q) \le \sum_{i \in V} \frac{p_t(i)}{P_t(i)} \le \alpha \ .$$

However, the former result would only recover a $\mathcal{O}(\sqrt{KT})$ regret bound (regardless of α) with the best choice of $q = 1/2$, which could be trivially achieved by ignoring side-observations of the losses, whereas the latter bound would only manage to achieve a $\mathcal{O}(\sqrt{\alpha T \ln K})$ regret bound, incurring the extra $\sqrt{\ln K}$ factor for all values of α. Other results in the literature (e.g., see [2, 4, 16, 19, 22, 25, 32, 35]) do not bring an improvement in this setting when bounding the $B_t(q)$ term and, hence, do not suffice for achieving the desired regret bound. The following lemma provides a novel and improved bound on quantities of the same form as $B_t(q)$ in terms of the independence number $\alpha_t = \alpha$ of the undirected graph G_t.

Lemma 1. *Let $G = (V, E)$ be any undirected graph with $|V| = K$ vertices and independence number $\alpha(G) = \alpha$. Let $b \in [0,1]$, $p \in \Delta_K$ and consider any nonempty subset $U \subseteq \{v \in V : v \in N_G(v)\}$. Then,*

$$\sum_{v \in U} \frac{p(v)^{1+b}}{\sum_{u \in N_G(v)} p(u)} \le \alpha^{1-b} \ .$$

Proof. First of all, observe that we can restrict ourselves to the subgraph $G[U]$ induced by U, i.e., the graph $G[U] = (U, E \cap (U \times U))$. This is because the neighbourhoods in this graph are such that $N_{G[U]}(v) \subseteq N_G(v)$ for all $v \in U$, and its independence number is $\alpha(G[U]) \le \alpha(G)$. Hence, it suffices to prove the claimed inequality for any undirected graph $G = (V, E)$ with all self-loops, any $p \in [0,1]^K$ such that $\|p\|_1 \le 1$, and the choice $U = V$. We assume this in what follows without loss of generality.

For any subgraph $H \subseteq G$ with vertices $V(H) \subseteq V$, denote the quantity we want to upper bound by

$$Q(H) = \sum_{v \in V(H)} \frac{p(v)^{1+b}}{\sum_{u \in N_G(v)} p(u)} \ .$$

Our aim is thus to provide an upper bound to $Q(G)$.

Consider a greedy algorithm that incrementally constructs a subset of vertices in the following way: at each step, it selects a vertex v that maximizes $p(v)^b / \big(\sum_{u \in N_G(v)} p(u)\big)$, it adds v to the solution, and it removes v from G together with its neighbourhood $N_G(v)$. This step is iterated on the remaining graph until no vertex is left.

Let $S = \{v_1, \dots, v_s\} \subseteq V$ be the solution returned by the above greedy algorithm on G. Also let $G_1, \dots, G_{s+1}$ be the sequence of graphs induced by the operations of the algorithm, where $G_1 = G$ and G_{s+1} is the empty graph, and let $N_r(v) = N_{G_r}(v)$ for $v \in V(G_r)$. At every step $r \in [s]$ of the greedy algorithm, the contribution to $Q(G)$ of the removed vertices $N_r(v_r)$ amounts to

$$\begin{aligned} Q(G_r) - Q(G_{r+1}) = \sum_{v \in N_r(v_r)} \frac{p(v)^{1+b}}{\sum_{u \in N_1(v)} p(u)} &\le \sum_{v \in N_r(v_r)} p(v) \frac{p(v_r)^b}{\sum_{u \in N_1(v_r)} p(u)} \\ &\le \frac{\sum_{v \in N_1(v_r)} p(v)}{\sum_{u \in N_1(v_r)} p(u)} p(v_r)^b = p(v_r)^b \ , \end{aligned}$$

where the last inequality is due to the fact that $N_i(v) \subseteq N_j(v)$ for all $i \ge j$ and $v \in V_i$. Therefore, we can observe that

$$Q(G) = \sum_{r=1}^{s} \big(Q(G_r) - Q(G_{r+1})\big) \le \sum_{v \in S} p(v)^b \ .$$

The solution S is an independent set of G by construction. Consider now any independent set $A \subseteq V$ of G. We have that

$$\begin{aligned} \sum_{v \in A} p(v)^b &\le \max_{x \in \Delta_K} \sum_{v \in A} x(v)^b = |A| \max_{x \in \Delta_K} \sum_{v \in A} \frac{x(v)^b}{|A|} \\ &\le |A| \max_{x \in \Delta_K} \left(\frac{1}{|A|} \sum_{v \in A} x(v) \right)^b \le |A|^{1-b} \le \alpha^{1-b} \ , \end{aligned} \tag{3}$$

where the second inequality follows by Jensen's inequality and the fact that $b \in [0, 1]$. □

Observe that this upper bound is tight for general probability distributions $p \in \Delta_K$ over the vertices V of any strongly observable undirected graph G (containing at least one self-loop), as it is exactly achieved by the distribution $p^\star \in \Delta_K$ defined as $p^\star(i) = \frac{1}{|S|} \mathbb{I}\{i \in S\}$ for some maximum independent set $S \subseteq V$ of G. Using this lemma, the following theorem provides our improved upper bound under the simplifying assumptions we made thus far.

Theorem 1. *Let $G_1, \dots, G_T$ be a sequence of undirected feedback graphs, where each G_t contains all self-loops and has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}} \ ,$$

and loss estimates (1)*, then its regret satisfies $R_T \le 2\sqrt{e\alpha T\,(2 + \ln(K/\alpha))}$.*

Proof. One can verify that for any $i \in V$, the loss estimate $\widehat{\ell}_t(i)$ defined in (1) satisfies $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$. Hence, using also that $\mathbb{E}_t\big[\widehat{\ell}_t(i)\big] = \ell_t(i)$, Lemma 2 in Appendix A implies that

$$R_T \le \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \in V} \frac{p_t(i)^{2-q}}{P_t(i)}\right] \tag{4}$$

$$\le \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q} \alpha^q T \ , \tag{5}$$

where the second inequality follows by Lemma 1 with $b = 1 - q$ since all actions $i \in V$ are such that $i \in N_G(i)$. Our choices for q and η allow us to further upper bound the right-hand side of (5) by

$$\begin{aligned}\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}T} &= \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\alpha K)-\frac{1}{2}\sqrt{\ln\left(K/\alpha\right)^2+4}\right)\left(2+\sqrt{\ln\left(K/\alpha\right)^2+4}\right)} \\ &\le \sqrt{2e\alpha T\left(2+\sqrt{\ln\left(K/\alpha\right)^2+4}\right)} \le 2\sqrt{e\alpha T\sqrt{\ln\left(K/\alpha\right)^2+4}} \\ &\le 2\sqrt{e\alpha T\left(2+\ln(K/\alpha)\right)}\ . \end{aligned}$$

□

The regret bound achieved in the above theorem achieves the optimal regret bound for the experts setting (i.e., $\alpha = 1$) and the bandits setting (i.e., $\alpha = K$) simultaneously. Moreover, it interpolates the intermediate cases for α ranging between 1 and K, introducing the multiplicative logarithmic factor only for graphs with independence number strictly smaller than K. We remark that the chosen values of q and η do in fact minimize the right-hand side of (5). Note that we relied on the knowledge of α to tune the parameter q. This is undesirable in general. We will show how to lift this requirement in Section 4. The same comment applies to Theorem 2, below.

We now show how to achieve the improved regret bound of Theorem 1 in the case of strongly observable undirected feedback graphs where some self-loops may be missing; i.e., there may be actions $i \in V$ such that $i \notin N_G(i)$. Using the loss estimator defined in (1) may lead to a large variance term due to the presence of actions without self-loops. One approach to deal with this—see, e.g., [35] or [28]—is to suitably alter the loss estimates of these actions.

Define $S_t = \{i \in V : i \notin N_t(i)\}$ as the subset of actions without self-loops in the feedback graph G_t at each time step $t \in [T]$. The idea is that we need to carefully handle some action $i \in S_t$ only in the case when the probability $p_t(i)$ of choosing i at round t is sufficiently large, say, larger than $1/2$. Define the set of such actions as $J_t = \{i \in S_t : p_t(i) > 1/2\}$ and observe that $|J_t| \le 1$. Similarly to [35], define new loss estimates

$$\widehat{\ell}_t(i) = \begin{cases} \frac{\ell_t(i)}{P_t(i)}\mathbb{I}\left\{I_t \in N_t(i)\right\} & \text{if } i \in V \setminus J_t \\ \frac{\ell_t(i)-1}{P_t(i)}\mathbb{I}\left\{I_t \in N_t(i)\right\}+1 & \text{if } i \in J_t \end{cases} \tag{6}$$

for which it still holds that $\mathbb{E}_t\left[\widehat{\ell}_t\right] = \ell_t$ and that $\mathbb{E}_t\left[\widehat{\ell}_t(i)^2\right] \le 1/P_t(i)$ for all $i \notin J_t$. This change, along with the use of Lemma 1 for the actions in $V \setminus S_t$, suffices in order to prove the following regret bound (see Appendix B for the proof) when the feedback graphs do not necessarily contain self-loops for all actions.

Theorem 2. *Let $G_1, \ldots, G_T$ be a sequence of strongly observable undirected feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 1 is run with input*

$$q = \frac{1}{2}\left(1+\frac{\ln(K/\alpha)}{\sqrt{\ln(K/\alpha)^2+4}+2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}}\ ,$$

and loss estimates (6)*, then its regret satisfies $R_T \le 6\sqrt{e\alpha T\left(2+\ln(K/\alpha)\right)}$.*

4 Adapting to Arbitrary Sequences of Graphs

In the previous section, we assumed for simplicity that all the graphs have the same independence number. This independence number was then used to tune q, the parameter of the Tsallis entropy regularizer used by the algorithm. In this section, we show how to extend our approach to the case when the independence numbers of the graphs are neither the same nor known a-priori by the learner. Had these independence numbers been known a-priori, one approach is to set q as in Theorem 2 but using the average independence number

$$\overline{\alpha}_T = \frac{1}{T}\sum_{t=1}^{T}\alpha_t\ .$$

Doing so would allow us to achieve a $\mathcal{O}\Big(\sqrt{\sum_{t=1}^{T}\alpha_t(1+\ln(K/\bar{\alpha}_T))}\Big)$ regret bound. We now show that we can still recover a bound of the same order without prior knowledge of $\bar{\alpha}_T$. For round t and any fixed $q \in [0,1]$, define

$$H_t(q) = \sum_{i \in V \setminus S_t} \frac{p_t(i)^{2-q}}{P_t(i)} \ .$$

We know from Lemma 1 that $H_t(q) \le \alpha_t^q$. Thus, we can leverage these observations and use a doubling trick (similar in principle to [3]) to guess the value of $\bar{\alpha}_T$. This approach is outlined in Algorithm 2. Starting with $r = 0$ and $T_r = 1$, the idea is to instantiate Algorithm 1 at time-step T_r with q and η set as in Theorem 2 but with 2^r replacing the independence number. Then, at $t \ge T_r$, we increment r and restart Algorithm 1 only if

$$\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1},$$

since (again thanks to Lemma 1) the left-hand side of the above inequality is always upper bounded by $\bar{\alpha}_T$. The following theorem shows that this approach essentially enjoys the same regret bound of Theorem 2 up to an additive $\log_2 \bar{\alpha}_T$ term.

Algorithm 2 q-FTRL for an arbitrary sequence of strongly observable undirected graphs

input: Time horizon T
define: For each $r \in \{0, \dots, \lfloor \log_2 K \rfloor\}$,

$$q_r = \frac{1}{2}\left(1 + \frac{\ln(K/2^r)}{\sqrt{\ln(K/2^r)^2 + 4} + 2}\right) \qquad \text{and} \qquad \eta_r = \sqrt{\frac{2q_r K^{1-q_r}}{11T(1-q_r)\,(2^r)^{q_r}}}$$

initialization: $T_0 \leftarrow 1$, $r \leftarrow 0$, instantiate Algorithm 1 with $q = q_0$, $\eta = \eta_0$, and loss estimates (6)
for $t = 1, \dots, T$ **do**
  Perform one step of the current instance of Algorithm 1
  if $\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1}$ **then**
    $r \leftarrow r + 1$
    $T_r \leftarrow t + 1$
    Restart Algorithm 1 with $q = q_r$, $\eta = \eta_r$, and loss estimates (6)
  end if
end for

Theorem 3. *Let $C = 4\sqrt{6e}\,\frac{\sqrt{\pi}+\sqrt{4-2\ln 2}}{\ln 2}$. Then, the regret of Algorithm 2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T}\alpha_t\left(2 + \ln\left(\frac{K}{\bar{\alpha}_T}\right)\right)} + \log_2 \bar{\alpha}_T \ .$$

Proof sketch. For simplicity, we sketch here the proof for the case when in every round t, all the nodes have self-loops; hence, $H_t(q) = B_t(q)$. See the full proof in Appendix C, which treats the general case in a similar manner. Let $n = \lceil \log_2 \bar{\alpha}_T \rceil$ and assume without loss of generality that $\bar{\alpha}_T > 1$. Since Lemma 1 implies that for any r and t, $B_t(q_r) \le \alpha_t^{q_r}$, we have as a consequence that for any $t \ge T_r$,

$$\frac{1}{T}\sum_{s=T_r}^{t} B_s(q_r)^{1/q_r} \le \frac{1}{T}\sum_{s=T_r}^{t}\alpha_s \le \bar{\alpha}_T \le 2^n \ .$$

Hence, the maximum value of r that the algorithm can reach is $n-1$. In doing so, we will execute n instances of Algorithm 1, each corresponding to a value of $r \in \{0, \dots, n-1\}$. For every such r, we upper bound the instantaneous regret at step $T_{r+1} - 1$ (the step when the restarting condition is satisfied) by 1, hence the added $\log_2 \bar{\alpha}_T$ term in the regret bound. For the rest of the interval; namely, for $t \in [T_r, T_{r+1} - 2]$, we have via (4) that the regret of Algorithm 1 is bounded by

$$\frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\,\mathbb{E}\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r) \ . \tag{7}$$

Define $T_{r:r+1} = T_{r+1} - T_r - 1$, and notice that

$$\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r) \le T_{r:r+1}\left(\frac{1}{T_{r:r+1}} \sum_{t=T_r}^{T_{r+1}-2} B_t(q_r)^{1/q_r}\right)^{q_r}$$
$$\le T_{r:r+1}\left(\frac{T}{T_{r:r+1}} 2^{r+1}\right)^{q_r} \le 2T\big(2^r\big)^{q_r} \ ,$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0,1)$, and the second follows from the restarting condition of Algorithm 2. After, plugging this back into (7), we can simply use the definitions of η_r and q_r and bound the resulting expression in a similar manner to the proof of Theorem 1. Overall, we get that

$$R_T \le 4\sqrt{3eT} \sum_{r=0}^{n-1} \sqrt{2^r \ln\big(e^2 K 2^{-r}\big)} + \log_2 \bar{\alpha}_T \ ,$$

from which the theorem follows by using Lemma 4 in Appendix A, which shows, roughly speaking, that the sum on the right-hand side is of the same order as its last term. □

Although Algorithm 2 requires knowledge of the time horizon, this can be dealt with by applying a standard doubling trick on T at the cost of a larger constant factor. It is also noteworthy that the bound we obtained is of the form $\sqrt{T\bar{\alpha}_T(1+\ln(K/\bar{\alpha}_T))}$ and not $\sqrt{\sum_t \alpha_t(1+\ln(K/\alpha_t))}$. Although both coincide with the bound of Theorem 2 when α_t is the same for all time steps, the latter is smaller via the concavity of $x(1+\ln(K/x))$ in x. It is not clear, however, whether there is a tuning of $q \in (0,1)$ that can achieve the second bound (even with prior knowledge of the entire sequence $\alpha_1, \ldots, \alpha_T$ of independence numbers).

5 An Improved Lower Bound via Multitask Learning

In this section we provide a new lower bound on the minimax regret showing that, apart from the bandits case, a logarithmic factor is indeed necessary in general. When the graph is fixed over time, it is known that a lower bound of order $\sqrt{\alpha T}$ holds for any value of α [3, 29]. Whereas for the experts case ($\alpha = 1$), the minimax regret is of order[3] $\sqrt{T \ln K}$ [9]. The following theorem provides, for the first time, a lower bound that interpolates between the two aforementioned bounds for the intermediate values of α.

Theorem 4. *Pick any $K \ge 2$ and any α such that $2 \le \alpha \le K$. Then, for any algorithm and for all $T \ge \frac{\alpha \log_\alpha K}{4\log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1, \ldots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t = 1, \ldots, T$ and*

$$R_T \ge \frac{1}{18\sqrt{2}} \sqrt{\alpha T \log_\alpha K}.$$

In essence, the proof of this theorem (see Appendix D) constructs a sequence of feedback graphs and losses that is equivalent to a hard instance of the multitask bandit problem (MTB) [12], an important special case of combinatorial bandits with a convenient structure for proving lower bounds [6, 15, 21]. We consider a variant of MTB in which, at the beginning of each round, the decision-maker selects an arm to play in each one of M stochastic bandit games. Subsequently, the decision-maker only observes (and suffers) the loss of the arm played in a single randomly selected game. For proving the lower bound, we use a class of stationary stochastic adversaries (i.e., environments), each generating graphs and losses in a manner that simulates an MTB instance.

Fix $2 \le \alpha \le K = |V|$ and assume for simplicity that $M = \log_\alpha K$ is an integer. We now construct an instance of online learning with time-varying feedback graphs $G_t = (V, E_t)$ with $\alpha(G_t) = \alpha$ that is equivalent to an MTB instance with M bandit games each containing α "base actions". Since $K = \alpha^M$, we can uniquely identify each action in V with a vector $a = \big(a(1), \ldots, a(M)\big)$ in

[3] As a lower bound, this is known to hold asymptotically as K and T grow. However, it can also be shown to hold non-asymptotically (though with worse leading constants); see [20, Theorem 3.22] or [11, Theorem 3.6].

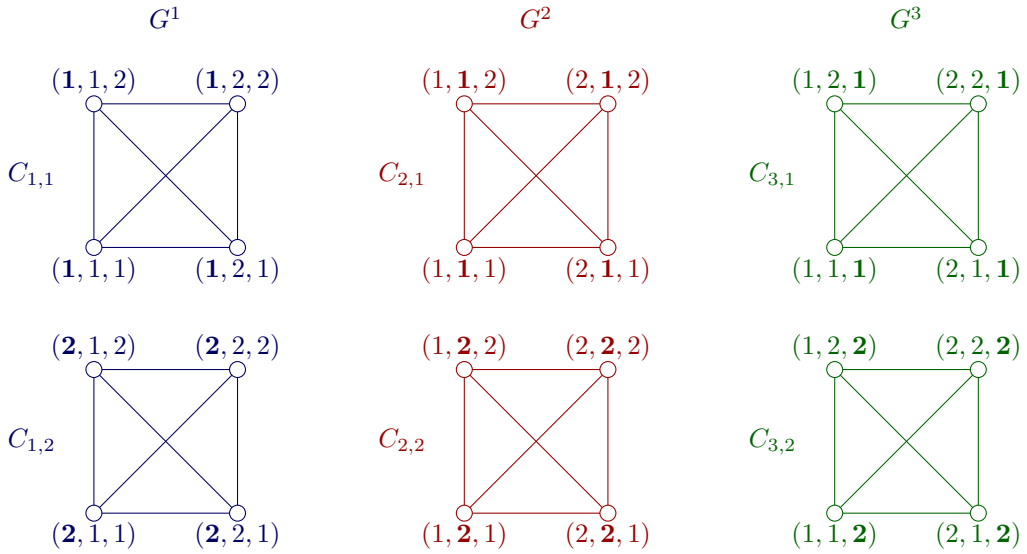

Figure 1: This figure shows an example of the multi-task bandit construction used to prove the lower bound. Here, $K = 8$ and $\alpha = 2$; thus, the number of games is $M = 3$. Each action is identified by a tuple of three numbers, each corresponding to a choice of one out of a pair of "base actions" in each game. Each of the three graphs in the figure corresponds to a game, such that two actions share an edge if and only if they choose the same base action in the corresponding game. At every round, a graph is randomly drawn, and all actions belonging to the same clique suffer the same loss.

$[\alpha]^M$. The action $a_t \in V$ chosen by the learner at round t is equivalent to a choice of base actions $a_t(1), \dots, a_t(M)$ in the M games. The feedback graph at every round is sampled uniformly at random from a set of M undirected graphs $\{G^i\}_{i=1}^M$, where $G^i = (V, E^i)$ is such that $(a, a') \in E^i$ if and only if $a(i) = a'(i)$. This means (see Figure 1) that each graph G^i consists of α isolated cliques $\{C_{i,j}\}_{j=1}^\alpha$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. Clearly, the independence number of any such graph is α. Drawing feedback graph $G_t = G^i$ corresponds to the activation of game i in the MTB instance. Hence, choosing $a_t \in V$ with feedback graph $G_t = G^i$ is equivalent to playing base action $a_t(i)$ in game i in the MTB. As for the losses, we enforce that, given a feedback graph G_t, all actions that belong to the same clique of the feedback graph are assigned the same loss. Namely, if $G_t = G^i$ and $a(i) = a'(i) = j$, then $\ell_t(a) = \ell_t(a')$, which can be seen as the loss $\ell_t(j)$ assigned to base action j in game G^i. To choose the distribution of the losses for the base actions, we apply the classic needle-in-a-haystack approach of [7] over the M games. More precisely, we construct a different environment for each action $a \in V$ in such a way that the distribution of the losses in each MTB game slightly favors (with a difference of a small $\varepsilon > 0$) the base action corresponding to a in that game. The proof then proceeds similarly to, for example, the proof of Theorem 5 in [6] or Theorem 7 in [17].

While both our upper and lower bounds achieve the desired goal of interpolating between the minimax rates of experts and bandits, the logarithmic factors in the two bounds are not exactly matching. In particular, if we compare $1 + \log_2(K/\alpha)$ and $\log_\alpha K$, we can see that although they coincide at $\alpha = 2$ and $\alpha = K$, the former is larger for intermediate values. It is reasonable to believe that the upper bound is of the correct order, seeing as it arose naturally as a result of choosing the best parameter for the Tsallis entropy regularizer, whereas achieving the extra logarithmic term in the lower bound required a somewhat contrived construction.

6 Conclusions and Future Work

In this work, we have shown that a proper tuning of the q-FTRL algorithm allows one to achieve a $\mathcal{O}\big(\sqrt{\alpha T(1 + \ln(K/\alpha))}\big)$ regret for the problem of online learning with undirected strongly observable feedback graphs. Our bound interpolates between the minimax regret rates of the bandits and the experts problems, the two extremes of the strongly observable graph feedback spectrum.

Furthermore, we have shown that an analogous bound can be achieved when the graphs vary over time, and without requiring any prior knowledge on the graphs. These results are complemented by our new lower bound of $\Omega\big(\sqrt{\alpha T(\ln K)/(\ln \alpha)}\big)$, which holds for $\alpha \geq 2$ and shows the necessity of a logarithmic factor in the minimax regret except for the bandits case. While our results provide the tightest characterization to date of the minimax rate for this setting, closing the small remaining gap (likely on the lower bound side) is an interesting problem. After the submission of this manuscript, a subsequent work [14] showed a lower bound for fixed feedback graphs composed of disjoint cliques that would imply worst-case optimality (up to constant factors) of our proposed algorithm for each pair of K and α—see Appendix E for a more detailed comparison with results therein. Extending our results to the case of directed strongly observable feedback graphs is a considerably harder task—see Appendix F for a preliminary discussion. Better understanding this more general setting is an interesting future direction.

Acknowledgements

KE, EE, and NCB gratefully acknowledge the financial support from the MUR PRIN grant 2022EKNE5K (Learning in Markets and Society), funded by the NextGenerationEU program within the PNRR scheme (M4C2, investment 1.1), the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence), and the EU Horizon CL4-2022-HUMAN-02 research and innovation action under grant agreement 101120237, project ELIAS (European Lighthouse of AI for Sustainability). TC gratefully acknowledges the support of the University of Ottawa through grant GR002837 (Start-Up Funds) and that of the Natural Sciences and Engineering Research Council of Canada (NSERC) through grants RGPIN-2023-03688 (Discovery Grants Program) and DGECR-2023-00208 (Discovery Grants Program, DGECR - Discovery Launch Supplement).

References

[1] J. D. Abernethy, C. Lee, and A. Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

[2] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 23–35. JMLR.org, 2015.

[3] N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

[4] N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 26, 2013.

[5] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

[6] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[9] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

[10] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. E. Schapire, and M. K. Warmuth. How to use expert advice. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 382–391, 1993.

[11] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] N. Cesa-Bianchi and G. Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[13] S.-H. Chang, P. C. Cosman, and L. B. Milstein. Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

[14] H. Chen, Y. He, and C. Zhang. On interpolating experts and multi-armed bandits. *arXiv preprint arXiv:2307.07264*, 2023.

[15] A. Cohen, T. Hazan, and T. Koren. Tight bounds for bandit combinatorial optimization. In *Conference on Learning Theory*, pages 629–642. PMLR, 2017.

[16] C. Dann, C. Wei, and J. Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *The Thirty Sixth Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR, 2023.

[17] K. Eldowa, N. Cesa-Bianchi, A. M. Metelli, and M. Restelli. Information-theoretic regret bounds for bandits with fixed expert advice. In *IEEE Information Theory Workshop*, pages 30–35. IEEE, 2023.

[18] L. Erez and T. Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34:28511–28521, 2021.

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

[20] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998.

[21] S. Ito, D. Hatano, H. Sumita, K. Takemura, T. Fukunaga, N. Kakimura, and K.-I. Kawarabayashi. Improved regret bounds for bandit combinatorial optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] S. Ito, T. Tsuchiya, and J. Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 35:28631–28643, 2022.

[23] T. Jin, J. Liu, and H. Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. *arXiv preprint arXiv:2302.13534*, 2023.

[24] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Proceedings of a Symposium on the Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[25] T. Kocák, G. Neu, M. Valko, and R. Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, volume 27, pages 613–621, 2014.

[26] J. Kwon and V. Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

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

[28] H. Luo, H. Tong, M. Zhang, and Y. Zhang. Improved high-probability regret for adversarial bandits with time-varying feedback graphs. In *International Conference on Algorithmic Learning Theory*, pages 1074–1100. PMLR, 2023.

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

[30] F. Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2019.

[31] C. Rouyer and Y. Seldin. Tsallis-inf for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

[32] C. Rouyer, D. van der Hoeven, N. Cesa-Bianchi, and Y. Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, 2022.

[33] Y. Seldin and G. Lugosi. A lower bound for multi-armed bandits with expert advice. In *13th European Workshop on Reinforcement Learning (EWRL)*, volume 2, page 7, 2016.

[34] N. M. Vural, H. Gokcesu, K. Gokcesu, and S. S. Kozat. Minimax optimal algorithms for adversarial bandit problem with multiple plays. *IEEE Transactions on Signal Processing*, 67(16):4383–4398, 2019.

[35] J. Zimmert and T. Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

[36] J. Zimmert and Y. Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

A Auxiliary Results

Lemma 2. *If Algorithm 1 is run with $q \in (0,1)$, learning rate $\eta > 0$, and non-negative loss estimates that satisfy $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$ for all $t = 1, \dots, T$, then its regret satisfies*

$$R_T \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{\ell}_t(i)^2\right] \ .$$

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^{T} \ell_t(i)$ be an action that minimizes the cumulative loss, and let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^*. Recall that for $t \in [T]$, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$, and notice that p_t is measurable with respect to the σ-algebra generated by $I_1, \dots, I_{t-1}$. Hence, using that

$$\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i \in V} p_t(i)\ell_t(i) \qquad \text{and} \qquad \mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t \ ,$$

we have, via the tower rule and the linearity of expectation, that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(i^*) = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \ell_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t \rangle\right],$$

from which we can obtain the desired result by using Lemma 3 (which holds even if the loss $\widehat{\ell}_t$ at each round $t \in [T]$ depends on the prediction p_t made at that round). □

Lemma 3. *Let $q \in (0,1)$, $\eta > 0$, and $(y_t)_{t=1}^T$ be an arbitrary sequence of non-negative loss vectors in $\mathbb{R}^K$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_K and the q-Tsallis regularizer ψ_q over this sequence of losses. That is, $p_1 = \arg\min_{p \in \Delta_K} \psi_q(p)$, and for $t \in [T]$,*

$$p_{t+1} = \operatorname*{arg\,min}_{p \in \Delta_K} \eta \sum_{s=1}^{t} \big\langle y_s, p \big\rangle + \psi_q(p) \ .$$

Then for any $u \in \Delta_K$,

$$\sum_{t=1}^{T} \langle p_t - u, y_t \rangle \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i \in V} p_t(i)^{2-q}\, y_t(i)^2 \ .$$

Proof. By Theorem 28.5 in [27], we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t - u, y_t \rangle &\le \frac{\psi_q(u) - \psi_q(p_1)}{\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) \\
&= \frac{K^{1-q} - 1}{(1-q)\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) \\
&\le \frac{K^{1-q}}{(1-q)\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) \ ,
\end{aligned}$$

where $D_{\psi_q}(\cdot, \cdot)$ is the Bregman divergence based on ψ_q. For bounding each summand in the second term, we follow a similar argument to that used in Theorem 30.2 in [27]. Namely, for each $i \in V$ and

round $t \in [T]$, define $\overline{y}_t(i) = \mathbb{I}\{p_{t+1}(i) \le p_t(i)\} y_t(i)$. We then have that

$$\begin{aligned}
&\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&\qquad \le \langle p_t - p_{t+1}, \overline{y}_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&\qquad = \frac{1}{\eta} \langle p_t - p_{t+1}, \eta \overline{y}_t \rangle - \frac{1}{2\eta} \big\| p_{t+1} - p_t \big\|^2_{\nabla^2 \psi_q(z_t)} \\
&\qquad \le \frac{\eta}{2} \big\| \overline{y}_t \big\|^2_{(\nabla^2 \psi_q(z_t))^{-1}} \\
&\qquad = \frac{\eta}{2q} \sum_{i \in V} z_t(i)^{2-q} \, \overline{y}_t(i)^2 \\
&\qquad = \frac{\eta}{2q} \sum_{i \in V} \big(\gamma_t p_{t+1}(i) + (1-\gamma_t) p_t(i) \big)^{2-q} \, \overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, \overline{y}_t(i)^2 + \gamma_t \frac{\eta}{2q} \sum_{i \in V} \big(p_{t+1}(i)^{2-q} - p_t(i)^{2-q} \big) \, \overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, \overline{y}_t(i)^2 \\
&\qquad \le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, y_t(i)^2 \ ,
\end{aligned}$$

where $z_t = \gamma_t p_{t+1} + (1-\gamma_t) p_t$ for some $\gamma_t \in [0,1]$; the first inequality holds due to the non-negativity of the losses, the second inequality is an application of the Fenchel-Young inequality, the second equality holds since the Hessian of ψ_q is a diagonal matrix with $(\nabla^2 \psi_q(x))_{i,i} = q x(i)^{q-2}$, the third inequality is an application of Jensen's inequality (since $q \in (0,1)$), and the fourth inequality holds since $\overline{y}_t(i) = 0$ for any i such that $p_{t+1}(i)^{2-q} > p_t(i)^{2-q}$. □

Lemma 4. *Let a and b be positive integers such that $2 \le a \le b$, and let $n = \lceil \log_2 a \rceil$. Then,*

$$\sum_{r=0}^{n-1} \sqrt{2^r \ln\big(e^2 b 2^{-r}\big)} \le \frac{\sqrt{2\pi} + 2\sqrt{2 - \ln 2}}{\ln 2} \sqrt{a \ln\left(\frac{e^2 b}{a}\right)} \ .$$

Proof. Since $n \le \log_2(2b)$ and $2^r \ln\big(e^2 b 2^{-r}\big)$ is monotonically increasing in r for $r \in [0, \log_2(eb)]$, we can bound the sum by an integral:

$$\sum_{r=0}^{n-1} \sqrt{2^r \ln\big(e^2 b 2^{-r}\big)} \le \int_0^n \sqrt{2^r \ln\big(e^2 b 2^{-r}\big)} \, \mathrm{d}r \ .$$

We proceed via a change of variable; let $x = e^2 b 2^{-r}$, and note that $\mathrm{d}r = -\frac{\mathrm{d}x}{x \ln 2}$. We then have that

$$\begin{aligned}
\int_0^n \sqrt{2^r \ln\big(e^2 b 2^{-r}\big)} \, \mathrm{d}r &= \sqrt{e^2 b} \int_0^n \sqrt{\frac{2^r}{e^2 b} \ln\big(e^2 b 2^{-r}\big)} \, \mathrm{d}r \\
&= -\frac{e\sqrt{b}}{\ln 2} \int_{e^2 b}^{e^2 b 2^{-n}} \sqrt{\frac{\ln x}{x^3}} \, \mathrm{d}x = \frac{e\sqrt{b}}{\ln 2} \int_{e^2 b 2^{-n}}^{e^2 b} \sqrt{\frac{\ln x}{x^3}} \, \mathrm{d}x \\
&= \frac{e\sqrt{b}}{\ln 2} \Big[-\sqrt{2\pi} \cdot \mathrm{erfc}\big(\sqrt{(\ln x)/2}\big) - 2\sqrt{(\ln x)/x} \Big]_{e^2 b 2^{-n}}^{e^2 b} \\
&\le \frac{e\sqrt{b}}{\ln 2} \left(\sqrt{2\pi} \cdot \mathrm{erfc}\big(\sqrt{\ln(e^2 b 2^{-n})/2}\big) + 2\sqrt{\frac{2^n \ln(e^2 b 2^{-n})}{e^2 b}} \right) \ ,
\end{aligned}$$

where $\operatorname{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x \exp(-z^2)\,\mathrm{d}z$ is the complementary Gaussian error function, which is always positive. By [13, Theorem 1], we have that $\operatorname{erfc}(x) \leq \exp(-x^2)$. Consequently,

$$\begin{aligned}
\int_0^n \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)}\,\mathrm{d}r &\leq \frac{e\sqrt{b}}{\ln 2}\left(\sqrt{2\pi}\sqrt{\frac{2^n}{e^2 b}} + 2\sqrt{\frac{2^n \ln(e^2 b 2^{-n})}{e^2 b}}\right) \\
&= \frac{\sqrt{2^n}}{\ln 2}\left(\sqrt{2\pi} + 2\sqrt{\ln(e^2 b 2^{-n})}\right) \\
&\leq \frac{\sqrt{2a}}{\ln 2}\left(\sqrt{2\pi} + 2\sqrt{\ln\left(\frac{e^2 b}{2a}\right)}\right) \\
&\leq \frac{\sqrt{2\pi} + 2\sqrt{2 - \ln 2}}{\ln 2}\sqrt{a \ln\left(\frac{e^2 b}{a}\right)}\ ,
\end{aligned}$$

where in the second inequality we used once again the fact that $2^r \ln\bigl(e^2 b 2^{-r}\bigr)$ is monotonically increasing in r for $r \in [0, \log_2(eb)]$ to replace n with $\log_2(a) + 1$, and the last inequality holds since $b/a \geq 1$. □

B Proofs of Section 3

In this section, we provide the proof of Theorem 2, which is restated below.

Theorem 2. *Let $G_1, \ldots, G_T$ be a sequence of strongly observable undirected feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}}\ ,$$

and loss estimates (6)*, then its regret satisfies $R_T \leq 6\sqrt{e\alpha T\,(2 + \ln(K/\alpha))}$.*

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^T \ell_t(i)$ and $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be its indicator vector. Whenever J_t is nonempty, let $j_t \in V$ be the only action such that $J_t = \{j_t\}$. Similarly to [35], let $z_t = \mathbb{I}\{J_t \neq \emptyset\}\,\mathbb{I}\{I_t \in N_t(j_t)\}\,\frac{1-\ell_t(j_t)}{1-p_t(j_t)}$ and define new losses $\widetilde{\ell}_t(i) = \widehat{\ell}_t(i) + z_t$ for each time step $t \in [T]$ and each action $i \in V$. Since $p_t, \mathbf{e}_{i^*} \in \Delta_K$, we have that $\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ for every $t \in [T]$. Then, using the fact that $\mathbb{E}_t\bigl[\widehat{\ell}_t\bigr] = \ell_t$, we get that

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^T \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right]\ ,$$

where the first equality holds via the same arguments made in the proof of Lemma 2. If we consider the optimization step of Algorithm 1, computing the same inner product over the new losses $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_T$ for some $p \in \Delta_K$ gives

$$\left\langle \sum_{s=1}^t \widetilde{\ell}_s, p\right\rangle = \sum_{s=1}^t z_s + \left\langle \sum_{s=1}^t \widehat{\ell}_s, p\right\rangle\ ,$$

where the sum $\sum_{s=1}^t z_s$ is constant with respect to p. This implies that the objective functions in terms of either $(\widehat{\ell}_t)_{t\in[T]}$ and $(\widetilde{\ell}_t)_{t\in[T]}$, respectively, are minimized by the same probability distributions. However, notice that, unlike $(\widehat{\ell}_t)_{t\in[T]}$, the loss vectors in $(\widetilde{\ell}_t)_{t\in[T]}$ are always non-negative. Consequently, similar to the proof of Lemma 2, we may apply Lemma 3 to upper bound the regret of Algorithm 1 in terms of the losses $(\widetilde{\ell}_t)_{t\in[T]}$. Doing so gives

$$\mathbb{E}\left[\sum_{t=1}^T \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \leq \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\sum_{t=1}^T \mathbb{E}\left[\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\left[\widetilde{\ell}_t(i)^2\right]\right]\ . \tag{8}$$

We can bound the second term by observing that $\widetilde{\ell}_t(j_t) = 1$ whenever $J_t \neq \emptyset$. Therefore,

$$\begin{aligned}\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big] &\le 2\sum_{i\in V\setminus J_t} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widehat{\ell}_t(i)^2\Big] + 2\,\mathbb{E}_t\left[z_t^2\right]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1\\ &\le 2\sum_{i\in V\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 2\,\mathbb{E}_t\left[z_t^2\right]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1\\ &\le 2\sum_{i\in V\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 3\ ,\end{aligned}$$

where the second inequality holds because $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$ for all $i \notin J_t$, and the third inequality follows from the fact that

$$\mathbb{E}_t\left[z_t^2\right]\sum_{i\in V\setminus J_t} p_t(i)^{2-q} = \mathbb{I}\left\{J_t \neq \emptyset\right\}\frac{\big(1-\ell_t(j_t)\big)^2}{1-p_t(j_t)}\sum_{i\in V\setminus J_t} p_t(i)^{2-q} \le 1\ .$$

We can handle the remaining sum by separating it over nodes $i \in S_t$, which satisfy $P_t(i) = 1 - p_t(i)$ because of strong observability, and those in $\overline{S}_t = V \setminus S_t$. In the first case, any node $i \in S_t \setminus J_t$ has $p_t(i) \le 1/2$ and thus

$$\sum_{i\in S_t\setminus J_t}\frac{p_t(i)^{2-q}}{P_t(i)} = \sum_{i\in S_t\setminus J_t}\frac{p_t(i)^{2-q}}{1-p_t(i)} \le 2\sum_{i\in S_t\setminus J_t} p_t(i)^{2-q} \le 2\ .$$

while in the second case we have that $\sum_{i\in\overline{S}_t} p_t(i)^{2-q}/P_t(i) \le \alpha^q$ by Lemma 1 with $U = \overline{S}_t$ and $b = 1 - q$. Overall, we have shown that

$$\sum_{i\in V} p_t(i)^{2-q}\,\mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big] \le 2\sum_{i\in\overline{S}_t}\frac{p_t(i)^{2-q}}{P_t(i)} + 7 \le 2\alpha^q + 7 \le 9\alpha^q\ . \tag{9}$$

Plugging back into (8), we obtain that

$$\begin{aligned}R_T &\le \frac{K^{1-q}}{\eta(1-q)} + \frac{9\eta}{2q}\alpha^q T\\ &= 3\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}T}\\ &\le 6\sqrt{e\alpha T\left(2+\ln(K/\alpha)\right)}\ ,\end{aligned}$$

where the equality is due to our choice of η, and the last inequality follows as in the proof of Theorem 1 together with our choice of q. □

C Proofs of Section 4

In this section, we provide the proof of Theorem 3, which is restated below.

Theorem 3. *Let $C = 4\sqrt{6e}\,\frac{\sqrt{\pi}+\sqrt{4-2\ln 2}}{\ln 2}$. Then, the regret of Algorithm 2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T}\alpha_t\left(2+\ln\left(\frac{K}{\bar{\alpha}_T}\right)\right)} + \log_2\bar{\alpha}_T\ .$$

Proof. Notice that if $\bar{\alpha}_T = 1$, the initial guess is correct and the algorithm will never restart. Moreover, since in this case we have that $\alpha_t = 1$ for all t, the theorem follows trivially from the regret bound of Theorem 2. Hence, we can assume for what follows that $\bar{\alpha}_T > 1$. Let $i^* \in \arg\min_{i\in[K]}\sum_{t=1}^{T}\ell_t(i)$ and $n = \lceil\log_2\bar{\alpha}_T\rceil$. Note that the maximum value of r that the algorithm can reach is $n-1$. To

see this, observe that Lemma 1 implies that for any r and t, $H_t(q_r) \leq \alpha_t^{q_r}$. Consequently, for any $t \geq T_r$,

$$\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} \leq \frac{1}{T}\sum_{s=T_r}^{t} \alpha_s \leq \bar{\alpha}_T \leq 2^n \ .$$

For $t \in [T]$, let r_t be the value of r at round t. Without loss of generality, we assume that r takes each value in $\{0, \dots, n-1\}$ for at least two rounds. Additionally, we define $T_n = T + 2$ for convenience. We start by decomposing the regret over the n intervals (each corresponding to a value of r in $\{0, \dots, n-1\}$) and bounding the instantaneous regret with 1 for each step in which we restart (i.e., at the last step of each but the last interval):

$$\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\
&\leq \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] + n - 1 \\
&\leq \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] + \log_2 \bar{\alpha}_T \ . \qquad (10)
\end{aligned}$$

For what follows, let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^* and let $\widetilde{\ell}_t$ be as defined in the proof of Theorem 2. Fix $r \in \{0, \dots, n-1\}$, we proceed by bounding the regret in the interval $[T_r, T_{r+1} - 2]$:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \leq 2^{r_t+1}\right\}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \leq 2^{r_t+1}\right\}\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \leq 2^{r_t+1}\right\}\langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \\
&\overset{(c)}{\leq} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\,\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\right] \\
&\overset{(d)}{=} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \leq 2^{r_t+1}\right\}\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\right]\right] \\
&\overset{(e)}{\leq} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\,\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \leq 2^{r_t+1}\right\}(2H_t(q_r) + 7)\right] \\
&= \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\,\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}(2H_t(q_r) + 7)\right] \ , \qquad (11)
\end{aligned}$$

where (a) follows since $\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i\in V} p_t(i)\ell_t(i)$, $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$, and the indicator at round t is measurable with respect to $\sigma(I_1, \dots, I_{t-1})$, that is, the σ-algebra generated by $I_1, \dots, I_{t-1}$; (b) follows since $\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ holds by the definition of $\widetilde{\ell}_t$; (c) is an application of Lemma 3, justifiable with the same argument leading to (8) in the proof of Theorem 2; (d) uses once

again that the indicator at round t is measurable with respect to $\sigma(I_1, \dots, I_{t-1})$; finally, (e) follows via (9). Define $T_{r:r+1} = T_{r+1} - T_r - 1$, and notice that

$$\begin{aligned}
\sum_{t=T_r}^{T_{r+1}-2} H_t(q_r) &= \frac{T_{r:r+1}}{T_{r:r+1}} \sum_{t=T_r}^{T_{r+1}-2} \big(H_t(q_r)^{1/q_r}\big)^{q_r} \\
&\le T_{r:r+1} \left(\frac{1}{T_{r:r+1}} \sum_{t=T_r}^{T_{r+1}-2} H_t(q_r)^{1/q_r} \right)^{q_r} \\
&\le T_{r:r+1} \left(\frac{T}{T_{r:r+1}} 2^{r+1} \right)^{q_r} \\
&\le 2T\big(2^r\big)^{q_r} \ ,
\end{aligned}$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0,1)$, and the second follows from the restarting condition of Algorithm 2. Next, we plug this inequality back into (11), and then, similar to the proof of Theorem 2, we use the definitions of η_r and q_r and bound the resulting expression to get that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \big(\ell_t(I_t) - \ell_t(i^*)\big) \right] &\le \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{11\eta_r}{2q_r} T \, (2^r)^{q_r} \\
&\le 2\sqrt{11eT2^r\Big(2 + \ln\big(K2^{-r}\big)\Big)} \le 4\sqrt{3eT2^r \ln\big(e^2 K 2^{-r}\big)} \ .
\end{aligned}$$

We then sum this quantity over r and use Lemma 4 to get that

$$\begin{aligned}
\mathbb{E}\left[\sum_{r=0}^{n-1} \sum_{t=T_r}^{T_{r+1}-2} \big(\ell_t(I_t) - \ell_t(i^*)\big) \right] &\le 4\sqrt{3eT} \sum_{r=0}^{n-1} \sqrt{2^r \ln\big(e^2 K 2^{-r}\big)} \\
&\le 4\sqrt{6e} \frac{\sqrt{\pi} + \sqrt{4 - 2\ln 2}}{\ln 2} \sqrt{\bar{\alpha}_T T \big(2 + \ln\left(K/\bar{\alpha}_T\right)\big)} \ ,
\end{aligned}$$

which, together with (10), concludes the proof. □

D Proof of the Lower Bound

In this section, we prove the lower bound provided in Section 5, which we restate below. As remarked before, our proof makes use of known techniques for proving lower bounds for the multitask bandit problem. In particular, parts of the proof are adapted from the proof of Theorem 7 in [17].

Theorem 4. *Pick any $K \ge 2$ and any α such that $2 \le \alpha \le K$. Then, for any algorithm and for all $T \ge \frac{\alpha \log_\alpha K}{4 \log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1, \dots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t = 1, \dots, T$ and*

$$R_T \ge \frac{1}{18\sqrt{2}} \sqrt{\alpha T \log_\alpha K}.$$

Proof. Once again, we define $M = \log_\alpha K$, which we assume for now to be an integer; we discuss in the end how to extend the proof to the case when it is not. The proof will be divided into five parts I–V. We begin by formalizing the class of environments described in Section 5 and stating two useful lemmas.

I. Preliminaries

We remind the reader that we identify each action in V with a vector $a = \big(a(1), \dots, a(M)\big) \in [\alpha]^M$. We will focus on a set of M undirected graphs $\mathcal{G} = \{G^i\}_{i=1}^M$, where G^i consists of α isolated cliques (with self-loops) $\{C_{i,j}\}_{j=1}^\alpha$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. As remarked before, all these graphs have independence number α. For convenience, we also use actions in V as functions from $\mathcal{G}$ to $[\alpha]$, with $a(G^i) = a(i)$.

An environment is identified by a function $\mu\colon [\alpha] \times \mathcal{G} \to [0,1]$ such that at every round t, after having drawn a graph G_t from the uniform distribution over $\mathcal{G}$ (denoted with $U_{\mathcal{G}}$), the environment latently draws for each $j \in [\alpha]$ and $G \in \mathcal{G}$, a Bernoulli random variable $\gamma_t(j;G)$ with mean $\mu(j;G_t)$. Subsequently, for defining the loss of action $a \in V$ at round t, we simply set $\ell_t(a) = \gamma_t(a(G_t);G_t)$, whose expectation, conditioned on G_t, is $\mu(a(G_t);G_t)$. To simplify the notation, we use $\mu(a;G)$ as shorthand for $\mu(a(G);G)$ and $\gamma_t(a;G)$ as shorthand for $\gamma_t(a(G);G)$. Denote by A_t the action picked by the player at round t, which is chosen prior to observing G_t. We will focus on the following notion of stochastic regret, which we define for environment μ as:

$$\overline{R}_T(\mu) = \max_{a \in V} \mathbb{E}_\mu \left[\sum_{t=1}^{T} (\ell_t(A_t) - \ell_t(a)) \right] \ ,$$

where $\mathbb{E}_\mu[\cdot]$ denotes the expectation with respect to the sequence of losses and graphs generated by environment μ, as well as the randomness in the choices of the player. We can use the tower rule to rewrite this expression as

$$\begin{aligned}
\overline{R}_T(\mu) &= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\ell_t(A_t) - \ell_t(a) \,\middle|\, G_t, A_t \right] \,\middle|\, A_t \right] \right] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\mu(A_t;G_t) - \mu(a;G_t) \,\middle|\, A_t \right] \right] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu \left[\sum_{i=1}^{M} U_{\mathcal{G}}(G^i)(\mu(A_t;G^i) - \mu(a;G^i)) \right] \\
&= \max_{a \in V} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_\mu \left[\sum_{t=1}^{T} (\mu(A_t;G^i) - \mu(a;G^i)) \right] \ . \qquad (12)
\end{aligned}$$

For a fixed algorithm, one can show via standard arguments that

$$\sup_{(\ell_t)_{t=1}^T, (G_t)_{t=1}^T} R_T \geq \sup_{\mu} \overline{R}_T(\mu) \ .$$

Hence, it suffices for our purposes to prove a lower bound for the right-hand side of this inequality.

In the following, we will have to be more precise about the probability measure with respect to which the expectation in (12) is defined. Let $\boldsymbol{\lambda}_t \in \{0,1\}^{K/\alpha}$ denote the vector of losses observed by the player at round t, which corresponds to the losses of the actions connected to A_t assuming that a systematic ordering of the actions makes it clear which coordinate of $\boldsymbol{\lambda}_t$ belongs to which action. Let $\mathbf{1}_{K/\alpha}$ and $\mathbf{0}_{K/\alpha}$ be the K/α dimensional[4] vectors of all ones and all zeros respectively. Clearly, we have that $\boldsymbol{\lambda}_t = \gamma_t(A_t;G_t)\mathbf{1}_{K/\alpha} = \ell_t(A_t)\mathbf{1}_{K/\alpha}$, which is a binary random variable taking values in $\{\mathbf{0}_{K/\alpha}, \mathbf{1}_{K/\alpha}\}$. Let $\mathbb{P}_\mu^\lambda$ be the probability distribution of $\boldsymbol{\lambda}_t$ in environment μ. Notice then that we have that

$$\mathbb{P}_\mu^\lambda(\gamma_t = \mathbf{1}_{K/\alpha} \,|\, G_t = G, A_t = a) = \mu(a;G) \ . \qquad (13)$$

Let $H_t = (A_1, G_1, \boldsymbol{\lambda}_1, \ldots, A_t, G_t, \boldsymbol{\lambda}_t) \in (V \times \mathcal{G} \times \{0,1\}^{K/\alpha})^t$ be the interaction trajectory after t steps. The policy π adopted by the player can be modelled as a sequence of probability kernels $\{\pi_t\}_{t=1}^T$ each mapping the trajectory so far to a distribution over the actions, i.e., A_t is sampled from $\pi_t(\cdot \,|\, H_{t-1})$. An environment μ and a policy π (implicit in the notation, and fixed throughout the rest of the proof) together define a distribution $\mathbb{P}_\mu$ over the set of possible trajectories of T steps such that:

$$\mathbb{P}_\mu(H_T) = \prod_{t=1}^{T} \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}_\mu^\lambda(\lambda_t \,|\, G_t, A_t) \ .$$

If P and Q are two distributions defined on the same space, let $D_{\mathrm{KL}}(P \,\|\, Q)$ and $\delta(P,Q)$ be the KL-divergence and the total variation distance respectively between P and Q. Furthermore, let $d(p \,\|\, q)$ be the KL-divergence between two Bernoulli random variables with means p and q. The following lemma provides an expression for the KL-divergence between two the probability distributions associated to two environments.

[4]Note that $K/\alpha = \alpha^{M-1}$ is an integer since $M(\geq 1)$ was assumed to be an integer.

Lemma 5. *For a fixed policy, let μ and μ' be two environments as described above. Then,*

$$D_{\mathrm{KL}}(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) = \frac{1}{M}\sum_{i=1}^{M}\sum_{a\in V} N_\mu(a;T)d\big(\mu(a;G^i)\,\big\|\,\mu'(a;G^i)\big) \ ,$$

where $N_\mu(a;T) = \mathbb{E}_\mu\Big[\sum_{t=1}^{T}\mathbb{I}\{A_t = a\}\Big]$.

Proof. The proof is similar to that of Lemma 15.1 in [27]. Namely, we have in our case that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) &= \mathbb{E}_\mu\left[\ln\frac{\mathbb{P}_\mu(H_T)}{\mathbb{P}_{\mu'}(H_T)}\right] \\
&= \mathbb{E}_\mu\left[\ln\frac{\prod_{t=1}^{T}\pi_t(A_t\,|\,H_{t-1})U_{\mathcal{G}}(G_t)\mathbb{P}_\mu^\lambda(\lambda_t\,|\,G_t,A_t)}{\prod_{t=1}^{T}\pi_t(A_t\,|\,H_{t-1})U_{\mathcal{G}}(G_t)\mathbb{P}_{\mu'}^\lambda(\lambda_t\,|\,G_t,A_t)}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_\mu\left[\ln\frac{\mathbb{P}_\mu^\lambda(\lambda_t\,|\,G_t,A_t)}{\mathbb{P}_{\mu'}^\lambda(\lambda_t\,|\,G_t,A_t)}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\ln\frac{\mathbb{P}_\mu^\lambda(\lambda_t\,|\,G_t,A_t)}{\mathbb{P}_{\mu'}^\lambda(\lambda_t\,|\,G_t,A_t)}\,\middle|\,G_t,A_t\right]\,\middle|\,A_t\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot\,|\,G_t,A_t)\,\|\,\mathbb{P}_{\mu'}^\lambda(\cdot\,|\,G_t,A_t))\Big|A_t\right]\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_\mu\left[\sum_{i=1}^{M}U_{\mathcal{G}}(G^i)D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot\,|\,G^i,A_t)\,\|\,\mathbb{P}_{\mu'}^\lambda(\cdot\,|\,G^i,A_t))\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\mathbb{E}_\mu\left[D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot\,|\,G^i,A_t)\,\|\,\mathbb{P}_{\mu'}^\lambda(\cdot\,|\,G^i,A_t))\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\sum_{a\in V}N_\mu(a;T)D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot\,|\,G^i,a)\,\|\,\mathbb{P}_{\mu'}^\lambda(\cdot\,|\,G^i,a)) \\
&= \frac{1}{M}\sum_{i=1}^{M}\sum_{a\in V}N_\mu(a;T)d\big(\mu(a;G^i)\,\big\|\,\mu'(a;G^i)\big) \ ,
\end{aligned}$$

where the last equality holds via (13). □

The following standard lemma, adapted from [27], will be used in the sequel.

Lemma 6. *Let P and Q be probability measures on the same measurable space $(\Omega,\mathcal{F})$. Let $a<b$ and $X:\Omega\to[a,b]$ be an $\mathcal{F}$-measurable random variable. Then,*

$$\left|\int_\Omega X(\omega)dP(\omega) - \int_\Omega X(\omega)dQ(\omega)\right| \le (b-a)\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\,\|\,Q)} \ .$$

Proof. We have, by Exercise 14.4 in [27], that

$$\left|\int_\Omega X(\omega)dP(\omega) - \int_\Omega X(\omega)dQ(\omega)\right| \le (b-a)\delta(P,Q) \ ,$$

from which the lemma follows by applying Pinsker's inequality. □

II. Choosing the environments

We will construct a collection of environments $\{\mu_a\}_{a\in V}$, each associated to an action, such that for any $i\in[M]$ and $j\in[\alpha]$,

$$\mu_a(j;G^i) = \frac{1}{2} - \varepsilon\mathbb{I}\{a(i)=j\} \ ,$$

where $0 < \varepsilon \leq \frac{1}{4}$ will be tuned later. In words, for a fixed graph, environment μ_a gives a slight advantage to actions that are connected to a in that graph, and thus agree with a in the corresponding game. Additionally, for every $a \in V$ and $i \in [M]$, we define environment μ_a^{-i} to be such that for any $s \in [M]$ and $j \in [\alpha]$,

$$\mu_a^{-i}(j; G^s) = \begin{cases} \frac{1}{2}, & \text{if } s = i \\ \mu_a(j; G^s), & \text{otherwise.} \end{cases}$$

Similar to [17], we will define, for every $i \in [M]$, an equivalence relation $\sim_i$ on the arms such that

$$a \sim_i a' \iff \forall s \in [M] \setminus \{i\}, a'(s) = a(s) \ ,$$

for any $a, a' \in V$. This means that two arms are equivalent according to $\sim_i$ if and only if their choices of base actions coincide in all games that are different from i. Let $V/\sim_i$ be the set of equivalence classes of $\sim_i$. It is easy to see that $V/\sim_i$ contains exactly α^{M-1} equivalence classes, and that each class consists of α actions, each corresponding to a different choice of base action in game i. Notice then that for an equivalence class $W \in V/\sim_i$, all environments μ_a^{-i} with $a \in V$ are indeed identical. In the sequel, this environment will also be referred to as μ_W^{-i}.

III. Lower-bounding the regret of a single environment

Note that in environment μ_a, we have that $a = \arg\min_{a' \in V} \sum_{i=1}^{M} \mu_a(a'; G^i)$. Consequently, starting from (12) we get that

$$\begin{aligned}
\overline{R}_T(\mu_a) &= \sum_{i=1}^{M} \frac{1}{M} \mathbb{E}_{\mu_a}\left[\sum_{t=1}^{T} (\mu_a(A_t; G^i) - \mu_a(a; G^i))\right] \\
&= \sum_{i=1}^{M} \frac{1}{M} \mathbb{E}_{\mu_a}\left[\sum_{t=1}^{T} \left(\frac{1}{2} - \varepsilon \mathbb{I}\{A_t(i) = a(i)\} - \left(\frac{1}{2} - \varepsilon\right)\right)\right] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^{M} \mathbb{E}_{\mu_a}\left[\sum_{t=1}^{T} (1 - \mathbb{I}\{A_t(i) = a(i)\})\right] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^{M} \big(T - N_{\mu_a}(i, a; T)\big) \ ,
\end{aligned}$$

where for environment μ, action a, and game i, $N_\mu(i, a; T) = \mathbb{E}_\mu\big[\sum_{t=1}^{T} \mathbb{I}\{A_t(i) = a(i)\}\big]$ is the expected number of times in environment μ that the action chosen by the policy agrees with action a in game i. Next, we use Lemma 6 to obtain that

$$\overline{R}_T(\mu_a) \geq \frac{\varepsilon}{M} \sum_{i=1}^{M} \left(T - N_{\mu_a^{-i}}(i, a; T) - T\sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(\mathbb{P}_{\mu_a^{-i}} \,\big\|\, \mathbb{P}_{\mu_a}\big)}\right) \ . \tag{14}$$

For bounding the KL-divergence term, we start from Lemma 5:

$$\begin{aligned}
D_{\mathrm{KL}}\big(\mathbb{P}_{\mu_a^{-i}} \,\big\|\, \mathbb{P}_{\mu_a}\big) &= \frac{1}{M}\sum_{s=1}^{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(\mu_a^{-i}(a';G^s) \,\big\|\, \mu_a(a';G^s)\big) \\
&= \frac{1}{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(\mu_a^{-i}(a';G^i) \,\big\|\, \mu_a(a';G^i)\big) \\
&= \frac{1}{M}\sum_{a'\in V} N_{\mu_a^{-i}}(a';T)\, d\big(1/2 \,\big\|\, 1/2-\varepsilon\mathbb{I}\{a'(i)=a(i)\}\big) \\
&= \frac{1}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\} N_{\mu_a^{-i}}(a';T)\, d\big(1/2 \,\big\|\, 1/2-\varepsilon\big) \\
&\le \frac{c\varepsilon^2}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\} N_{\mu_a^{-i}}(a';T) \\
&= \frac{c\varepsilon^2}{M}\sum_{a'\in V} \mathbb{I}\{a'(i)=a(i)\}\, \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^{T} \mathbb{I}\{A_t=a'\} \\
&= \frac{c\varepsilon^2}{M}\, \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^{T} \mathbb{I}\{A_t(i)=a(i)\} \\
&= \frac{c\varepsilon^2}{M} N_{\mu_a^{-i}}(i,a;T) \ ,
\end{aligned}$$

where the second equality holds since the two environments only differ in G^i, and the inequality holds for $\varepsilon \le \frac{1}{4}$ with $c = 8\log\frac{4}{3}$. Plugging back into (14) gets us that

$$\overline{R}_T(\mu_a) \ge \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M} N_{\mu_a^{-i}}(i,a;T)}\right) \ . \tag{15}$$

IV. Summing up

Fix $i \in [M]$. Notice that for $W \in V/\sim_i$,

$$\sum_{a\in W} \mathbb{I}\{A_t(i)=a(i)\} = 1$$

since each action in W corresponds to a different choice of base action in game i. Hence,

$$\begin{aligned}
\frac{1}{\alpha^M}\sum_{a\in V} N_{\mu_a^{-i}}(i,a;T) &= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W} N_{\mu_a^{-i}}(i,a;T) \\
&= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W} N_{\mu_W^{-i}}(i,a;T) \\
&= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i} \mathbb{E}_{\mu_W^{-i}}\left[\sum_{t=1}^{T}\sum_{a\in W}\mathbb{I}\{A_t(i)=a(i)\}\right] \\
&= \frac{1}{\alpha^M}\alpha^{M-1}T = \frac{T}{\alpha} \ .
\end{aligned}$$

Using this together with (15) allows us to conclude that

$$\begin{aligned}
\sup_{\mu} \overline{R}_T(\mu) &\geq \frac{1}{\alpha^M} \sum_{a \in V} \overline{R}_T(\mu_a) \\
&\geq \frac{1}{\alpha^M} \sum_{a \in V} \frac{\varepsilon}{M} \sum_{i=1}^{M} \left(T - N_{\mu_a^{-i}}(i, a; T) - \varepsilon T \sqrt{\frac{c}{2M} N_{\mu_a^{-i}}(i, a; T)} \right) \\
&\geq \frac{\varepsilon}{M} \sum_{i=1}^{M} \left(T - \frac{1}{\alpha^M} \sum_{a \in V} N_{\mu_a^{-i}}(i, a; T) - \varepsilon T \sqrt{\frac{c}{2M\alpha^M} \sum_{a \in V} N_{\mu_a^{-i}}(i, a; T)} \right) \\
&= \frac{\varepsilon}{M} \sum_{i=1}^{M} \left(T - \frac{T}{\alpha} - \varepsilon T \sqrt{\frac{cT}{2M\alpha}} \right) \\
&= \varepsilon T \left(1 - \frac{1}{\alpha} - \varepsilon \sqrt{\frac{cT}{2M\alpha}} \right) \\
&\geq \varepsilon T \left(\frac{1}{2} - \varepsilon \sqrt{\frac{cT}{2M\alpha}} \right) \ ,
\end{aligned}$$

where the third inequality holds due to the concavity of the square root, and the last inequality holds by our assumption that $\alpha \geq 2$. Setting $\varepsilon = \frac{1}{4}\sqrt{\frac{2M\alpha}{cT}}$ yields that

$$\sup_{\mu} \overline{R}_T(\mu) \geq \frac{1}{16}\sqrt{\frac{2}{c}} \cdot \sqrt{\alpha T M} \geq \frac{1}{18}\sqrt{\alpha T M} = \frac{1}{18}\sqrt{\alpha T \log_\alpha K} \ ,$$

whereas it holds that $\varepsilon \leq \frac{1}{4}$ thanks to the assumption made on T in the statement of the theorem.

V. The case when $\log_\alpha K$ is not an integer

If M is not an integer,[5] we can use the same construction as before for the first $\alpha^{\lfloor M \rfloor}$ actions and force the remaining actions to behave identically to some action in the construction. That is, we can designate a certain action such that, in all environments, all the excess actions receive the same loss as this action and are connected to it, to each other, and to every action that happens to share an edge with this designated action in a given graph (in other words, we are expanding the designated action into a clique). This way, the independence number of all the graphs in the construction is still α, and the excess actions do not provide any extra utility to the learner; playing one of them is exactly like playing the designated action, and the construction does not hide this from the player. We can then obtain the same bound as before but in terms of $\lfloor M \rfloor$, thus costing us an extra $1/\sqrt{2}$ factor to recover the desired bound (using that $\lfloor M \rfloor \geq M/2$). □

E Comparison with [14]

In [14], the authors consider a special case of the undirected feedback graph problem where the graph (fixed and known) is composed of α disjoint cliques with self-loops. For $j \in [\alpha]$, let m_j denote the number of actions in the j-th clique, implying that $\sum_{j=1}^{\alpha} m_j = K$ (the number of arms). For this problem, [14, Theorem 4] provides a lower bound of order $\sqrt{T \sum_{j=1}^{\alpha} \ln(m_j + 1)}$. In particular, if the cliques are balanced (i.e., $m_1 = \cdots = m_\alpha = K/\alpha$), the lower bound becomes of order $\sqrt{\alpha T \ln(1 + K/\alpha)}$, thus matching the regret bound of Algorithm 1. This means that, for any value of $1 \leq \alpha \leq K$, there are feedback graphs on K nodes with independence number α such that no other algorithm can achieve a better minimax regret guarantee than that of our proposed algorithm.

We emphasize that this does not imply *graph-specific* minimax optimality. Indeed, as shown in [14], when the cliques are unbalanced, the regret guarantee of our algorithm can be inferior to that of the algorithm they proposed, which matches the $\sqrt{T \sum_{j=1}^{\alpha} \ln(m_j + 1)}$ bound. However, beyond

[5] Note that M is never smaller than 1 since $\alpha \leq K$.

the disjoint cliques case, their algorithm requires computing a minimum clique cover for the given feedback graph G, which is known to be NP-hard [24]. More importantly, their reliance on a clique cover leads to a dependence of the regret on the clique cover number $\theta(G)$ instead of the independence number $\alpha(G)$. One can argue that the ratio between $\theta(G)$ and $\alpha(G)$ can be $\Omega(K/(\ln K)^2)$ for most graphs on a sufficiently large number K of vertices (e.g., see [29, Section 6]). Finally, it is not clear how to generalize their approach to time-varying feedback graphs (informed or uninformed). Hence, despite the contributions of our work and those of [14], the problem of characterizing the minimax regret rate at a graph-based granularity still calls for further investigation.

F Directed Strongly Observable Feedback Graphs

In this section, we consider the case of directed strongly observable graphs. For a directed graph $G = (V, E)$, let $N_G^{\text{in}}(i) = \{j \in V : (j, i) \in E\}$ be the in-neighbourhood of node $i \in V$ in G, and let $N_G^{\text{out}}(i) = \{j \in V : (i, j) \in E\}$ be its out-neighbourhood. A directed graph G is strongly observable if for every $i \in V$, at least one of the following holds: $i \in N_G^{\text{in}}(i)$ or $j \in N_G^{\text{in}}(i)$ for all $j \neq i$. The independence number $\alpha(G)$ is still defined in the same manner as before; that is, the cardinality of the largest set of nodes such that no two nodes share an edge, regardless of orientation. The interaction protocol is the same as in the undirected case, except that, in each round $t \in [T]$, the learner only observes the losses of the actions in $N_{G_t}^{\text{out}}(I_t)$, which is the out-neighbourhood in graph G_t of the action I_t picked by the learner. As before, we will use $N_t^{\text{in}}(i)$ and $N_t^{\text{out}}(i)$ to denote $N_{G_t}^{\text{in}}(i)$ and $N_{G_t}^{\text{out}}(i)$ respectively. For this setting, a bound of $\mathcal{O}\big(\sqrt{\alpha T} \cdot \ln(KT)\big)$ was proven in [2] for the EXP3.G algorithm. Later, [35] proved a bound of $\mathcal{O}\big(\sqrt{\alpha T (\ln K)^3}\big)$ for OSMD with a variant of the q-Tsallis entropy regularizer where q was chosen as $1 - 1/(\ln K)$.

To use Algorithm 1 in the directed case, one can define loss estimates analogous to (6) by using the in-neighbourhood in place of the neighbourhood in the relevant quantities. Namely, let $S_t = \big\{i \in V : i \notin N_t^{\text{in}}(i)\big\}$, $J_t = \{i \in S_t : p_t(i) > 1/2\}$, and $P_t(i) = \sum_{j \in N_t^{\text{in}}(i)} p_t(j)$. The loss estimates (again due to [35]) can then be given by

$$\widehat{\ell}_t(i) = \begin{cases} \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\big\{I_t \in N_t^{\text{in}}(i)\big\} & \text{if } i \in V \setminus J_t \\ \frac{\ell_t(i)-1}{P_t(i)} \mathbb{I}\big\{I_t \in N_t^{\text{in}}(i)\big\} + 1 & \text{if } i \in J_t \ . \end{cases}$$

Algorithm 1 with these loss estimates can be analyzed in a similar manner to the proof of Theorem 2, with the major difference being the way that the variance term is handled for actions with self-loops. Namely, the relevant term is

$$\sum_{i \in V : i \in N_t^{\text{in}}(i)} \frac{p_t(i)^{2-q}}{\sum_{j \in N_t^{\text{in}}(i)} p_t(j)},$$

on which we elaborate more in the following.

Let $p \in \Delta_K$ and $\beta \in (0, 1/2)$ be such that $\min_{i \in V} p(i) \geq \beta$. We first consider the variance term given by the negative Shannon entropy regularizer. It is known [2] that such a variance term, restricted to nodes with a self-loop in the strongly observable feedback graph $G = (V, E)$, has an upper bound of the form

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} \leq 4\alpha(G) \ln\left(\frac{4K}{\alpha(G)\beta}\right) \ . \tag{16}$$

In addition to the fact that this variance bound has a linear dependence on the independence number $\alpha(G)$ of G, we observe that there is a logarithmic factor in K/α and $1/\beta$ given by the fact that we now consider directed graphs. The main problem is that, in general, we cannot hope to improve upon the above logarithmic factor as it can be shown to be unavoidable unless we manage to restrict the probability distributions we consider. Indeed, it is possible to show [3, Fact 4] that there exist probability distributions $p \in \Delta_K$ and directed strongly observable graphs G for which $\alpha(G) = 1$ and

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} = \frac{K+1}{2} = \frac{1}{2} \log_2\left(\frac{4}{\min_i p(i)}\right) = \alpha(G) \log^{\omega(1)}\left(\frac{K}{\alpha(G)}\right) \ .$$

A usual way to avoid this is to introduce some explicit exploration to the probability distributions in order to force a lower bound on the probabilities of all nodes, e.g., as in EXP3.G [2]. This would bring

the linear dependence on K down to α in the above bad case, while, on the other hand, introducing a $\ln(KT)$ factor which then worsens the overall dependence on the time horizon T.

Consider now the variance term given by the analysis of the q-FTRL algorithm. As already argued in Section 3, we can reuse the variance bound in (16) for the case of negative Shannon entropy because

$$\sum_{i \in V : i \in N_G^{\mathrm{in}}(i)} \frac{p(i)^{2-q}}{\sum_{j \in N_G^{\mathrm{in}}(i)} p(j)} \leq \sum_{i \in V : i \in N_G^{\mathrm{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\mathrm{in}}(i)} p(j)}$$

for any $q \in (0,1)$, and such a bound is the best known so far for the general case of directed strongly observable graphs. However, we can be more clever in the way we utilize it. Similarly to the proof of [35, Theorem 14], we can gain an advantage from the adoption of q-FTRL by splitting the sum in the variance term into two sums according to some adequately chosen threshold β on the probabilities of the individual nodes. More precisely, by choosing $\beta \approx \exp\bigl(-\ln(K/\alpha)\ln K\bigr)$ and $q = 1 - 1/(\ln K)$, we can prove that

$$\sum_{i \in V : i \in N_G^{\mathrm{in}}(i)} \frac{p(i)^{2-q}}{\sum_{j \in N_G^{\mathrm{in}}(i)} p(j)} = \mathcal{O}\left(\alpha \ln K \left(1 + \ln \frac{K}{\alpha}\right)\right) \ .$$

We can further argue that, by following a similar analysis as in the proofs of Theorems 1 and 2, this variance bound would allow to show that the regret of q-FTRL is $\mathcal{O}\Bigl(\sqrt{\alpha T\bigl(1 + \ln(K/\alpha)\bigr)} \cdot \ln K\Bigr)$, where there is an additional $\ln K$ factor when compared to our regret bound in the undirected case (Theorem 2).

The presence of extra logarithmic factors is to be expected in the directed case, as many edges between distinct nodes might reduce the independence number of the graph, while providing information in one direction only. However, the undirected graph G' obtained from any directed strongly observable graph G by reciprocating edges between distinct nodes has the same independence number $\alpha(G') = \alpha(G)$ but the regret guarantee given by the more general analysis of q-FTRL would introduce a spurious $\ln K$ multiplicative factor. We remark that all the currently available upper bounds on the variance term (either with negative Shannon entropy or negative q-Tsallis entropy regularizers) do not exactly reflect the phenomenon of a gradually disappearing logarithmic factor when the graph is closer to being undirected (i.e., has fewer unreciprocated edges).

Taking these observations into account, we believe that it should be possible to achieve tighter guarantees that match our intuition, by improving the currently available tools. The bound on the variance term, for instance, is one part of the analysis that might be improvable. We might want to have a similar bound as (16) but with a sublinear dependence on α that varies according to the parameter q of the negative q-Tsallis entropy; e.g., ignoring logarithmic factors, we could expect it to become of order α^q as we managed to prove for the undirected case (Lemma 1). Doing so could allow a better tuning of q that might lead to improved logarithmic factors in the regret.