JMLR: Workshop and Conference Proceedings vol 40:1–13, 2015

Online Learning with Feedback Graphs: Beyond Bandits

Noga Alon NOGAA@POST.TAU.AC.IL
Tel Aviv University, Tel Aviv, Israel, and Microsoft Research, Herzliya, Israel

Nicolò Cesa-Bianchi NICOLO.CESA-BIANCHI@UNIMI.IT
Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy

Ofer Dekel OFERD@MICROSOFT.COM
Microsoft Research, Redmond, WA, USA

Tomer Koren TOMERK@TECHNION.AC.IL
Technion—Israel Institute of Technology, Haifa, Israel, and Microsoft Research, Herzliya, Israel

Abstract

We study a general class of online learning problems where the feedback is specified by a graph. This class includes online prediction with expert advice and the multi-armed bandit problem, but also several learning problems where the online player does not necessarily observe his own loss. We analyze how the structure of the feedback graph controls the inherent difficulty of the induced T-round learning problem. Specifically, we show that any feedback graph belongs to one of three classes: *strongly observable* graphs, *weakly observable* graphs, and *unobservable* graphs. We prove that the first class induces learning problems with $\widetilde{\Theta}(\alpha^{1/2}T^{1/2})$ minimax regret, where α is the independence number of the underlying graph; the second class induces problems with $\widetilde{\Theta}(\delta^{1/3}T^{2/3})$ minimax regret, where δ is the domination number of a certain portion of the graph; and the third class induces problems with linear minimax regret. Our results subsume much of the previous work on learning with feedback graphs and reveal new connections to partial monitoring games. We also show how the regret is affected if the graphs are allowed to vary with time.

1. Introduction

Online learning can be formulated as a repeated game between a randomized player and an arbitrary, possibly adversarial, environment (see, e.g., Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2011). We focus on the version of the game where, on each round, the player chooses one of K actions and incurs a corresponding loss. The loss associated with each action on each round is a number between 0 and 1, assigned in advance by the environment. The player's performance is measured using the game-theoretic notion of regret, which is the difference between his cumulative loss and the cumulative loss of the best fixed action in hindsight. We say that the player is *learning* if his regret after T rounds is $o(T)$.

After choosing an action, the player observes some feedback, which enables him to learn and improve his choices on subsequent rounds. A variety of different feedback models are discussed in online learning. The most common is *full feedback*, where the player gets to see the loss of all the actions at the end of each round. This feedback model is often called *prediction with expert advice* (Cesa-Bianchi et al., 1997; Littlestone and Warmuth, 1994; Vovk, 1990). For example, imagine a single-minded stock market investor who invests all of his wealth in one of K stocks on each day.

At the end of the day, the investor incurs the loss associated with the stock he chose, but he also observes the loss of all the other stocks.

Another common feedback model is *bandit feedback* (Auer et al., 2002), where the player only observes the loss of the action that he chose. In this model, the player's choices influence the feedback that he receives, so he has to balance an exploration-exploitation trade-off. On the one hand, the player wants to exploit what he has learned from the previous rounds by choosing an action that is expected to have a small loss; on the other hand, he wants to explore by choosing an action that will give him the most informative feedback. The canonical example of online learning with bandit feedback is online advertising. Say that we operate an Internet website and we present one of K ads to each user that views the site. Our goal is to maximize the number of clicked ads and therefore we incur a unit loss whenever a user doesn't click on an ad. We know whether or not the user clicked on the ad we presented, but we don't know whether he would have clicked on any of the other ads.

Full feedback and bandit feedback are special cases of a general framework introduced by Mannor and Shamir (2011), where the feedback model is specified by a *feedback graph*. A feedback graph is a directed graph whose nodes correspond to the player's K actions. A directed edge from action i to action j (when $i = j$ this edge is called a *self-loop*) indicates that whenever the player chooses action i he gets to observe the loss associated with action j. The full feedback model is obtained by setting the feedback graph to be the directed clique (including all self-loops, see Fig. 1a). The bandit feedback model is obtained by the graph that only includes the self-loops (see Fig. 1b). Feedback graphs can describe many other interesting online learning scenarios, as discussed below.

Our main goal is to understand how the structure of the feedback graph controls the inherent difficulty of the induced online learning problem. While regret measures the performance of a specific player or algorithm, the inherent difficulty of the game itself is measured by the *minimax regret*, which is the regret incurred by an optimal player that plays against the worst-case environment. Freund and Schapire (1997) prove that the minimax regret of the full feedback game is $\Theta(\sqrt{T \ln K})$ while Auer et al. (2002) prove that the minimax regret of the bandit feedback game is $\widetilde{\Theta}(\sqrt{KT})$. Both of these settings correspond to feedback graphs where all of the vertices have self-loops—we say that the player in these settings is *self-aware*: he observes his own loss value on each round. The minimax regret rates induced by self-aware feedback graphs were extensively studied by Alon et al. (2014). In this paper, we focus on the intriguing situation that occurs when the feedback graph is missing some self-loops, namely, when the player does not always observe his own loss. He is still accountable for the loss on each round, but he does not always know how much loss he incurred. As revealed by our analysis, the absence of self-loops can have a significant impact on the minimax regret of the induced game.

An example of a concrete setting where the player is not always self-aware is the *apple tasting* problem (Helmbold et al., 2000). In this problem, the player examines a sequence of apples, some of which may be rotten. For each apple, he has two possible actions: he can either discard the apple (action 1) or he can ship the apple to the market (action 2). The player incurs a unit loss whenever he discards a good apple and whenever he sends a rotten apple to the market. However, the feedback is asymmetric: whenever the player chooses to discard an apple, he first tastes the apple and obtains full feedback; on the other hand, whenever he chooses to send the apple to the market, he doesn't taste it and receives no feedback at all. The feedback graph that describes the apple tasting problem is shown in Fig. 1d. Another problem that is closely related to apple tasting is the *revealing action* or *label efficient* problem (Cesa-Bianchi and Lugosi, 2006, Example 6.4). In this problem, one

action is a special action, called the revealing action, which incurs a constant unit loss. Whenever the player chooses the revealing action, he receives full feedback. Whenever the player chooses any other action, he observes no feedback at all (see Fig. 1e).

Yet another interesting example where the player is not self-aware is obtained by setting the feedback graph to be the *loopless clique* (the directed clique minus the self-loops, see Fig. 1c). This problem is the complement to the bandit problem: when the player chooses an action, he observes the loss of all the other actions, but he does not observe his own loss. To motivate this, imagine a police officer who wants to prevent crime. On each day, the officer chooses to stand in one of K possible locations. Criminals then show up at some of these locations: if a criminal sees the officer, he runs away before being noticed and the crime is prevented; otherwise, he goes ahead with the crime. The officer gets a unit reward for each crime he prevents,[1] and at the end of each day he receives a report of all the crimes that occurred that day. By construction, the officer does not know if his presence prevented a planned crime, or if no crime was planned for that location. In other words, the officer observes everything but his own reward.

Our main result is a full characterization of the minimax regret of online learning problems defined by feedback graphs. Specifically, we categorize the set of all feedback graphs into three distinct sets. The first is the set of *strongly observable* feedback graphs, which induce online learning problems whose minimax regret is $\widetilde{\Theta}(\alpha^{1/2}T^{1/2})$, where α is the independence number of the feedback graph. This slow-growing minimax regret rate implies that the problems in this category are easy to learn. The set of strongly observable feedback graphs includes the set of self-aware graphs, so this result extends the characterization given in Alon et al. (2014). The second category is the set of *weakly observable* feedback graphs, which induce learning problems whose minimax regret is $\widetilde{\Theta}(\delta^{1/3}T^{2/3})$, where δ is a new graph-dependent quantity called the weak domination number of the feedback graph. The minimax regret of these problems grows at a faster rate of $T^{2/3}$ with the number of rounds, which implies that the induced problems are hard to learn. The third category is the set of *unobservable* graphs, which induce unlearnable $\Theta(T)$ online problems.

Our characterization bears some surprising implications. For example, the minimax regret for the loopless clique is the same, up to constant factors, as the $\Theta(\sqrt{T \ln K})$ minimax regret for the full feedback graph. However, if we start with the full feedback graph (the directed clique with self-loops) and remove a self-loop and an incoming edge from any node (see Fig. 1f), we are left with a weakly observable feedback graph, and the minimax regret jumps to order $T^{2/3}$. Another interesting property of our characterization is how the two learnable categories of feedback graphs depend on completely different graph-theoretic quantities: the independence number α and the weak domination number δ.

The setting of online learning with feedback graphs is closely related to the more general setting of partial monitoring (see, e.g., Cesa-Bianchi and Lugosi, 2006, Section 6.4), where the player's feedback is specified by a feedback matrix, rather than a feedback graph. Partial monitoring games have also been categorized into three classes: easy problems with $T^{1/2}$ regret, hard problems with $T^{2/3}$ regret, and unlearnable problems with linear regret (Bartók et al., 2014, Theorem 2). If the loss values are chosen from a finite set (say $\{0, 1\}$), then bandit feedback, apple tasting feedback, and the revealing action feedback models are all known to be special cases of partial monitoring. In fact, it can be shown that any problem in our setting (at least one with binary losses) can be reduced to the partial monitoring setting (see the full paper Alon et al., 2015). Nevertheless, the characterization

1. It is easier to describe this example in terms of maximizing rewards, rather than minimizing losses. In our formulation of the problem, a reward of r is mathematically equivalent to a loss of $1 - r$.

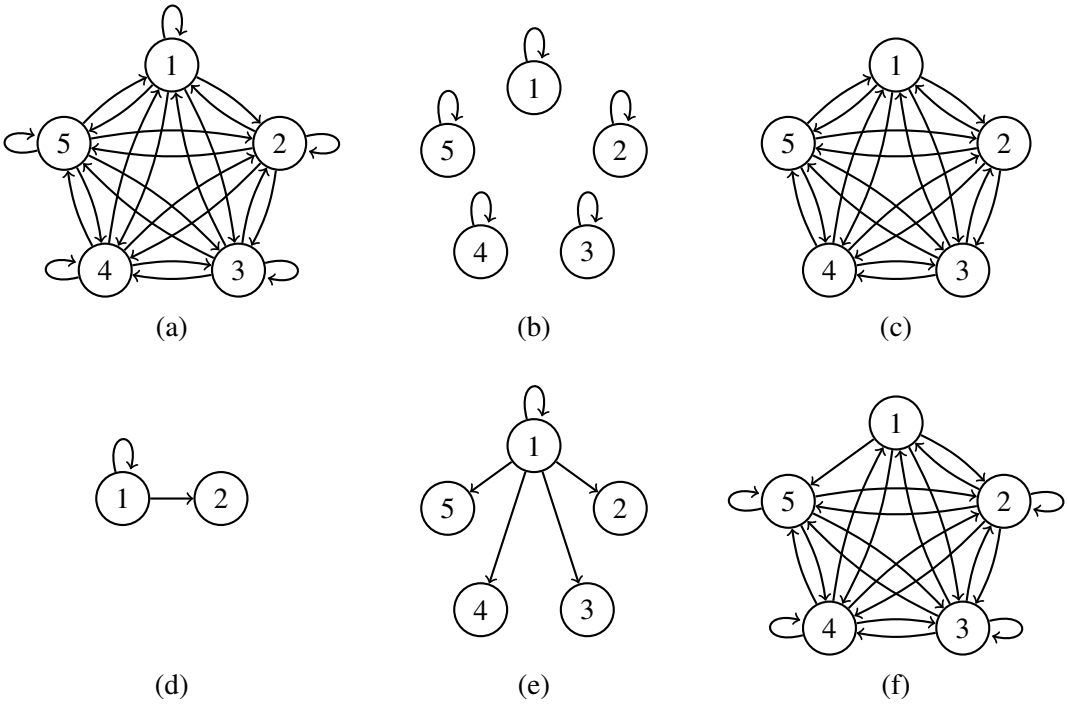

Figure 1: Examples of feedback graphs: (a) *full feedback*, (b) *bandit feedback*, (c) *loopless clique*, (d) *apple tasting*, (e) *revealing action*, (f) a clique minus a self-loop and another edge.

presented in this paper has several clear advantages over the more general characterization of partial monitoring games. First, our regret bounds are minimax optimal not only with respect to T, but also with respect to the relevant properties of the feedback graph. Second, we obtain our upper bounds with a simple and efficient algorithm. Third, our characterization is stated in terms of simple and intuitive combinatorial properties of the problem.

The setting discussed above can be generalized by allowing the feedback graphs to change arbitrarily from round to round (see Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014). We defer the treatment of this more general case to the full version (Alon et al., 2015) of the paper, where we extend our analysis to the case where the feedback graph is neither fixed nor known in advance, and discuss whether this generalization increases the minimax regret of the induced online learning problem.

The paper is organized as follows. In Section 2 we define the problem setting and state our main results. In Section 3 we describe our player algorithm and prove upper bounds on the minimax regret. Finally, in Section 4 we prove matching lower bounds on the minimax regret.

2. Problem Setting and Main Results

Let $G = (V, E)$ be a directed feedback graph over the set of actions $V = \{1, \ldots, K\}$. For each $i \in V$, let $N^{\text{in}}(i) = \{j \in V : (j, i) \in E\}$ be the in-neighborhood of i in G, and let $N^{\text{out}}(i) =$

$\{j \in V : (i,j) \in E\}$ be the out-neighborhood of i in G. If i has a self-loop, that is $(i,i) \in E$, then $i \in N^{\mathrm{in}}(i)$ and $i \in N^{\mathrm{out}}(i)$.

Before the game begins, the environment privately selects a sequence of loss functions $\ell_1, \ell_2 \ldots$, where $\ell_t : V \mapsto [0,1]$ for each $t \geq 1$. On each round $t = 1, 2, \ldots$, the player randomly chooses an action $I_t \in V$ and incurs the loss $\ell_t(I_t)$. At the end of round t, the player receives the feedback $\{\big(j, \ell_t(j)\big) : j \in N^{\mathrm{out}}(I_t)\}$. In words, the player observes the loss associated with each vertex in the out-neighborhood of the chosen action I_t. In particular, if I_t has no self-loop, then the player's loss $\ell_t(I_t)$ remains unknown, and if the out-neighborhood of I_t is empty, then the player does not observe any feedback on that round. The player's *expected regret* against a specific loss sequence $\ell_1, \ldots, \ell_T$ is defined as $\mathbb{E}\big[\sum_{t=1}^T \ell_t(I_t)\big] - \min_{i \in V} \sum_{t=1}^T \ell_t(i)$. The inherent difficulty of the T-round online learning problem induced by the feedback graph G is measured by the *minimax regret*, denoted by $R(G,T)$ and defined as the minimum over all randomized player strategies, of the maximum over all loss sequences, of the player's expected regret.

2.1. Main Results

The main result of this paper is a complete characterization of the minimax regret when the feedback graph G is fixed and known to the player. Our characterization relies on various properties of G, which we define below.

Definition (Observability)**.** In a directed graph $G = (V,E)$ a vertex $i \in V$ is *observable* if $N^{\mathrm{in}}(i) \neq \emptyset$. A vertex is *strongly observable* if either $\{i\} \subseteq N^{\mathrm{in}}(i)$, or $V \setminus \{i\} \subseteq N^{\mathrm{in}}(i)$, or both. A vertex is *weakly observable* if it is observable but not strongly. A graph G is observable if all its vertices are observable and it is strongly observable if all its vertices are strongly observable. A graph is weakly observable if it is observable but not strongly.

In words, a vertex is observable if it has at least one incoming edge (possibly a self-loop), and it is strongly observable if it has either a self-loop or incoming edges from *all* other vertices. Note that a graph with all of the self-loops is necessarily strongly observable. However, a graph that is missing some of its self-loops may or may not be observable or strongly observable.

Definition (Weak Domination)**.** In a directed graph $G = (V,E)$ with a set of weakly observable vertices $W \subseteq V$, a *weakly dominating set* $D \subseteq V$ is a set of vertices that dominates W. Namely, for any $w \in W$ there exists $d \in D$ such that $w \in N^{\mathrm{out}}(d)$. The *weak domination number* of G, denoted by $\delta(G)$, is the size of the smallest weakly dominating set.

Our characterization also relies on a more standard graph-theoretic quantity. An *independent set* $S \subseteq V$ is a set of vertices that are not connected by any edges. Namely, for any $u, v \in S$, $u \neq v$ it holds that $(u,v) \notin E$. The *independence number* $\alpha(G)$ of G is the size of its largest independent set. Our characterization of the minimax regret rates is given by the following theorem.

Theorem 1. *Let $G = (V,E)$ be a feedback graph with $|V| \geq 2$, fixed and known in advance. Let $\alpha = \alpha(G)$ denote its independence number and let $\delta = \delta(G)$ denote its weak domination number. Then the minimax regret of the T-round online learning problem induced by G, for $T \geq |V|^3$, is*

(i) $R(G,T) = \widetilde{\Theta}(\alpha^{1/2} T^{1/2})$ *if G is strongly observable;*

(ii) $R(G,T) = \widetilde{\Theta}(\delta^{1/3} T^{2/3})$ *if G is weakly observable;*

(iii) $R(G,T) = \Theta(T)$ *if G is not observable.*

Algorithm 1: EXP3.G: online learning with a feedback graph

Parameters: Feedback graph $G = (V, E)$, learning rate $\eta > 0$,
exploration set $U \subseteq V$, exploration rate $\gamma \in [0, 1]$

Let u be the uniform distribution over U;
Initialize q_1 to the uniform distribution over V;
For round $t = 1, 2, \ldots$
Compute $p_t = (1 - \gamma)q_t + \gamma u$;
Draw $I_t \sim p_t$, play I_t and incur loss $\ell_t(I_t)$;
Observe $\{(i, \ell_t(i)) : i \in N^{\mathrm{out}}(I_t)\}$;
Update

$$\forall\, i \in V \qquad \widehat{\ell}_t(i) \;=\; \frac{\ell_t(i)}{P_t(i)}\, \mathbb{I}\{i \in N^{\mathrm{out}}(I_t)\}\,, \qquad \text{with} \qquad P_t(i) \;=\; \sum_{j \in N^{\mathrm{in}}(i)} p_t(j)\,; \tag{1}$$

$$\forall\, i \in V \qquad q_{t+1}(i) \;=\; \frac{q_t(i)\exp(-\eta\widehat{\ell}_t(i))}{\sum_{j \in V} q_t(j)\exp(-\eta\widehat{\ell}_t(j))}\,; \tag{2}$$

As mentioned above, this characterization has some interesting consequences. Any strongly observable graph can be turned into a weakly observable graph by removing at most two edges. Doing so will cause the minimax regret rate to jump from order $\sqrt{T}$ to order $T^{2/3}$. Even more remarkably, removing these edges will cause the minimax regret to switch from depending on the independence number to depending on the weak domination number. A striking example of this abrupt change is the *loopy star* graph, which is the union of the directed star (Fig. 1e) and all of the self-loops (Fig. 1b). In other words, this example is a multi-armed bandit problem with a revealing action. The independence number of this graph is $K - 1$, while its weak domination number is 1. Since the loopy star is strongly observable, it induces a game with minimax regret $\widetilde{\Theta}(\sqrt{TK})$. However, removing a single loop from the feedback graph turns it into a weakly observable graph, and its minimax regret rate changes to $\widetilde{\Theta}(T^{2/3})$ (with no polynomial dependence on K).

3. The EXP3.G Algorithm

The upper bounds for weakly and strongly observable graphs in Theorem 1 are both achieved by an algorithm we introduce, called EXP3.G (see Algorithm 1), which is a variant of the EXP3-SET algorithm for undirected feedback graphs (Alon et al., 2013).

Similarly to EXP3 and EXP3.SET, our algorithm uses importance sampling to construct unbiased loss estimates with controlled variance. Indeed, notice that $P_t(i) = \mathbb{P}(i \in N^{\mathrm{out}}(I_t))$ is simply the probability of observing the loss $\ell_t(i)$ upon playing $I_t \sim p_t$. Hence, $\widehat{\ell}_t(i)$ is an unbiased estimate of the true loss $\ell_t(i)$, and for all t and $i \in V$ we have

$$\mathbb{E}_t[\widehat{\ell}_t(i)] \;=\; \ell_t(i) \quad \text{and} \quad \mathbb{E}_t[\widehat{\ell}_t(i)^2] \;=\; \frac{\ell_t(i)^2}{P_t(i)}\,. \tag{3}$$

The purpose of the exploration distribution u is to control the variance of the loss estimates by providing a lower bound on $P_t(i)$ for those $i \in V$ in the support of u; this ingredient will turn out to be essential to our analysis.

We now state the upper bounds on the regret achieved by Algorithm 1.

Theorem 2. *Let $G = (V, E)$ be a feedback graph with $K = |V|$, independence number $\alpha = \alpha(G)$ and weakly dominating number $\delta = \delta(G)$. Let D be a weakly dominating set such that $|D| = \delta$. The expected regret of Algorithm 1 on the online learning problem induced by G satisfies the following:*

(i) if G is strongly observable, then for $U = V$, $\gamma = \min\{(\frac{1}{\alpha T})^{1/2}, \frac{1}{2}\}$ and $\eta = 2\gamma$, the expected regret against any loss sequence is $\mathcal{O}(\alpha^{1/2} T^{1/2} \ln(KT))$;

(ii) if G is weakly observable and $T \geq K^3 \ln(K)/\delta^2$, then for $U = D$, $\gamma = \min\{(\frac{\delta \ln K}{T})^{1/3}, \frac{1}{2}\}$ and $\eta = \frac{\gamma^2}{\delta}$, the expected regret against any loss sequence is $\mathcal{O}\big((\delta \ln K)^{1/3} T^{2/3}\big)$.

In the previously studied self-aware case (i.e., strongly observable with self-loops), our result matches the bounds of Alon et al. (2014); Kocák et al. (2014). The tightness of our bounds in all cases is discussed in Section 4 below.

Our result in the weakly observable case involves computing a maximal weakly dominating set of the feedback graph and providing it as input to Algorithm 1. We remark that computing a dominating set of maximal size is equivalent to solving a set cover problem, which is NP-hard. Nevertheless, the latter can be efficiently approximated to within a logarithmic factor via a simple greedy algorithm (e.g., Vazirani, 2001), leading to an additional $O(\log K)$ factor in the regret of Algorithm 1 when implemented efficiently.

3.1. A Tight Bound for the Loopless Clique

One of the simplest examples of a feedback graph that is not self-aware is the loopless clique (Fig. 1c). This graph is strongly observable with an independence number of 1, so Theorem 2 guarantees that the regret of Algorithm 1 in the induced game is $\mathcal{O}(\sqrt{T} \ln(KT))$. However, in this case we can do better than Theorem 2 and prove (see Alon et al., 2015) that the regret of the same algorithm is actually $\mathcal{O}(\sqrt{T \ln K})$, which is the same as the regret rate of the full feedback game (Fig. 1a). In other words, if we start with full feedback and then hide the player's own loss, the regret rate remains the same (up to constants).

Theorem 3. *For any sequence of loss functions $\ell_1, \ldots, \ell_T$, where $\ell_t : V \mapsto [0, 1]$, the regret of Algorithm 1, with the loopless clique feedback graph and with parameters $\eta = \sqrt{(\ln K)/(2T)}$ and $\gamma = 2\eta$, is upper-bounded by $5\sqrt{T \ln K}$.*

3.2. Refined Second-order Bound for Hedge

Our analysis of EXP3.G builds on a new second-order regret bound for the classic Hedge algorithm.[2] Recall that Hedge (Freund and Schapire, 1997) operates in the full feedback setting (see Fig. 1a), where at time t the player has access to losses $\ell_s(i)$ for all $s < t$ and $i \in V$. Hedge draws action I_t from the distribution p_t defined by

$$\forall\, i \in V\,, \qquad q_t(i) \;=\; \frac{\exp\big(-\eta \sum_{s=1}^{t-1} \ell_s(i)\big)}{\sum_{j \in V} \exp\big(-\eta \sum_{s=1}^{t-1} \ell_s(j)\big)}\,, \tag{4}$$

where η is a positive learning rate. The following novel regret bound is key to proving that our algorithm achieves tight bounds on the regret (to within logarithmic factors).

2. A second-order regret bound controls the regret with an expression that depends on a quantity akin to the second moment of the losses.

Lemma 4. *Let $q_1, \ldots, q_T$ be the probability vectors defined by Eq. (4) for a sequence of loss functions $\ell_1, \ldots, \ell_T$ such that $\ell_t(i) \geq 0$ for all $t = 1, \ldots, T$ and $i \in V$. For each t, let S_t be a subset of V such that $\ell_t(i) \leq 1/\eta$ for all $i \in S_t$. Then, for any $i^\star \in V$ it holds that*

$$\sum_{t=1}^{T} \sum_{i \in V} q_t(i)\ell_t(i) - \sum_{t=1}^{T} \ell_t(i^\star) \;\leq\; \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \left(\sum_{i \in S_t} q_t(i)\big(1 - q_t(i)\big)\ell_t(i)^2 + \sum_{i \notin S_t} q_t(i)\ell_t(i)^2 \right).$$

See Alon et al. (2015) for a proof of this result. The standard second-order regret bound of Hedge (see, e.g., Cesa-Bianchi et al., 2007) is obtained by setting $S_t = \emptyset$ for all t. Therefore, our bound features a slightly improved dependence (i.e., the $1 - q_t(i)$ factors) on actions whose losses do not exceed $1/\eta$. Indeed, in the analysis of EXP3.G, we apply the above lemma to the loss estimates $\widehat{\ell}_t(i)$, and include in the sets S_t all strongly observable vertices i that do not have a self-loop. This allows us to gain a finer control on the variances $\ell_t(i)^2 / P_t(i)$ of such vertices.

3.3. Proof of Theorem 2

We now turn to prove Theorem 2. For the proof, we need the following graph-theoretic result, which is a variant of Alon et al. (2014, Lemma 16).

Lemma 5. *Let $G = (V, E)$ be a directed graph with $|V| = n$, in which $i \in N^{\mathrm{in}}(i)$ for all vertices $i \in V$. Assign each $i \in V$ a positive weight w_i such that $\sum_{i \in V} w_i \leq 1$ and $w_i \geq \epsilon$ for all $i \in V$ for some constant $0 < \epsilon < \frac{1}{2}$. Then*

$$\sum_{i \in V} \frac{w_i}{\sum_{j \in N^{\mathrm{in}}(i)} w_j} \;\leq\; 4\alpha \ln \frac{4n}{\alpha\epsilon}\,,$$

where $\alpha = \alpha(G)$ is the independence number of G.

The proof of the lemma can be found in Alon et al. (2015). We proceed to prove Theorem 2.

Proof of Theorem 2. Without loss of generality, we may assume that $K \geq 2$. The proof proceeds by applying Lemma 4 and upper bounding the second-order terms it introduces. Indeed, since the distributions $q_1, q_2, \ldots$ generated by Algorithm 1 via Eq. (2) are of the form given by Eq. (4), with the losses ℓ_t replaced by the nonnegative loss estimates $\widehat{\ell}_t$, we may apply Lemma 4 to these distributions and loss estimates. The way we apply the lemma differs between the strongly observable and weakly observable cases, and we treat each separately.

First, assume that G is strongly observable, implying that the exploration distribution u is uniform on V. Notice that for any $i \in V$ without a self-loop, namely with $i \notin N^{\mathrm{in}}(i)$, we have $j \in N^{\mathrm{in}}(i)$ for all $j \neq i$, and so $P_t(i) = 1 - p_t(i)$. On the other hand, by the definition of p_t and since $\eta = 2\gamma$ and $K \geq 2$, we have $p_t(i) = (1-\gamma)q_t(i) + \frac{\gamma}{K} \leq 1 - \gamma + \frac{\gamma}{2} = 1 - \eta$, so that $P_t(i) \geq \eta$. Thus, we can apply Lemma 4 with $S_t = S = \{i : i \notin N^{\mathrm{in}}(i)\}$ to the vectors $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$ and take expectations, and obtain that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in V} q_t(i)\mathbb{E}_t[\widehat{\ell}_t(i)] - \sum_{t=1}^{T} \mathbb{E}_t[\widehat{\ell}_t(i^\star)] \right] \;&\leq\; \frac{\ln K}{\eta} \\
&+ \eta \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \in S} q_t(i)(1 - q_t(i))\mathbb{E}_t[\widehat{\ell}_t(i)^2] + \sum_{i \notin S} q_t(i)\mathbb{E}_t[\widehat{\ell}_t(i)^2] \right]
\end{aligned}$$

for any fixed $i^\star \in V$. Recalling Eq. (3) and $P_t(i) = 1 - p_t(i)$ for all $i \in S$, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V} q_t(i)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(i^\star) \;\le\; \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\in S} q_t(i)\frac{1-q_t(i)}{1-p_t(i)} + \sum_{i\notin S}\frac{q_t(i)}{P_t(i)}\right].$$

The sum over $i \in S$ on the right-hand side is bounded as follows:

$$\sum_{t=1}^{T}\sum_{i\in S} q_t(i)\frac{1-q_t(i)}{1-p_t(i)} \;\le\; 2\sum_{t=1}^{T}\sum_{i\in S} q_t(i) \;\le\; 2T\,.$$

For the second sum, recall that any $i \notin S$ has a self-loop in the feedback graph, and also that $p_t(i) \ge \frac{\gamma}{K}$ as a result of mixing in the uniform distribution over V. Hence, we can use $p_t(i) \ge (1-\gamma)q_t(i) \ge \frac{1}{2}q_t(i)$ and apply Lemma 5 with $\epsilon = \frac{\gamma}{K}$ that yields

$$\sum_{i\notin S}\frac{q_t(i)}{P_t(i)} \;\le\; 2\sum_{i\notin S}\frac{p_t(i)}{P_t(i)} \;\le\; 8\alpha\ln\frac{4K^2}{\alpha\gamma}\,.$$

Putting everything together, and using the fact that $p_t(i) \le q_t(i) + \gamma u(i)$ to obtain

$$\sum_{i\in V} p_t(i)\ell_t(i) \;\le\; \sum_{i\in V} q_t(i)\ell_t(i) + \gamma\,, \tag{5}$$

results in the regret bound

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V} p_t(i)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(i^\star) \;\le\; \gamma T + \frac{\ln K}{\eta} + 2\eta T\left(1 + 4\alpha\ln\frac{4K^2}{\alpha\gamma}\right).$$

Substituting the chosen values of η and γ gives the first claim of the theorem.

Next, assume that G is only weakly observable. Let $D \subseteq V$ be a weakly dominating set supporting the exploration distribution u, with $|D| = \delta$. Similarly to the strongly observable case, we apply Lemma 4 to the vectors $\widehat{\ell}_1, \ldots, \widehat{\ell}_T$, but in this case we set $S_t = \emptyset$ for all t. Using Eqs. (3) and (5) and proceeding exactly as in the strongly observable case, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V} p_t(i)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(i^\star) \;\le\; \gamma T + \frac{\ln K}{\eta} + \eta\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\in V}\frac{q_t(i)}{P_t(i)}\right]$$

for any fixed $i^\star \in V$. In order to bound the expectation in the right-hand side, consider again the set $S = \{i : i \notin N^{\mathrm{in}}(i)\}$ of vertices without a self-loop, and observe that $P_t(i) = \sum_{j\in N^{\mathrm{in}}(i)} p_t(j) \ge \frac{\gamma}{\delta}$ for all $i \in S$. Indeed, if i is weakly observable then there exists some $k \in D$ such that $k \in N^{\mathrm{in}}(i)$ and $p_t(k) \ge \frac{\gamma}{\delta}$ because the exploration distribution u is uniform over D; if i is strongly observable then the same holds since i does not have a self-loop and thus must be dominated by all other vertices in the graph. Hence,

$$\sum_{i\in V}\frac{q_t(i)}{P_t(i)} \;=\; \sum_{i\in S}\frac{q_t(i)}{P_t(i)} + \sum_{i\notin S}\frac{q_t(i)}{P_t(i)} \;\le\; \frac{\delta}{\gamma} + 2K\,,$$

where we used $P_t(i) \geq p_t(i) \geq (1-\gamma)q_t(i) \geq \frac{1}{2}q_t(i)$ to bound the sum over the vertices having a self-loop. Therefore, we may write

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in V} p_t(i)\ell_t(i)\right] - \sum_{t=1}^{T}\ell_t(i^\star) \;\leq\; \gamma T + \frac{\ln K}{\eta} + \frac{\eta\delta}{\gamma}T + 2\eta K T\,.$$

Substituting our choices of η and γ, we obtain the second claim of the theorem. □

4. Lower Bounds

In this section we prove lower bounds on the minimax regret for non-observable and weakly observable graphs. Together with Theorem 2 and the known lower bound of $\Omega\big(\sqrt{\alpha(G)T}\big)$ for strongly observable graphs (Alon et al., 2014, Theorem 5),[3] these results complete the proof of Theorem 1. We remark that their lower bound applies when $T \geq \alpha(G)^3$, which includes our regime of interest. We begin with a simple lower bound for non-observable feedback graphs.

Theorem 6. *If $G = (V, E)$ is not observable and $|V| \geq 2$, then for any player algorithm there exists a sequence of loss functions $\ell_1, \ell_2, \ldots : V \mapsto [0,1]$ such that the player's expected regret is at least $\frac{1}{4}T$.*

The proof is straightforward: if G is not observable, then it is possible to find a vertex of G with no incoming edges; the environment can then set the loss of this vertex to be either 0 or 1 on all rounds of the game, and the player has no way of knowing which is the case. For the formal proof, refer to Alon et al. (2015).

Next, we prove a lower bound for weakly observable feedback graphs.

Theorem 7. *If $G = (V, E)$ is weakly observable with $K = |V| \geq 2$ and weak domination number $\delta = \delta(G)$, then for any randomized player algorithm and for any time horizon T there exists a sequence of loss functions $\ell_1, \ldots, \ell_T : V \mapsto [0,1]$ such that the player's expected regret is at least $\frac{1}{150}\big(\delta/\ln^2 K\big)^{1/3}T^{2/3}$.*

The proof relies on the following graph-theoretic result, relating the notions of domination and independence in directed graphs.

Lemma 8. *Let $G = (V, E)$ be a directed graph over $|V| = n$ vertices, and let $W \subseteq V$ be a set of vertices whose minimal dominating set is of size k. Then, W contains an independent set U of size at least $\frac{1}{50}k/\ln n$, with the property that any vertex of G dominates at most $\ln n$ vertices of U.*

Proof. If $k < 50\ln n$ the statement is vacuous; hence, in what follows we assume $k \geq 50\ln n$. Let $\beta = (2\ln n)/k < 1$. Our first step is to prove that W contains a non-empty set R such that each vertex of G dominates at most β fraction of R, namely such that $|N^{\text{out}}(v)\cap R| \leq \beta|R|$ for all $v \in V$. To prove this, consider the following iterative process: initialize $R = W$, and as long as there exists a vertex $v \in V$ such that $|N^{\text{out}}(v)\cap R| > \beta|R|$, remove all the vertices v dominates from R. Notice

3. While Alon et al. (2014) only consider the special case of graphs that have self-loops at all vertices, their lower bound applies to any strongly observable graph: we can simply add any missing self-loops to the graph, without changing its independence number α; the resulting learning problem, whose minimax regret is $\Omega\big(\sqrt{\alpha T}\big)$, is only easier for the player who may ignore the additional feedback.

that the process cannot continue for k (or more) iterations, since each step the size of R decreases at least by a factor of $1-\beta$, so after $k-1$ steps we have $|R| \le n(1-\beta)^{k-1} < ne^{-\beta k/2} = 1$. On the other hand, the process cannot end with $R = \emptyset$, as in that case the vertices v found along the way form a dominating set of W whose size is less than k, which is a contradiction to our assumption. Hence, the set R at the end of process must be non-empty and satisfy $|N^{\text{out}}(v) \cap R| \le \beta|R|$ for all $v \in V$, as claimed.

Next, consider a random set $S \subseteq R$ formed by picking a multiset $\widetilde{S}$ of $m = \lfloor \frac{1}{10\beta} \rfloor$ elements from R independently and uniformly at random (with replacement), and discarding any repeating elements. Notice that $m \le \frac{1}{10}|R|$, as $|R| \ge \frac{1}{\beta}|N^{\text{out}}(v) \cap R|$ for any $v \in V$, and for some v the right-hand side is non-zero. The proof proceeds via the probabilistic method: we will show that with positive probability, S contains an independence set as required, which would give the theorem.

We first observe the following properties of the set S.

Claim. *With probability at least $\frac{3}{4}$, it holds that $|S| \ge \frac{1}{10}m$.*

To see this, note that each element from R is not included in $\widetilde{S}$ with probability $(1-\frac{1}{r})^m \le e^{-m/r}$ with $r = |R|$. Since $m \le \frac{1}{10}r$, the expected size of S is at least $r(1-e^{-m/r}) = re^{-m/r}(e^{m/r}-1) \ge me^{-m/r} \ge \frac{9}{10}m$, where both inequality use $e^x \ge x+1$. Since always $|S| \le m$, Markov's inequality shows that $|S| \ge \frac{1}{10}m$ with probability at least $\frac{3}{4}$; otherwise, we would have $\mathbb{E}\big[|S|\big] \le \frac{1}{10}m + m\,\mathbb{P}\big(|S| \ge \frac{1}{10}m\big) < \frac{9}{10}m$.

Claim. *With probability at least $\frac{3}{4}$, we have $|N^{\text{out}}(v) \cap S| \le \ln n$ for all $v \in V$.*

Indeed, fix some $v \in V$ and recall that v dominates at most a β fraction of the vertices in R, so each element of $\widetilde{S}$ (that was chosen uniformly at random from R) is dominated by v with probability at most β. Hence, the random variable $\widetilde{X}_v = |N^{\text{out}}(v) \cap \widetilde{S}|$ has a binomial distribution $\mathsf{Bin}(m,p)$ with $p \le \beta$. By a standard binomial tail bound,

$$\mathbb{P}(\widetilde{X}_v \ge \ln n) \;\le\; \binom{m}{\ln n}\beta^{\ln n} \;\le\; (m\beta)^{\ln n} \;\le\; e^{-2\ln n} \;=\; \frac{1}{n^2}\,.$$

The same bound holds also for the random variable $X_v = |N^{\text{out}}(v) \cap S|$, that can only be smaller than $\widetilde{X}_v$. Our claim now follows from a union bound over all $v \in V$.

Claim. *With probability at least $\frac{3}{4}$, we have $\frac{1}{|S|}\sum_{v\in S}|N^{\text{out}}(v) \cap S| \le \frac{1}{2}$.*

To obtain this, we note that for each $v \in V$ the random variable $X_v = |N^{\text{out}}(v) \cap S|$ defined above has $\mathbb{E}[X_v] \le \mathbb{E}[\widetilde{X}_v] \le m\beta \le \frac{1}{10}$, and therefore $\mathbb{E}\big[\frac{1}{|S|}\sum_{v\in S}X_v\big] \le \frac{1}{10}$. By Markov's inequality we then have $\frac{1}{|S|}\sum_{v\in S}X_v > \frac{1}{2}$ with probability less than $\frac{1}{5}$, which gives the claim.

The three claims together imply that there exists a set $S \subseteq W$ of size at least $\frac{1}{10}m$, such that any $v \in V$ dominates at most $\ln n$ vertices of S, and the average degree of the induced undirected graph over S is at most 1. Hence, by Turán's Theorem,[4] S contains an independent set U of size $\frac{1}{20}m \ge \frac{1}{50}k/\ln n$. This concludes the proof, as each $v \in V$ dominates at most $\ln n$ vertices of U. □

Given Lemma 8, the idea of the proof is quite intuitive; here we only give a sketch of the proof, and defer the formal details to Alon et al. (2015).

4. Turán's Theorem (e.g., Alon and Spencer, 2008) states that in any undirected graph whose average degree is d, there is an independent set of size $n/(d+1)$.

Proof of Theorem 7 (sketch). First, we use the lemma to find an independent set U of weakly observable vertices of size $\widetilde{\Omega}(\delta)$, with the crucial property that each vertex in the entire graph dominates at most $\widetilde{\mathcal{O}}(1)$ vertices of U. Then, we embed in the set U a hard instance of the stochastic multiarmed bandit problem, in which the optimal action has expected loss smaller by ϵ than the expected loss of the other actions in U. To all other vertices of the graph, we assign the maximal loss of 1. Hence, unless the player is able to detect the optimal action, his regret cannot be better than $\Omega(\epsilon T)$.

The main observation is that, due to the properties of the set U, in order to obtain accurate estimates of the losses of all actions in U the player has to use $\widetilde{\Omega}(\delta)$ different actions outside of U and pick each for $\Omega(1/\epsilon^2)$ times. Since each such action entails a constant instantaneous regret, the player has to pay an $\Omega(\delta/\epsilon^2)$ penalty in his cumulative regret for exploration. The overall regret is thus of order $\Omega\big(\min\{\epsilon T, \delta/\epsilon^2\}\big)$, which is maximized at $\epsilon = (\delta/T)^{1/3}$ and gives the stated lower bound. □

Acknowledgments

We thank Sébastien Bubeck for helpful discussions during various stages of this work, and Gábor Bartók for clarifying the connections to observability in partial monitoring. Part of this work was done while NCB and TK were visiting OD at Microsoft Research, whose support is gratefully acknowledged.

References

N. Alon and J. H. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2008.

N. Alon, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems 26*, pages 1610–1618. Curran Associates, Inc., 2013.

N. Alon, N. Cesa-Bianchi, C. Gentile, S. Mannor, Y. Mansour, and O. Shamir. Nonstochastic multiarmed bandits with graph-structured feedback. *CoRR*, abs/1409.8428, 2014.

N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren. Online learning with feedback graphs: Beyond bandits. *arXiv preprint arXiv:1502.07617*, 2015.

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.

G. Bartók, D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

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

N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2-3):321–352, 2007.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

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

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*, pages 613–621, 2014.

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

S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 684–692. Curran Associates, Inc., 2011.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

V. V. Vazirani. *Approximation algorithms*. Springer Science & Business Media, 2001.

V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, pages 371–386, 1990.