Online Learning with Feedback Graphs Without the Graphs

Alon Cohen ALON.COHEN@TECHNION.AC.IL
Tamir Hazan TAMIR.HAZAN@TECHNION.AC.IL
Tomer Koren TOMERK@TECHNION.AC.IL
Technion—Israel Institute of Technology, Haifa, Israel

Abstract

We study an online learning framework introduced by Mannor and Shamir (2011) in which the feedback is specified by a graph, in a setting where the graph may vary from round to round and is *never fully revealed* to the learner. We show a large gap between the adversarial and the stochastic cases. In the adversarial case, we prove that even for dense feedback graphs, the learner cannot improve upon a trivial regret bound obtained by ignoring any additional feedback besides her own loss. In contrast, in the stochastic case we give an algorithm that achieves $\widetilde{\Theta}(\sqrt{\alpha T})$ regret over T rounds, provided that the independence numbers of the hidden feedback graphs are at most α. We also extend our results to a more general feedback model, in which the learner does not necessarily observe her own loss, and show that, even in simple cases, concealing the feedback graphs might render a learnable problem unlearnable.

1. Introduction

Online learning is a general framework for sequential decision-making under uncertainty. In its most basic form, it can be described as follows. A learner has to iteratively choose an action from a set of K available actions, and suffer a loss associated with that action. The losses of the actions on each round are assigned in advance by an arbitrary, possibly adversarial, environment. The learner's goal is to minimize her regret over T rounds of the game, which is the difference between her cumulative loss and that of the best fixed action in hindsight.

After making each decision, the learner receives some form of feedback about the losses. Traditionally, the literature considers two types of feedback: full feedback (Littlestone and Warmuth, 1994; Vovk, 1990; Cesa-Bianchi et al., 1997), where the learner observes the losses associated with all of her possible actions, and bandit feedback (Auer et al., 2002b), where the learner only observes the loss of the action she has actually taken.

Full feedback and bandit feedback are special cases of a general framework introduced by Mannor and Shamir (2011), in which the feedback model is specified by a sequence $G_1, \ldots, G_T$ of *feedback graphs*, one for each round t of the game. Each feedback graph G_t is a directed graph whose nodes correspond to the learner's K possible actions; a directed edge $u \to v$ in this graph indicates that whenever the learner chooses action u on round t, in addition to observing the loss of action u, she also gets to observe the loss associated with the action v on that round.

Online learning with feedback graphs was further studied by several authors. Alon et al. (2013), and subsequently Kocák et al. (2014); Alon et al. (2015), gave regret-minimization algorithms that achieve $\widetilde{O}(\sqrt{\alpha T})$ regret, where α is a bound on the *independence numbers* of the graphs $G_1, \ldots, G_T$. Up to logarithmic factors, their results recover and interpolate between the classic bounds of $O(\sqrt{T \log K})$ with full feedback (Freund and Schapire, 1997) and $O(\sqrt{KT})$ with bandit feedback (Auer et al., 2002b; Audibert and Bubeck, 2009). The $\widetilde{O}(\sqrt{\alpha T})$ bound turns out to be tight for *any* feedback graph (when it is fixed throughout the game and known in advance), in light of a matching lower bound due to Mannor and Shamir (2011).

However, all of the optimal algorithms mentioned above require the full structure of the feedback graph in order to operate. While some require the entire graph G_t for performing their updates only at the end of round t (e.g., Alon et al., 2013; Kocák et al., 2014; Alon et al., 2015),[1] others actually need the description of G_t at the beginning of the round before making their decision (e.g., Alon et al.,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[1] More precisely, these algorithms do not need the entire graph but rather the incoming neighborhood of each of the actions for which the associated loss has been observed.

2014). In fact, none of the algorithms previously proposed in the literature is able to provide non-trivial regret guarantees without the feedback graphs being disclosed.

The assumption that the entire observation system is revealed to the learner on each round, even if only after making her prediction, is rather unnatural. In principle, the learner need not be even aware of the fact that there is a graph underlying the feedback model; the feedback graph is merely a technical notion for us to specify a set of observations for each of the possible actions. Ideally, the only signal we would like the learner to receive following each round is the set of observations that corresponds to the action she has taken on that round (in addition to her own loss).

As a motivating example for situations where receiving the entire observation system is unrealistic, consider the following online pricing problem that faces any vendor selling goods over the internet. On each round, the seller has to announce a price for his product. Then, a buyer arrives and decides whether or not to purchase the product at this price based on his private value; the only feedback the seller receives is whether or not the buyer purchased the product at the announced price. However, when a purchase takes place, the seller also knows that the buyer would have bought the product at any price lower than the price that she announced. While this feedback structure can be thought of as a directed graph over the seller's actions (i.e., prices), the graph itself is never fully revealed to the seller as its structure discloses the buyer's private value.

1.1. Our contributions

In this paper, we study online learning with feedback graphs in a setting where the feedback graphs are *never revealed* to the learner in their entirety. That is, in this setting the only feedback available to the learner at the end of round t is the out-neighborhood of her chosen action in the graph G_t, along with the loss associated with each of the actions in this neighborhood and the loss of the action that she chose. We address the following questions: how this lack of full disclosure affects the learner's regret? Is it possible to achieve any non-trivial regret guarantee in this setting, i.e., one that improves on the trivial $O(\sqrt{KT})$ bound? In particular, can we obtain bounds that scale with the independence numbers of the feedback graphs?

Our main results show that not knowing the entire feedback graphs can have a significant impact on the learner's achievable regret. First, we show that in a standard adversarial online learning setting, where we assume nothing about the process generating the losses and the feedback graphs (i.e., both are possibly chosen by an adversary), any strategy of the learner must suffer $\Omega(\sqrt{KT})$ regret in the worst case, even if the independence numbers of $G_1, \ldots, G_T$ are all bounded by some small absolute constant. Namely, by hiding the feedback graphs from the learner, the problem surprisingly becomes as hard as the K-armed bandit problem, even when the feedback available to the learner is "almost full": each of the feedback graphs is "almost a clique." In other words, the side observations received by the learner are effectively useless; she may as well ignore them and use a standard bandit algorithm such as EXP3 (Auer et al., 2002b) to perform optimally.

Second, and in contrast to the adversarial setting, we show that in a stochastic setting where the losses of each action are known to be drawn i.i.d. from some unknown probability distribution, side observations can still be very useful. We show that the learner is able to achieve an optimal regret bound of the form $\widetilde{O}(\sqrt{\alpha T})$, even if the graphs $G_1, \ldots, G_T$ are chosen adversarially and are never fully revealed to the learner, as long as their independence numbers are all bounded by α. We give an efficient elimination-based algorithm achieving this bound, that does not require knowing the value of α in advance. This result is optimal up to logarithmic factors, even when the feedback graph is fixed throughout the game and known in advance, due to a lower bound of Mannor and Shamir (2011).

For our algorithm in the stochastic case, we also prove a distribution-dependent regret bound that scales logarithmically with T. The bound we prove is of the form $O(\sum_{v \in V'}(1/\Delta_v) \log T)$, where Δ_v is the gap of action v, and V' is the subset of $\widetilde{O}(\alpha)$ actions with smallest gaps. This bound is a substantial improvement over standard regret bounds of stochastic multi-armed bandit algorithms such as UCB (Auer et al., 2002a): whereas the regret of the latter algorithms is typically bounded by a sum $\sum_{v \in V}(1/\Delta_v)$ taken over *all* K actions, the sum in our bound is taken only over the subset of $\widetilde{O}(\alpha)$ actions with the smallest gaps. This result cannot be improved even when the feedback graph is fixed throughout the game, and has an optimal dependence on α as well as on the gaps Δ_v, thus resolving an open question of Alon et al. (2014).

Finally, we extend our results to a more general feedback model recently studied by Alon et al. (2015), in which the learner does not necessarily observe her own loss after making predictions (namely, each action may or may not have a self-loop in each feedback graph). Alon et al. (2015) gave a necessary and sufficient condition for attaining $\Theta(\sqrt{T})$ regret in this more general model—a graph-theoretic condition they call strong observability. The extension of our results to their model bears some surprising consequences: even in the strongly observable case with only two actions, not revealing the entire feedback graphs to the learner might make the problem unlearnable! Nevertheless, in the case of stochastic losses, our positive results do extend to the more general feedback model.

1.2. Additional related work

Online learning with feedback graphs was previously considered in the stochastic setting by Caron et al. (2012), who gave results depending on the graph clique structure. Their analysis, however, only applies when the feedback graph is fixed throughout the game, and can only bound the regret in terms of a quantity akin to the clique-partition number of this graph, which is always larger than its independence number (the gap between the two can be very large; see Alon et al., 2014).

More recently, Wu et al. (2015) and Kocák et al. (2016) have investigated a noisy version of the feedback graph model, where feedback is specified by a weighted directed graph with edge weights indicating the quality (e.g., the noise level or variance) of the feedback received on adjacent vertices. Wu et al. (2015) provided finite-time problem-dependent lower bounds for this setting; Kocák et al. (2016) generalized the notion of independence number to the noisy case and gave new efficient algorithms in this setting.

2. Setup and Main Results

We consider a general online learning model with graph-structured feedback, which can be described as a game between a learner and an environment that proceeds for T rounds. Before the game begins, the environment privately determines a sequence of loss functions $\ell_1, ..., \ell_T : V \mapsto [0,1]$ defined over a set $V = \{1, ..., K\}$ of K actions, which we view as a sequence of loss vectors $\ell_1, ..., \ell_T \in [0,1]^K$. In addition, the environment fixes a sequence of directed graphs $G_1, \ldots, G_T$ over V as vertices.

We will consider two different cases, that we refer to as the adversarial setting and the stochastic setting:

- In the adversarial setting, the loss vectors $\ell_1, ..., \ell_T$ and the feedback graphs $G_1, \ldots, G_T$ are chosen by the environment in an arbitrary way.
- In the stochastic setting, the environment privately selects a loss distribution $\mathcal{D}$ over $[0,1]^K$ and an arbitrary sequence $G_1, \ldots, G_T$; thereafter, the loss vectors $\ell_1, \ldots, \ell_T$ are sampled i.i.d. from $\mathcal{D}$.

Iteratively on rounds $t = 1, 2, ..., T$, the learner randomly chooses an action $v_t \in V$ and incurs the loss $\ell_t(v_t)$. At the end of each round t, the learner receives a feedback comprised of $\{(v, \ell_t(v)) : (v_t \to v) \in G_t)\}$, that includes the loss $\ell_t(v_t)$ incurred by the learner (i.e., we assume that $(v \to v) \in G_t$ for all t and $v \in V$). In words, the learner observes the losses associated with v_t and the actions in the out-neighborhood of v_t in the feedback graph G_t. The feedback graph G_t itself *is never revealed* in its entirety to the learner.

The goal of the learner throughout the T rounds of the game is to minimize her expected regret, which is defined as

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(v_t) - \sum_{t=1}^{T} \ell_t(v^\star)\right], \qquad (1)$$

where $v^\star = \min_{v \in V} \mathbb{E}[\sum_{t=1}^{T} \ell_t(v)]$ is the best action in hindsight. Here, the expectations are taken over the random choices of the learner and, in the stochastic setting, also over the randomness of the losses.

For the stochastic setting we require additional notation. For each $v \in V$, we denote by $\mu(v)$ the mean of the loss of action v under $\mathcal{D}$. We denote $\mu^\star = \mu(v^\star)$, and let $\Delta_v = \mu(v) - \mu^\star$ for all $v \in V$. We refer to Δ_v as the *gap* of action v, and assume for simplicity that $v^\star$ is unique so that $\Delta_v > 0$ for all $v \neq v^\star$.

For stating our results, we need a standard graph-theoretic definition. An *independent set* in a graph $G = (V, E)$ (either directed or undirected) is a set of vertices that are not connected by any edges. Namely, $S \subseteq V$ is independent if for any $u, v \in S$, $u \neq v$, it holds that $(u \to v) \notin E$. The *independence number* $\alpha(G)$ of G is the size of the largest independent set in G.

2.1. Main results

We now state the main results of this paper. Our first result deals with the adversarial case and shows that when the feedback graphs are not revealed to the learner at the end of each round, her regret might be very large even when the independence numbers of the graphs are small—they are all bounded by a constant.

Theorem 1. *In the adversarial setting, any online learning algorithm must suffer at least $\Omega(\sqrt{KT})$ regret in the worst case, even when all feedback graphs $G_1, \ldots, G_T$ have independence numbers $\leq O(1)$.*

The lower bound in the theorem is tight: it can be matched by simply running a standard bandit algorithm (e.g., EXP3 of Auer et al., 2002b), ignoring all observed feedback besides the loss of the action played.

Our next result shows that in the stochastic case, the learner is still able to attain non-trivial regret despite the fact that the feedback graphs are never fully revealed to her.

Theorem 2. *In the stochastic setting, Algorithm 1 described in Section 4 attains an expected regret of at most $\widetilde{O}(\sqrt{\alpha T})$, provided that the independence numbers of the feedback graphs $G_1, \ldots, G_T$ are all bounded by α.*

This regret bound is optimal up to logarithmic factors, since the lower bound of $\Omega(\sqrt{\alpha T})$ found in Mannor and Shamir (2011) applies in our stochastic setting.

In the stochastic setting we also give a distribution-dependent analysis of Algorithm 1 which depends on the gaps of the actions under the distribution $\mathcal{D}$.

Theorem 3. *In the stochastic setting, Algorithm 1 described in Section 4 attains an expected regret of*

$$O\left(\sum_{v \in V'} \frac{1}{\Delta_v} \log T\right),$$

where V' is the set of $\widetilde{O}(\alpha)$ actions with the smallest gaps (excluding $v^\star$), provided that the the independence numbers of the graphs $G_1, \ldots, G_T$ are all bounded by α.

We also extend our results to a more general class of feedback graphs, in which each vertex may or may not have a self-loop. For the statements of these additional results, see the full version of the paper (Cohen et al., 2016).

2.2. Discussion of the results

Our results show that there is a large gap between the achievable regret rates in the adversarial and stochastic settings, in terms of the dependence on the properties of the feedback graphs.

In the adversarial case, the environment is free to simultaneously choose the sequences of loss values and feedback graphs in conjunction with each other; for example, they can be drawn from a *joint* distribution over sequences of loss values *and* sequences of directed graphs. The environment may use this freedom to manipulate the feedback observed by the learner and bias her observations in a malicious way. In the stochastic setting, on the other hand, the loss values are drawn from the underlying distribution only after the environment commits to some arbitrary sequence of graphs, so that the feedback graphs are *probabilistically independent* of the realizations of the losses.

In fact, as our arguments in Section 3 reveal, there exists a randomized construction of loss vectors and feedback graphs that inflicts $\Omega(\sqrt{KT})$ on any learner, in which the loss vectors are i.i.d. However, the stochastic process that generates the feedback graphs in that construction is correlated with the actual realizations of the i.i.d. losses. This is a crucial aspect of our construction, as implied by our upper bound in the stochastic case.

3. Lower Bound for Adversarial Losses

In this section we deal with the adversarial setting and prove Theorem 1: we show an $\Omega(\sqrt{KT})$ lower bound on the performance of any online learning algorithm, where both the losses of the actions and the feedback graphs can be chosen arbitrarily. In fact, our result relies heavily on the fact that the two can be behave in a correlated manner.

	$u \not\to v^\star$	$u \to v^\star$	
$\ell_t(v^\star) = 0$	2ϵ	$\frac{1}{2} - \epsilon$	$\frac{1}{2} + \epsilon$
$\ell_t(v^\star) = 1$	0	$\frac{1}{2} - \epsilon$	$\frac{1}{2} - \epsilon$
	2ϵ	$1 - 2\epsilon$	

	$u \not\to v$	$u \to v$	
$\ell_t(v) = 0$	ϵ	$\frac{1}{2} - \epsilon$	$\frac{1}{2}$
$\ell_t(v) = 1$	ϵ	$\frac{1}{2} - \epsilon$	$\frac{1}{2}$
	2ϵ	$1 - 2\epsilon$	

Figure 1. Summary of the joint distribution of the loss of action $v^\star$ and an edge between u and $v^\star$ (top), and of the joint distribution of the loss of action $v \neq v^\star$ and an edge between u and v (bottom). The grayed-out entries indicate probabilities that cannot be estimated by the learner; the remaining entries do not permit the learner to distinguish between $v^\star$ and v.

Let us sketch the idea behind the lower bound. By Yao's minimax principle, in order to prove a lower bound on the learner's regret it is enough to demonstrate a randomized strategy for the environment that forces any deterministic learner to incur $\Omega(\sqrt{KT})$ regret. We construct our environment's strategy as follows.

First, before the game begins, the environment chooses an action $v^\star$ uniformly at random from V. At each round, the loss of all actions $v \neq v^\star$ is distributed Bernoulli(1/2), while the loss of action $v^\star$ is distributed Bernoulli$(1/2 - \epsilon)$ with $\epsilon = (1/8)\sqrt{K/T}$. All of the loss values in the construction are drawn independently of each other.

The feedback graphs $G_1, \ldots, G_T$ are chosen i.i.d. from the following distribution. Any edge $u \to v$ for $v \neq v^\star$ appears with probability $1 - 2\epsilon$ independently from all other edges and the losses of the actions. Edges of the form $u \to v^\star$ appear mutually independently given the loss of action $v^\star$: if the loss of $v^\star$ is 1, each edge appears with probability 1; if the loss of $v^\star$ is 0, each edge appears with probability $(1-2\epsilon)/(1+2\epsilon)$. See Figure 1 for a summary of the edge probabilities in this construction. The idea behind the construction is as following. Suppose that the learner plays some action $u \neq v^\star$, the distributions of the observed losses of every other actions are identical, including that of $v^\star$. In other words, her only option of finding $v^\star$ is by sampling it directly and observing its loss. Hence, the construction is capable of simulating a K-armed bandit problem whose minimax regret is $\Omega(\sqrt{KT})$.

For the construction above we prove the following theorem.

Theorem 4. *Assume that $K \geq 2$ and $T \geq K^2$. Any deterministic learner must suffer an expected regret of at least $(1/32)\sqrt{KT}$ against the environment constructed above.*

To prove the theorem, we shall need a few definitions. Let $\mathbb{P}, \mathbb{Q}$ be a couple of distributions over the same space and sigma-algebra $\mathcal{F}$. We define the *total variation distance* between $\mathbb{P}$ and $\mathbb{Q}$ as $D_{\mathrm{TV}}(\mathbb{P}\,, \mathbb{Q}) = \sup_{E \in \mathcal{F}} |\mathbb{P}[E] - \mathbb{Q}[E]|$. If $\mathbb{P}$ and $\mathbb{Q}$ are discrete distributions, we define the KL divergence between $\mathbb{P}$ and $\mathbb{Q}$ as $D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \sum_x \log\left(\mathbb{P}[x]/\mathbb{Q}[x]\right)\mathbb{P}[x]$ assuming the support of $\mathbb{P}$ is contained in that of $\mathbb{Q}$, and where the sum is taken over the support of $\mathbb{P}$. We can now turn to the proof of the theorem.

Proof of Theorem 4. Let us introduce the random variables T_v whose value is the number of times the learner plays action v. We also introduce the notations $\mathbb{P}_v$ and $\mathbb{E}_v$ indicating probability and expectation with respect to the marginal distributions under which $v^\star = v$. Then, we have

$$\begin{aligned}
\mathrm{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(v_t) - \sum_{t=1}^{T} \ell_t(v^\star)\right] \\
&= \frac{1}{K}\sum_{v \in V} \mathbb{E}_v\left[\sum_{t=1}^{T} \ell_t(v_t) - \sum_{t=1}^{T} \ell_t(v)\right] \\
&= \frac{1}{K}\sum_{v \in V} \epsilon \cdot \mathbb{E}_v[T - T_v] \\
&= \epsilon\left(T - \frac{1}{K}\sum_{v \in V} \mathbb{E}_v[T_v]\right)\,, \qquad (2)
\end{aligned}$$

and in order to proceed we shall upper bound $\mathbb{E}_v[T_v]$.

Introduce a new distribution, in which the losses of the actions are independent Bernoulli(1/2) variables, and the feedback graphs are such that each directed edge appears with probability $1 - 2\epsilon$ independently of the other edges and the losses of the actions. We will refer to this new law using $\mathbb{P}_0$ and $\mathbb{E}_0$. Let λ_t be the losses and edges observed at time t, and similarly $\lambda^{(t)} = (\lambda_1, ..., \lambda_t)$ are the losses and edges observed up until time t (inclusive). Then, since the sequence $\lambda^{(T)}$ determines the actions of the learner over the entire game, and by Pinsker's inequality,

$$\begin{aligned}
\mathbb{E}_v[T_v] - \mathbb{E}_0[T_v] &\le T \cdot D_{\mathrm{TV}}\Big(\mathbb{P}_v[\lambda^{(T)}]\,, \mathbb{P}_0[\lambda^{(T)}]\Big) \\
&\le T\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}\big(\mathbb{P}_0[\lambda^{(T)}] \,\big\|\, \mathbb{P}_v[\lambda^{(T)}]\big)}. \qquad (3)
\end{aligned}$$

Moreover, by the chain rule of KL-divergence, $D_{\mathrm{KL}}\big(\mathbb{P}_0[\lambda^{(T)}] \,\|\, \mathbb{P}_v[\lambda^{(T)}]\big)$ equals

$$\sum_{t=1}^{T} \sum_{\lambda^{(t-1)}} \mathbb{P}_0[\lambda^{(t-1)}] D_{\mathrm{KL}}\Big(\mathbb{P}_0[\lambda_t | \lambda^{(t-1)}] \,\Big\|\, \mathbb{P}_v[\lambda_t | \lambda^{(t-1)}]\Big)\,. \qquad (4)$$

Consider a single term in the sum. Recall that $\lambda^{(t-1)}$ determines the action v_t chosen by the learner on round t. If $v_t \neq v$ then, by our construction, the losses and edges of the graph observed by the learner are distributed exactly the same under $\mathbb{P}_v$ and $\mathbb{P}_0$, and the KL divergence is 0. If $v_t = v$ then the losses of all other actions are distributed Bernoulli(1/2), and independently of the loss of action v and the observed edges. The latter is so under both $\mathbb{P}_v$ and $\mathbb{P}_0$. Moreover, the observed edges are distributed Bernoulli$(1 - 2\epsilon)$ independently of the loss of action v under both $\mathbb{P}_v$ and $\mathbb{P}_0$. Namely, the only element that is distributed differently under $\mathbb{P}_v$ and $\mathbb{P}_0$ is the loss of action v, and the latter is distributed independently from all other observed variables. Recall that the loss of action v is distributed as Bernoulli(1/2) under $\mathbb{P}_0$ and as Bernoulli$(1/2 - \epsilon)$ under $\mathbb{P}_v$. Therefore, $D_{\mathrm{KL}}\big(\mathbb{P}_0[\lambda_t|\lambda^{(t-1)}] \,\|\, \mathbb{P}_v[\lambda_t|\lambda^{(t-1)}]\big)$ is upper-bounded by

$$D_{\mathrm{KL}}\left(\frac{1}{2} \,\middle\|\, \frac{1}{2} - \epsilon\right) = -\frac{1}{2}\log(1 - 4\epsilon^2) \le 4\epsilon^2\,,$$

where the last inequality holds since $\epsilon < 1/4$ by assumption. Plugging the above back into Eq. (4),

$$\begin{aligned}
D_{\mathrm{KL}}\Big(\mathbb{P}_0[\lambda^{(T)}] \,\Big\|\, \mathbb{P}_v[\lambda^{(T)}]\Big) &\le \sum_{t=1}^{T} \mathbb{P}_0[v_t = v] 4\epsilon^2 \\
&= 4\epsilon^2 \mathbb{E}_0[T_v]\,,
\end{aligned}$$

and the latter into Eq. (3), we get that $\mathbb{E}_v[T_v] \le \mathbb{E}_0[T_v] + T\epsilon\sqrt{2\mathbb{E}_0[T_v]}$.

Now, $K \ge 2$ by assumption, and therefore

$$\begin{aligned}
\frac{1}{K}\sum_{v \in V} \mathbb{E}_v[T_v] &\le \frac{1}{K}\sum_{v \in V} \mathbb{E}_0[T_v] + \frac{1}{K}\sum_{v \in V} T\epsilon\sqrt{2\mathbb{E}_0[T_v]} \\
&\le \frac{1}{K}\sum_{v \in V} \mathbb{E}_0[T_v] + T\epsilon\sqrt{\frac{1}{K}\sum_{v \in V} 2\mathbb{E}_0[T_v]} \\
&= \frac{T}{K} + T\epsilon\sqrt{\frac{2T}{K}} \\
&\le \frac{T}{2} + T\epsilon\sqrt{\frac{2T}{K}}\,.
\end{aligned}$$

Let us now return to Eq. (2). We can lower bound the regret as

$$\mathrm{R}_T \ge \epsilon\left(T - \frac{T}{2} - T\epsilon\sqrt{\frac{2T}{K}}\right) = \epsilon T\left(\frac{1}{2} - \epsilon\sqrt{\frac{2T}{K}}\right)\,.$$

By our choice of ϵ, we have that $\epsilon\sqrt{2T/K}$ is at most $1/4$, and so

$$\mathrm{R}_T \ge \frac{T}{8}\sqrt{\frac{K}{T}}\left(\frac{1}{2} - \frac{1}{4}\right) = \frac{1}{32}\sqrt{KT}\,,$$

as claimed. □

To show that Theorem 1 holds, we need to show that the learner suffers a large regret against an environment that selects feedback graphs with constant independence numbers. While the independence numbers of the graphs that we have constructed might, in principle, be large, we can show that with very high probability they are uniformly bounded by a constant.

Lemma 5. *Suppose that $|V| = K \geq 2$ and $T \geq K^2$. Let $G_1, ..., G_T$ be a sequence of graphs as constructed above. With probability at least $1-\epsilon/8$, the independence numbers of all graphs are at most 9.*

Theorem 1 now follows by combining Theorem 4 and Lemma 5; for technical details, see Cohen et al. (2016).

4. Algorithms for Stochastic Losses

In this section we present and analyze our algorithm for the stochastic setting. The algorithm, given in Algorithm 1, is reminiscent of elimination-based algorithms for the stochastic multi-armed bandit problem (e.g., Even-Dar et al., 2002; Karnin et al., 2013). For this algorithm, we prove the following guarantee on the expected regret, which implies Theorem 2.

Theorem 6. *Assume that $K \geq 2$. Suppose that Algorithm 1 is run on a sequence of feedback graphs with independence numbers $\leq \alpha$. Then the expected regret of the algorithm is at most $\widetilde{O}(\sqrt{\alpha T})$.*

Algorithm 1 works in phases $r = 1, 2, \ldots$. It maintains a subset of actions V_r, where initially $V_1 = V$. At each phase r, the algorithm estimates the mean losses of all actions in V_r to within ϵ_r accuracy, by invoking a procedure called AlphaSample n_r times. It then filters out from V_r the actions that are known to be $2\epsilon_r$-suboptimal with sufficient confidence, and repeats this process, decreasing the accuracy parameter ϵ_r after each phase.

The key for achieving optimal regret lies in the the procedure AlphaSample, that appears as Algorithm 2. Each call to this procedure allows us to observe the losses of all actions in V_r once, while spending only $\widetilde{O}(\alpha)$ rounds in expectation. The exact details of AlphaSample are discussed in Section 4.2 below, and here we just state its guarantee.

Lemma 7. AlphaSample *returns one sample of the loss of each action in V_r and terminates after at most $10\alpha \log K$ rounds of the game in expectation, provided that the independence numbers of all feedback graphs $G_1, \ldots, G_T$ are at most α.*

To prove Theorem 6 we need one additional lemma. It shows that, at each phase, the elimination procedure of the algorithm succeeds with high probability. Namely, after phase r, the algorithm is left with actions that are at most $4\epsilon_r$-suboptimal.

Algorithm 1

```
input Set $V$ of $K$ actions, number of rounds $T$
initialize $r \leftarrow 1$, $V_1 = V$, $\epsilon_1 = 1/4$
while $|V_r| > 1$ and $T$ rounds have not elapsed do
  Set $n_r = \lceil 2\log(2KT)/\epsilon_r^2 \rceil$
  Invoke AlphaSample($V_r$) for $n_r$ times, and
    compute empirical mean $m_r(v)$ of each action
    $v \in V_r$ using collected samples
  Compute $m_r^\star = \min_{v \in V_r} m_r(v)$
  Eliminate actions:
    $V_{r+1} = \{v \in V_r : m_r(v) \leq m_r^\star + 2\epsilon_r\}$
  Set $\epsilon_{r+1} = \epsilon_r/2$, $r \leftarrow r+1$
end while
Play the action left in $V_r$ until $T$ rounds have passed
```

Lemma 8. *For all r, with probability at least $1 - 1/T$ we have $\mu(v) \leq \mu^\star + 4\epsilon_r$ for all $v \in V_{r+1}$.*

We can now proceed with the proof of the theorem.

Proof of Theorem 6. Let us start by bounding the number of phases R the algorithm makes. Let the random variable T_r denote the number of game rounds elapsed during phase r. Since the algorithm runs for T rounds we must have that

$$\sum_{r=1}^{R} T_r \leq T \,. \tag{5}$$

In particular, since AlphaSample takes at least one round to complete, we have that $T_r \geq n_r \geq 2\log(2KT)4^{r+1}$ and we get the crude bound of

$$R \leq \bar{r} = \frac{1}{2}\log_2\left(\frac{3T}{32\log(2KT)} + 1\right) \,. \tag{6}$$

We turn to bound the expected regret of the algorithm. By Lemma 8 and the union bound, the total probability of failure of the mean estimations is at most $\bar{r}/T$. Then the expected regret of the algorithm is at most the expected regret conditioned on the success of the estimation of the means plus $(\bar{r}/T) \cdot T = \bar{r} = O(\log T)$ by Eq. (6), and since the regret is bounded by T with probability 1. Thus it remains to bound the regret conditioned on the success of the mean estimations.

For convenience, define $\epsilon_0 = 1/2$. On phase r, by Lemma 8 we have an instantaneous regret of at most $4\epsilon_{r-1} = 8\epsilon_r$ per round. If only one action is left in V_r then it must be $v^\star$ and therefore after the final phase the algorithm suffers zero instantaneous expected regret. Overall, the expected regret is at most

$$\mathbb{E}\left[\sum_{r=1}^{R} T_r \cdot 8\epsilon_r\right] \le 8\sqrt{\mathbb{E}\left[\sum_{r=1}^{R} T_r\right]} \cdot \sqrt{\mathbb{E}\left[\sum_{r=1}^{R} T_r \epsilon_r^2\right]}$$

by the Cauchy-Schwartz inequality. Note that $\sum_{r=1}^{R} T_r \le T$ by Eq. (5). Additionally, by Lemma 7 each call to AlphaSample spends at most $m = 10\alpha \log K$ rounds in expectation and thus $\mathbb{E}[T_r] \le m n_r$. Hence,

$$\mathbb{E}\left[\sum_{r=1}^{R} T_r \epsilon_r^2\right] \le \sum_{r=1}^{\bar{r}} m n_r \epsilon_r^2 \le m\bar{r}(2\log(2KT) + 1) \,.$$

The first inequality holds since the number of phases is at most $\bar{r}$. The right-hand side is $O(\alpha \log(K) \log^2(KT))$ by Eq. (6) and the definition of m. □

4.1. Gap-based analysis

We can also provide a distribution-dependent analysis of Algorithm 1 that yields a logarithmic regret bound, albeit with an explicit dependence on the gaps Δ_v.

Denote by $V^{(n)}$ the set of n actions with smallest gaps, excluding $v^\star$ and breaking ties arbitrarily. Our main result in this section is the following theorem, which gives Theorem 3. Recall that we assume $v^\star$ is the unique optimal action, and so the gaps of all other actions are positive.

Theorem 9. *Suppose that $K \ge 2$ and $T \ge K$, and let $\tau = \lceil 10\alpha \log K \rceil$. Suppose that Algorithm 1 is run on a sequence of feedback graphs with independence numbers $\le \alpha$. Then the expected regret of Algorithm 1 is at most*

$$O\left(\sum_{v \in V^{(\tau)}} \frac{1}{\Delta_v} \log T\right) .$$

We can explain the intuition behind the bound as follows. Each call to AlphaSample spends at most $\widetilde{O}(\alpha)$ rounds while producing samples of all K actions. Thus, in the worst case, after a quick pruning phase the algorithm is left with the "hardest" $\tau = \widetilde{O}(\alpha)$ actions and has to tell them apart; in this last phase, the additional observations provided by the feedback graphs might not help the algorithm at all (e.g., the remaining τ actions might form an independent set in all graphs). We now turn to the proof of the theorem.

Proof of Theorem 9. As in the proof of Theorem 6, we have that the expected regret of the algorithm is at most the expected regret conditioned on the success of the mean estimations plus $O(\log T)$, and thus it remains to bound the regret conditioned on the success of the mean estimations. Conditioned on the success of the algorithm, the regret of the algorithm is at most the regret of an algorithm that has finished running with $V_r = \{v^\star\}$. Thus we can assume that T is large enough for that to happen.

If $\tau \le K - 1$, we begin by bounding the regret until the algorithm eliminates all actions besides the ones in $V^{(\tau)}$. Let $\bar{\Delta}$ be the largest gap of an action from $V^{(\tau)}$. Let $\bar{r} = \lfloor \log_2(2/\bar{\Delta}) \rfloor$. Thus, it takes $\bar{r} + 1$ phases in order for ϵ_r to be less than $\bar{\Delta}/4$. The regret up to round $\bar{r}$ is bounded using the following lemma.

Lemma 10. *Let $m = 10\alpha \log K$. The expected regret of Algorithm 1 up to round $\bar{r}$ is at most $(128m/\bar{\Delta}) \log(2KT)$.*

We proceed with the analysis of the expected regret after phase $\bar{r}$. This is given by this next lemma.

Lemma 11. *The expected regret of Algorithm 1 from round $\bar{r} + 1$ until the end of the game is at most $\sum_{v \in V^{(\tau)}} (128/\Delta_v) \log(2KT)$.*

If $\tau > K - 1$ then the regret of the algorithm is given by Lemma 11. Otherwise, the proof of the theorem is completed by noticing that the regret of the algorithm up to round $\bar{r}$ is at most the regret from round $\bar{r} + 1$ thereafter. Since $\bar{\Delta} \ge \Delta_v$ for all $v \in V^{(\tau)}$ we get that

$$\frac{m}{\bar{\Delta}} \le \frac{m}{|V^{(\tau)}|} \sum_{v \in V^{(\tau)}} \frac{1}{\Delta_v} \le \sum_{v \in V^{(\tau)}} \frac{1}{\Delta_v} ,$$

by definition of m and $V^{(\tau)}$. This in total gives a regret bound of $O(\sum_{v \in V^{(\tau)}} \Delta_v^{-1} \log(KT))$. Finally, we use our assumption that $T \ge K$ to simplify the bound. □

Proof of Lemma 10. By Lemma 7, each call to AlphaSample spends at most m rounds in expectation. By Lemma 8, the instantaneous regret for each round on phase r is at most $4\epsilon_{r-1} = 8\epsilon_r$. Then the expected regret up to round $\bar{r}$ is at most $\sum_{r=1}^{\bar{r}} m \cdot n_r \cdot 8\epsilon_r \le 32m \log(2KT) \sum_{r=1}^{\bar{r}} \epsilon_r^{-1}$, and we have

$$\sum_{r=1}^{\bar{r}} \frac{1}{\epsilon_r} = \sum_{r=1}^{\bar{r}} 2^{r+1} \le 2^{\bar{r}+2} \le \frac{4}{\bar{\Delta}} \,.$$

□

Proof of Lemma 11. Let us denote $\bar{r}_v = \lfloor \log_2(2/\Delta_v) \rfloor$, the number of phases until v is removed from V_r. Let w be the action with the minimum nonzero gap. We shall assume that the game is finished after $\bar{r}_w$ phases.

Note that after we have eliminated all actions not in $V^{(\tau)}$, each call to AlphaSample is finished after at most $|V_r|$ steps. Thus, the expected regret for the remaining phases is at most

$$\sum_{r=\bar{r}+1}^{\bar{r}_w} \frac{32\log(2KT)}{\epsilon_r} |V_r| = 32\log(2KT) \sum_{v \in V^{(\tau)}} \sum_{r=\bar{r}+1}^{\bar{r}_v} \frac{1}{\epsilon_r} ,$$

and for all $v \in V^{(\tau)}$, $\sum_{r=\bar{r}+1}^{\bar{r}_v} \epsilon_r^{-1} \leq \sum_{r=0}^{\bar{r}_v} 2^{r+1}$, which in turn equals $2^{\bar{r}_v+2} \leq 4/\Delta_v$. □

4.2. Efficient sampling scheme

In this section, we discuss the AlphaSample randomized sampling procedure. This procedure allows us to collect one sample of the loss for each action while spending only $\widetilde{O}(\alpha)$ rounds in expectation. AlphaSample is described in Algorithm 2.

Let us now explain the intuition behind the procedure. At each round, the procedure samples the loss of an action uniformly at random from a subset of actions U. As each sample is uniform over U, the procedure observes the losses of $\Omega(|U|/\alpha)$ actions in expectation. The actions that have been observed are then removed from U and the process continues recursively until U is empty. This phase is complete after an expected $\widetilde{O}(\alpha)$ rounds.

The main result regarding AlphaSample is the following theorem, from which Lemma 7 would follow immediately (see Cohen et al., 2016).

Theorem 12. *Algorithm 2 returns one sample of the loss of each action in U and terminates after at most $4\alpha \log(K/\delta)$ rounds with probability at least $1 - \delta$, provided that all graphs $G_1, \ldots, G_T$ have independence numbers $\leq \alpha$.*

To analyze the number of rounds that the algorithm spends, we shall define the following random process. Consider an infinite sequence $U_1, U_2, \ldots$ such that $U_1 = U$. For every $r > 0$, if U_r is not empty we sample an action uniformly at random from U_r, and we let U_{r+1} be U_r after removing the actions whose losses were observed. Otherwise, we let U_{r+1} be the empty set.

The following lemma lower bounds the expected number of actions whose losses are observed at each iteration of the process.

Lemma 13. *Let $r > 0$. Let N be the number of actions seen when sampling uniformly at random from U_r. Then, $\mathbb{E}[N|U_r] \geq |U_r|/(2\alpha)$.*

The main tool used in the proof of the lemma is the following version of Turán's theorem (see, e.g., Alon and Spencer, 2008).

Theorem 14 (Turán)**.** *Let $G = (V, E)$ be an undirected graph and α be the independence number of G. Then,*

$$\alpha \geq \frac{|V|}{1 + 2|E|/|V|}.$$

Proof of Lemma 13. Fix some feedback graph $G = (V, E)$ with independence number $\leq \alpha$, and let $d_{\text{out}}(v)$ be the out-degree of vertex v. Note that the independence number of the subgraph over U can only decrease, namely it is also at most α. As such, we shall think of $d_{\text{out}}(v)$ as the out-degree of v in the subgraph.

Algorithm 2 AlphaSample

```
input Set of actions $U \subseteq V$
initialize $S \leftarrow \emptyset$
while $|U| > 0$ do
  Play an action $u \in U$ uniformly at random,
    and let $W(u)$ be the set of actions observed
  Collect samples of losses of each $w \in W(u)$ into $S$
  Update $U \leftarrow U \setminus W(u)$
end while
return $S$
```

We would like to apply Turán's theorem to the subgraph, which is a directed graph. We do so by constructing an undirected version of the subgraph, namely one in which we ignore the orientation of the edges. Note that the number of edges in the undirected version can only decrease. Therefore,

$$\mathbb{E}[N|U_r] = 1 + \frac{1}{|U_r|} \sum_{v \in U_r} d_{\text{out}}(v) = 1 + \frac{|E|}{|U_r|} \geq \frac{|U_r|}{2\alpha} ,$$

where the inequality follows from Turán's theorem (Theorem 14). □

Proof of Theorem 12. By the construction of the random process, the probability that Algorithm 2 spends more than t rounds of the game is exactly the probability that U_{t+1} is not empty. To bound this probability we claim that for any $r > 0$,

$$\mathbb{E}[|U_{r+1}|] \leq K \exp\left(-\frac{r}{2\alpha}\right) . \tag{7}$$

Indeed, fix some $i > 0$. By Lemma 13 we have that

$$\mathbb{E}[|U_{i+1}||U_i] = |U_i| - \mathbb{E}[N|U_i] \leq |U_i| \left(1 - \frac{1}{2\alpha}\right) .$$

Taking expectation with respect to U_i and then applying this argument recursively, we get that

$$\mathbb{E}[|U_{r+1}|] \leq |U_1| \left(1 - \frac{1}{2\alpha}\right)^r \leq K \exp\left(-\frac{r}{2\alpha}\right) .$$

Now, let $t_1 = \lfloor 2\alpha \log(K/\delta) \rfloor + 1$. We will show that the probability that U_{t_1+1} is not empty is at most δ. By Markov's inequality and Eq. (7),

$$\begin{aligned} \mathbb{P}[|U_{t_1+1}| > 0] &\leq \mathbb{E}[|U_{t_1+1}|] \\ &\leq K \exp(-t_1/(2\alpha)) \\ &< \delta . \end{aligned}$$

To conclude, with probability at least $1 - \delta$, the number of rounds that the algorithm spends is at most $t_1 \leq 4\alpha \log(K/\delta)$, since $K \geq 2$ by assumption. □

Acknowledgements

TK would like to thank Nicolò Cesa-Bianchi and Ofer Dekel for stimulating discussions in the early stages of this research.

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 multi-armed 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. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40, pages 23–35, 2015.

J. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, pages 217–226, 2009.

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

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA*, pages 142–151, 2012.

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.

A. Cohen, T. Hazan, and T. Koren. Online learning with feedback graphs without the graphs. *arXiv preprint arXiv:1605.07018*, 2016.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.

Y. Freund and R. E. 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.

Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1238–1246, 2013.

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.

T. Kocák, G. Neu, and M. Valko. Online learning with noisy side observations. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics (AISTATS), to appear*, 2016.

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

S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems 24*, pages 684–692, 2011.

V. G. Vovk. Aggregating strategies. In *Proc. Third Workshop on Computational Learning Theory*, pages 371–383. Morgan Kaufmann, 1990.

Y. Wu, A. György, and C. Szepesvári. Online learning with gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368, 2015.