# From Bandits to Experts: A Tale of Domination and Independence

**Noga Alon**
Tel-Aviv University, Israel
nogaa@tau.ac.il

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

**Claudio Gentile**
University of Insubria, Italy
claudio.gentile@uninsubria.it

**Yishay Mansour**
Tel-Aviv University, Israel
mansour@tau.ac.il

## Abstract

We consider the partial observability model for multi-armed bandits, introduced by Mannor and Shamir [14]. Our main result is a characterization of regret in the directed observability model in terms of the dominating and independence numbers of the observability graph (which must be accessible before selecting an action). In the undirected case, we show that the learner can achieve optimal regret without even accessing the observability graph before selecting an action. Both results are shown using variants of the Exp3 algorithm operating on the observability graph in a time-efficient manner.

## 1 Introduction

Prediction with expert advice —see, e.g., [13, 16, 6, 10, 7]— is a general abstract framework for studying sequential prediction problems, formulated as repeated games between a player and an adversary. A well studied example of prediction game is the following: In each round, the adversary privately assigns a loss value to each action in a fixed set. Then the player chooses an action (possibly using randomization) and incurs the corresponding loss. The goal of the player is to control regret, which is defined as the excess loss incurred by the player as compared to the best fixed action over a sequence of rounds. Two important variants of this game have been studied in the past: the expert setting, where at the end of each round the player observes the loss assigned to each action for that round, and the bandit setting, where the player only observes the loss of the chosen action, but not that of other actions.

Let $K$ be the number of available actions, and $T$ be the number of prediction rounds. The best possible regret for the expert setting is of order $\sqrt{T \log K}$. This optimal rate is achieved by the Hedge algorithm [10] or the Follow the Perturbed Leader algorithm [12]. In the bandit setting, the optimal regret is of order $\sqrt{TK}$, achieved by the INF algorithm [2]. A bandit variant of Hedge, called Exp3 [3], achieves a regret with a slightly worse bound of order $\sqrt{TK \log K}$.

Recently, Mannor and Shamir [14] introduced an elegant way for defining intermediate observability models between the expert setting (full observability) and the bandit setting (single observability). An intuitive way of representing an observability model is through a directed graph over actions: an arc[1] from action $i$ to action $j$ implies that when playing action $i$ we get information also about the loss of action $j$. Thus, the expert setting is obtained by choosing a complete graph over actions (playing any action reveals all losses), and the bandit setting is obtained by choosing an empty edge set (playing an action only reveals the loss of that action).

[1] According to the standard terminology in directed graph theory, throughout this paper a directed edge will be called an *arc*.

The main result of [14] concerns undirected observability graphs. The regret is characterized in terms of the independence number $\alpha$ of the undirected observability graph. Specifically, they prove that $\sqrt{T\alpha \log K}$ is the optimal regret (up to logarithmic factors) and show that a variant of Exp3, called ELP, achieves this bound when the graph is known ahead of time, where $\alpha \in \{1, \ldots, K\}$ interpolates between full observability ($\alpha = 1$ for the clique) and single observability ($\alpha = K$ for the graph with no edges). Given the observability graph, ELP runs a linear program to compute the desired distribution over actions. In the case when the graph changes over time, and at each time step ELP observes the current observability graph before prediction, a bound of $\sqrt{\sum_{t=1}^{T} \alpha_t \log K}$ is shown, where $\alpha_t$ is the independence number of the graph at time $t$. A major problem left open in [14] was the characterization of regret for directed observability graphs, a setting for which they only proved partial results.

Our main result is a full characterization (to within logarithmic factors) of regret in the case of directed observability graphs. Our upper bounds are proven using a new algorithm, called Exp3-DOM. This algorithm is efficient to run even when the graph changes over time: it just needs to compute a small dominating set of the current observability graph (which must be given as side information) before prediction.[2] As in the undirected case, the regret for the directed case is characterized in terms of the independence numbers of the observability graphs (computed ignoring edge directions). We arrive at this result by showing that a key quantity emerging in the analysis of Exp3-DOM can be bounded in terms of the independence numbers of the graphs. This bound (Lemma 13 in the appendix) is based on a combinatorial construction which might be of independent interest.

We also explore the possibility of the learning algorithm receiving the observability graph only after prediction, and not before. For this setting, we introduce a new variant of Exp3, called Exp3-SET, which achieves the same regret as ELP for undirected graphs, but without the need of accessing the current observability graph before each prediction. We show that in some random directed graph models Exp3-SET has also a good performance. In general, we can upper bound the regret of Exp3-SET as a function of the maximum acyclic subgraph of the observability graph, but this upper bound may not be tight. Yet, Exp3-SET is much simpler and computationally less demanding than ELP, which needs to solve a linear program in each round.

There are a variety of real-world settings where partial observability models corresponding to directed and undirected graphs are applicable. One of them is route selection. We are given a graph of possible routes connecting cities: when we select a route $r$ connecting two cities, we observe the cost (say, driving time or fuel consumption) of the "edges" along that route and, in addition, we have complete information on any sub-route $r'$ of $r$, but not vice versa. We abstract this in our model by having an observability graph over routes $r$, and an arc from $r$ to any of its sub-routes $r'$.[3]

Sequential prediction problems with partial observability models also arise in the context of recommendation systems. For example, an online retailer, which advertises products to users, knows that users buying certain products are often interested in a set of related products. This knowledge can be represented as a graph over the set of products, where two products are joined by an edge if and only if users who buy any one of the two are likely to buy the other as well. In certain cases, however, edges have a preferred orientation. For instance, a person buying a video game console might also buy a high-def cable to connect it to the TV set. Vice versa, interest in high-def cables need not indicate an interest in game consoles.

Such observability models may also arise in the case when a recommendation system operates in a network of users. For example, consider the problem of recommending a sequence of products, or contents, to users in a group. Suppose the recommendation system is hosted on an online social network, on which users can befriend each other. In this case, it has been observed that social relationships reveal similarities in tastes and interests [15]. However, social links can also be asymmetric (e.g., followers of celebrities). In such cases, followers might be more likely to shape their preferences after the person they follow, than the other way around. Hence, a product liked by a celebrity is probably also liked by his/her followers, whereas a preference expressed by a follower is more often specific to that person.

[2] Computing an approximately minimum dominating set can be done by running a standard greedy set cover algorithm, see Section 2.

[3] Though this example may also be viewed as an instance of combinatorial bandits [8], the model studied here is more general. For example, it does not assume linear losses, which could arise in the routing example from the partial ordering of sub-routes.

## 2 Learning protocol, notation, and preliminaries

As stated in the introduction, we consider an adversarial multi-armed bandit setting with a finite action set $V = \{1, \ldots, K\}$. At each time $t = 1, 2, \ldots$, a player (the "learning algorithm") picks some action $I_t \in V$ and incurs a bounded loss $\ell_{I_t,t} \in [0, 1]$. Unlike the standard adversarial bandit problem [3, 7], where only the played action $I_t$ reveals its loss $\ell_{I_t,t}$, here we assume all the losses in a subset $S_{I_t,t} \subseteq V$ of actions are revealed after $I_t$ is played. More formally, the player observes the pairs $(i, \ell_{i,t})$ for each $i \in S_{I_t,t}$. We also assume $i \in S_{i,t}$ for any $i$ and $t$, that is, any action reveals its own loss when played. Note that the bandit setting ($S_{i,t} = \{i\}$) and the expert setting ($S_{i,t} = V$) are both special cases of this framework. We call $S_{i,t}$ the *observation set* of action $i$ at time $t$, and write $i \xrightarrow{t} j$ when at time $t$ playing action $i$ also reveals the loss of action $j$. Hence, $S_{i,t} = \{j \in V \;:\; i \xrightarrow{t} j\}$. The family of observation sets $\{S_{i,t}\}_{i \in V}$ we collectively call the *observation system* at time $t$.

The adversaries we consider are nonoblivious. Namely, each loss $\ell_{i,t}$ at time $t$ can be an arbitrary function of the past player's actions $I_1, \ldots, I_{t-1}$. The performance of a player $A$ is measured through the regret

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big]$$

where $L_{A,T} = \ell_{I_1,1} + \cdots + \ell_{I_T,T}$ and $L_{k,T} = \ell_{k,1} + \cdots + \ell_{k,T}$ are the cumulative losses of the player and of action $k$, respectively. The expectation is taken with respect to the player's internal randomization (since losses are allowed to depend on the player's past random actions, also $L_{k,t}$ may be random).[4] The observation system $\{S_{i,t}\}_{i \in V}$ is also adversarially generated, and each $S_{i,t}$ can be an arbitrary function of past player's actions, just like losses are. However, in Section 3 we also consider a variant in which the observation system is randomly generated according to a specific stochastic model.

Whereas some algorithms need to know the observation system at the beginning of each step $t$, others need not. From this viewpoint, we consider two online learning settings. In the first setting, called the *informed* setting, the full observation system $\{S_{i,t}\}_{i \in V}$ selected by the adversary is made available to the learner *before* making its choice $I_t$. This is essentially the "side-information" framework first considered in [14]. In the second setting, called the *uninformed setting*, no information whatsoever regarding the time-$t$ observation system is given to the learner prior to prediction. We find it convenient to adopt the same graph-theoretic interpretation of observation systems as in [14]. At each step $t = 1, 2, \ldots$, the observation system $\{S_{i,t}\}_{i \in V}$ defines a directed graph $G_t = (V, D_t)$, where $V$ is the set of actions, and $D_t$ is the set of arcs, i.e., ordered pairs of nodes. For $j \neq i$, arc $(i, j) \in D_t$ if and only if $i \xrightarrow{t} j$ (the self-loops created by $i \xrightarrow{t} i$ are intentionally ignored). Hence, we can equivalently define $\{S_{i,t}\}_{i \in V}$ in terms of $G_t$. Observe that the outdegree $d_i^+$ of any $i \in V$ equals $|S_{i,t}| - 1$. Similarly, the indegree $d_i^-$ of $i$ is the number of action $j \neq i$ such that $i \in S_{j,t}$ (i.e., such that $j \xrightarrow{t} i$). A notable special case of the above is when the observation system is symmetric over time: $j \in S_{i,t}$ if and only if $i \in S_{j,t}$ for all $i, j$ and $t$. In words, playing $i$ at time $t$ reveals the loss of $j$ if and only if playing $j$ at time $t$ reveals the loss of $i$. A symmetric observation system is equivalent to $G_t$ being an undirected graph or, more precisely, to a directed graph having, for every pair of nodes $i, j \in V$, either no arcs or length-two directed cycles. Thus, from the point of view of the symmetry of the observation system, we also distinguish between the *directed* case ($G_t$ is a general directed graph) and the *symmetric* case ($G_t$ is an undirected graph for all $t$).

The analysis of our algorithms depends on certain properties of the sequence of graphs $G_t$. Two graph-theoretic notions playing an important role here are those of *independent sets* and *dominating sets*. Given an undirected graph $G = (V, E)$, an independent set of $G$ is any subset $T \subseteq V$ such that no two $i, j \in T$ are connected by an edge in $E$. An independent set is *maximal* if no proper superset thereof is itself an independent set. The size of a largest (maximal) independent set is the *independence number* of $G$, denoted by $\alpha(G)$. If $G$ is directed, we can still associate with it an independence number: we simply view $G$ as undirected by ignoring arc orientation. If $G = (V, D)$ is a directed graph, then a subset $R \subseteq V$ is a dominating set for $G$ if for all $j \notin R$ there exists some $i \in R$ such that arc $(i, j) \in D$. In our bandit setting, a time-$t$ dominating set $R_t$ is a subset of actions with the property that the loss of any remaining action in round $t$ can be observed by playing

[4] Although we defined the problem in terms of losses, our analysis can be applied to the case when actions return rewards $g_{i,t} \in [0, 1]$ via the transformation $\ell_{i,t} = 1 - g_{i,t}$.

**Algorithm 1:** Exp3-SET algorithm (for the uninformed setting)

**Parameter:** $\eta \in [0,1]$
**Initialize:** $w_{i,1} = 1$ for all $i \in V = \{1,\dots,K\}$
**For** $t = 1,2,\dots$:

1. Observation system $\{S_{i,t}\}_{i\in V}$ is generated but not disclosed ;
2. Set $p_{i,t} = \dfrac{w_{i,t}}{W_{i,t}}$ for each $i \in V$, where $W_t = \sum_{j\in V} w_{j,t}$ ;
3. Play action $I_t$ drawn according to distribution $p_t = (p_{1,t},\dots,p_{K,t})$ ;
4. Observe pairs $(i,\ell_{i,t})$ for all $i \in S_{I_t,t}$;
5. Observation system $\{S_{i,t}\}_{i\in V}$ is disclosed ;
6. For any $i \in V$ set $w_{i,t+1} = w_{i,t}\, \exp\bigl(-\eta\,\widehat{\ell}_{i,t}\bigr)$, where
$$\widehat{\ell}_{i,t} = \frac{\ell_{i,t}}{q_{i,t}}\,\mathbb{I}\{i \in S_{I_t,t}\} \qquad \text{and} \qquad q_{i,t} = \sum_{j\,:\,j \xrightarrow{t} i} p_{j,t}\ .$$

some action in $R_t$. A dominating set is *minimal* if no proper subset thereof is itself a dominating set. The domination number of directed graph $G$, denoted by $\gamma(G)$, is the size of a smallest (minimal) dominating set of $G$.

Computing a minimum dominating set for an arbitrary directed graph $G_t$ is equivalent to solving a minimum set cover problem on the associated observation system $\{S_{i,t}\}_{i\in V}$. Although minimum set cover is NP-hard, the well-known Greedy Set Cover algorithm [9], which repeatedly selects from $\{S_{i,t}\}_{i\in V}$ the set containing the largest number of uncovered elements so far, computes a dominating set $R_t$ such that $|R_t| \le \gamma(G_t)\,(1+\ln K)$.

Finally, we can also lift the notion of independence number of an undirected graph to directed graphs through the notion of *maximum acyclic subgraphs*: Given a directed graph $G = (V,D)$, an acyclic subgraph of $G$ is any graph $G' = (V',D')$ such that $V' \subseteq V$, and $D' = D \cap \bigl(V' \times V'\bigr)$, with no (directed) cycles. We denote by $\mathtt{mas}(G) = |V'|$ the maximum size of such $V'$. Note that when $G$ is undirected (more precisely, as above, when $G$ is a directed graph having for every pair of nodes $i,j \in V$ either no arcs or length-two cycles), then $\mathtt{mas}(G) = \alpha(G)$, otherwise $\mathtt{mas}(G) \ge \alpha(G)$. In particular, when $G$ is itself a directed acyclic graph, then $\mathtt{mas}(G) = |V|$.

## 3 Algorithms without Explicit Exploration: The Uninformed Setting

In this section, we show that a simple variant of the Exp3 algorithm [3] obtains optimal regret (to within logarithmic factors) in the symmetric and uninformed setting. We then show that even the harder adversarial directed setting lends itself to an analysis, though with a weaker regret bound.

Exp3-SET (Algorithm 1) runs Exp3 without mixing with the uniform distribution. Similar to Exp3, Exp3-SET uses loss estimates $\widehat{\ell}_{i,t}$ that divide each observed loss $\ell_{i,t}$ by the probability $q_{i,t}$ of observing it. This probability $q_{i,t}$ is simply the sum of all $p_{j,t}$ such that $j \xrightarrow{t} i$ (the sum includes $p_{i,t}$). Next, we bound the regret of Exp3-SET in terms of the key quantity

$$Q_t = \sum_{i\in V} \frac{p_{i,t}}{q_{i,t}} = \sum_{i\in V} \frac{p_{i,t}}{\sum_{j\,:\,j \xrightarrow{t} i} p_{j,t}}\ . \tag{1}$$

Each term $p_{i,t}/q_{i,t}$ can be viewed as the probability of drawing $i$ from $p_t$ conditioned on the event that $i$ was observed. Similar to [14], a key aspect to our analysis is the ability to deterministically and nonvacuously[5] upper bound $Q_t$ in terms of certain quantities defined on $\{S_{i,t}\}_{i\in V}$. We do so in two ways, either irrespective of how small each $p_{i,t}$ may be (this section) or depending on suitable lower bounds on the probabilities $p_{i,t}$ (Section 4). In fact, forcing lower bounds on $p_{i,t}$ is equivalent to adding exploration terms to the algorithm, which can be done only when knowing $\{S_{i,t}\}_{i\in V}$ before each prediction —an information available only in the informed setting.

[5] An obvious upper bound on $Q_t$ is $K$.

The following result is the building block for all subsequent results in the uninformed setting.[6]

**Theorem 1** *The regret of Exp3-SET satisfies*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[Q_t]\ .$$

As we said, in the adversarial and symmetric case the observation system at time $t$ can be described by an undirected graph $G_t=(V,E_t)$. This is essentially the problem of [14], which they studied in the easier informed setting, where the same quantity $Q_t$ above arises in the analysis of their ELP algorithm. In their Lemma 3, they show that $Q_t \le \alpha(G_t)$, irrespective of the choice of the probabilities $p_t$. When applied to Exp3-SET, this immediately gives the following result.

**Corollary 2** *In the symmetric setting, the regret of Exp3-SET satisfies*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\alpha(G_t)]\ .$$

*In particular, if for constants $\alpha_1,\dots,\alpha_T$ we have $\alpha(G_t)\le\alpha_t$, $t=1,\dots,T$, then setting $\eta=\sqrt{(2\ln K)\big/\sum_{t=1}^{T}\alpha_t}$, gives*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \sqrt{2(\ln K)\sum_{t=1}^{T}\alpha_t}\ .$$

The bounds proven in Corollary 2 are equivalent to those proven in [14] (Theorem 2 therein) for the ELP algorithm. Yet, our analysis is much simpler and, more importantly, our algorithm is simpler and more efficient than ELP, which requires solving a linear program at each step. Moreover, unlike ELP, Exp-SET does not require prior knowledge of the observation system $\{S_{i,t}\}_{i\in V}$ at the beginning of each step.

We now turn to the directed setting. We start by considering a setting in which the observation system is stochastically generated. Then, we turn to the harder adversarial setting.

The Erdős-Renyi model is a standard model for random directed graphs $G=(V,D)$, where we are given a density parameter $r\in[0,1]$ and, for any pair $i,j\in V$, arc $(i,j)\in D$ with independent probability $r$.[7] We have the following result.

**Corollary 3** *Let $G_t$ be generated according to the Erdős-Renyi model with parameter $r\in[0,1]$. Then the regret of Exp3-SET satisfies*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \frac{\ln K}{\eta}+\frac{\eta T}{2r}\left(1-(1-r)^K\right)\ .$$

*In the above, the expectations $\mathbb{E}[\cdot]$ are w.r.t. both the algorithm's randomization and the random generation of $G_t$ occurring at each round. In particular, setting $\eta=\sqrt{\frac{2r\ln K}{T(1-(1-r)^K)}}$, gives*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \sqrt{\frac{2(\ln K)T\left(1-(1-r)^K\right)}{r}}\ .$$

Note that as $r$ ranges in $[0,1]$ we interpolate between the bandit ($r=0$)[8] and the expert ($r=1$) regret bounds.

When the observation system is generated by an adversary, we have the following result.

**Corollary 4** *In the directed setting, the regret of Exp3-SET satisfies*

$$\max_{k\in V}\mathbb{E}\big[L_{A,T}-L_{k,T}\big] \le \frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\mathtt{mas}(G_t)]\ .$$

[6] All proofs are given in the supplementary material to this paper.
[7] Self loops, i.e., arcs $(i,i)$ are included by default here.
[8] Observe that $\lim_{r\to 0^+}\frac{1-(1-r)^K}{r}=K$.

*In particular, if for constants* $m_1, \ldots, m_T$ *we have* $\texttt{mas}(G_t) \le m_t$, $t = 1, \ldots, T$, *then setting* $\eta = \sqrt{(2 \ln K) / \sum_{t=1}^T m_t}$, *gives*

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \sqrt{2(\ln K) \sum_{t=1}^T m_t}\ .$$

Observe that Corollary 4 is a strict generalization of Corollary 2 because, as we pointed out in Section 2, $\texttt{mas}(G_t) \ge \alpha(G_t)$, with equality holding when $G_t$ is an undirected graph.

As far as lower bounds are concerned, in the symmetric setting, the authors of [14] derive a lower bound of $\Omega\big(\sqrt{\alpha(G)T}\big)$ in the case when $G_t = G$ for all $t$. We remark that similar to the symmetric setting, we can derive a lower bound of $\Omega\big(\sqrt{\alpha(G)T}\big)$. The simple observation is that given a directed graph $G$, we can define a new graph $G'$ which is made undirected just by reciprocating arcs; namely, if there is an arc $(i, j)$ in $G$ we add arcs $(i, j)$ and $(j, i)$ in $G'$. Note that $\alpha(G) = \alpha(G')$. Since in $G'$ the learner can only receive more information than in $G$, any lower bound on $G$ also applies to $G'$. Therefore we derive the following corollary to the lower bound of [14] (Theorem 4 therein).

**Corollary 5** *Fix a directed graph* $G$*, and suppose* $G_t = G$ *for all* $t$*. Then there exists a (randomized) adversarial strategy such that for any* $T = \Omega\big(\alpha(G)^3\big)$ *and for any learning strategy, the expected regret of the learner is* $\Omega\big(\sqrt{\alpha(G)T}\big)$.

Moreover, standard results in the theory of Erdős-Renyi graphs, at least in the symmetric case (e.g., [11]), show that, when the density parameter $r$ is constant, the independence number of the resulting graph has an inverse dependence on $r$. This fact, combined with the abovementioned lower bound of [14] gives a lower bound of the form $\sqrt{\frac{T}{r}}$, matching (up to logarithmic factors) the upper bound of Corollary 3.

One may wonder whether a sharper lower bound argument exists which applies to the general directed adversarial setting and involves the larger quantity $\texttt{mas}(G)$. Unfortunately, the above measure does not seem to be related to the optimal regret: Using Claim 1 in the appendix (see proof of Theorem 3) one can exhibit a sequence of graphs each having a large acyclic subgraph, on which the regret of Exp3-SET is still small.

The lack of a lower bound matching the upper bound provided by Corollary 4 is a good indication that something more sophisticated has to be done in order to upper bound $Q_t$ in (1). This leads us to consider more refined ways of allocating probabilities $p_{i,t}$ to nodes. In the next section, we show an allocation strategy that delivers optimal (to within logarithmic factors) regret bounds using prior knowledge of the graphs $G_t$.

## 4 Algorithms with Explicit Exploration: The Informed Setting

We are still in the general scenario where graphs $G_t$ are adversarially generated and directed, but now $G_t$ is made available before prediction. We start by showing a simple example where our analysis of Exp3-SET inherently fails. This is due to the fact that, when the graph induced by the observation system is directed, the key quantity $Q_t$ defined in (1) cannot be nonvacuously upper bounded independent of the choice of probabilities $p_{i,t}$. A way around it is to introduce a new algorithm, called Exp3-DOM, which controls probabilities $p_{i,t}$ by adding an exploration term to the distribution $p_t$. This exploration term is supported on a dominating set of the current graph $G_t$. For this reason, Exp3-DOM requires prior access to a dominating set $R_t$ at each time step $t$ which, in turn, requires prior knowledge of the entire observation system $\{S_{i,t}\}_{i \in V}$.

As announced, the next result shows that, even for simple directed graphs, there exist distributions $p_t$ on the vertices such that $Q_t$ is linear in the number of nodes while the independence number is 1.[9] Hence, nontrivial bounds on $Q_t$ can be found only by imposing conditions on distribution $p_t$.

[9] In this specific example, the maximum acyclic subgraph has size $K$, which confirms the looseness of Corollary 4.

**Algorithm 2:** Exp3-DOM algorithm (for the uninformed setting)

**Input:** Exploration parameters $\gamma^{(b)} \in (0,1]$ for $b \in \{0,1,\dots,\lfloor \log_2 K \rfloor\}$
**Initialization:** $w_{i,1}^{(b)} = 1$ for all $i \in V$ and $b \in \{0,1,\dots,\lfloor \log_2 K \rfloor\}$
**For** $t = 1,2,\dots$ :

1. Observation system $\{S_{i,t}\}_{i\in V}$ is generated *and disclosed* ;
2. Compute a dominating set $R_t \subseteq V$ for $G_t$ associated with $\{S_{i,t}\}_{i\in V}$ ;
3. Let $b_t$ be such that $|R_t| \in \left[2^{b_t}, 2^{b_t+1}-1\right]$;
4. Set $W_t^{(b_t)} = \sum_{i\in V} w_{i,t}^{(b_t)}$;
5. Set $p_{i,t}^{(b_t)} = \left(1-\gamma^{(b_t)}\right)\dfrac{w_{i,t}^{(b_t)}}{W_t^{(b_t)}} + \dfrac{\gamma^{(b_t)}}{|R_t|}\mathbb{I}\{i \in R_t\}$;
6. Play action $I_t$ drawn according to distribution $p_t^{(b_t)} = \left(p_{1,t}^{(b_t)},\dots,p_{V,t}^{(b_t)}\right)$ ;
7. Observe pairs $(i,\ell_{i,t})$ for all $i \in S_{I_t,t}$;
8. For any $i \in V$ set $w_{i,t+1}^{(b_t)} = w_{i,t}^{(b_t)}\,\exp\left(-\gamma^{(b_t)}\,\widehat{\ell}_{i,t}^{(b_t)}/2^{b_t}\right)$, where

$$\widehat{\ell}_{i,t}^{(b_t)} = \frac{\ell_{i,t}}{q_{i,t}^{(b_t)}}\,\mathbb{I}\{i \in S_{I_t,t}\} \qquad \text{and} \qquad q_{i,t}^{(b_t)} = \sum_{j\,:\,j\xrightarrow{t} i} p_{j,t}^{(b_t)} \,.$$

**Fact 6** *Let $G = (V,D)$ be a total order on $V = \{1,\dots,K\}$, i.e., such that for all $i \in V$, arc $(j,i) \in D$ for all $j = i+1,\dots,K$. Let $p = (p_1,\dots,p_K)$ be a distribution on $V$ such that $p_i = 2^{-i}$, for $i < K$ and $p_k = 2^{-K+1}$. Then*

$$Q = \sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j\,:\,j\longrightarrow i} p_j} = \sum_{i=1}^{K} \frac{p_i}{\sum_{j=i}^{K} p_j} = \frac{K+1}{2} \,.$$

We are now ready to introduce and analyze the new algorithm Exp3-DOM for the informed and directed setting. Exp3-DOM (see Algorithm 2) runs $\mathcal{O}(\log K)$ variants of Exp3 indexed by $b = 0,1,\dots,\lfloor \log_2 K \rfloor$. At time $t$ the algorithm is given observation system $\{S_{i,t}\}_{i\in V}$, and computes a dominating set $R_t$ of the directed graph $G_t$ induced by $\{S_{i,t}\}_{i\in V}$. Based on the size $|R_t|$ of $R_t$, the algorithm uses instance $b_t = \lfloor \log_2 |R_t| \rfloor$ to pick action $I_t$. We use a superscript $b$ to denote the quantities relevant to the variant of Exp3 indexed by $b$. Similarly to the analysis of Exp3-SET, the key quantities are

$$q_{i,t}^{(b)} = \sum_{j\,:\,i\in S_{j,t}} p_{j,t}^{(b)} = \sum_{j\,:\,j\xrightarrow{t} i} p_{j,t}^{(b)} \qquad \text{and} \qquad Q_t^{(b)} = \sum_{i\in V} \frac{p_{i,t}^{(b)}}{q_{i,t}^{(b)}}\,, \qquad b = 0,1,\dots,\lfloor \log_2 K \rfloor \,.$$

Let $T^{(b)} = \left\{t = 1,\dots,T \,:\, |R_t| \in [2^b, 2^{b+1}-1]\right\}$. Clearly, the sets $T^{(b)}$ are a partition of the time steps $\{1,\dots,T\}$, so that $\sum_b |T^{(b)}| = T$. Since the adversary adaptively chooses the dominating sets $R_t$, the sets $T^{(b)}$ are random. This causes a problem in tuning the parameters $\gamma^{(b)}$. For this reason, we do not prove a regret bound for Exp3-DOM, where each instance uses a fixed $\gamma^{(b)}$, but for a slight variant (described in the proof of Theorem 7 —see the appendix) where each $\gamma^{(b)}$ is set through a doubling trick.

**Theorem 7** *In the directed case, the regret of Exp3-DOM satisfies*

$$\max_{k\in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \left( \frac{2^b \ln K}{\gamma^{(b)}} + \gamma^{(b)} \mathbb{E}\left[ \sum_{t\in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}}\right)\right]\right) \,. \tag{2}$$

*Moreover, if we use a doubling trick to choose $\gamma^{(b)}$ for each $b = 0, \ldots, \lfloor \log_2 K \rfloor$, then*

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] = \mathcal{O}\left( (\ln K)\, \mathbb{E}\left[ \sqrt{\sum_{t=1}^{T} \Big( 4|R_t| + Q_t^{(b_t)} \Big)} \right] + (\ln K) \ln(KT) \right) . \tag{3}$$

Importantly, the next result shows how bound (3) of Theorem 7 can be expressed in terms of the sequence $\alpha(G_t)$ of independence numbers of graphs $G_t$ whenever the Greedy Set Cover algorithm [9] (see Section 2) is used to compute the dominating set $R_t$ of the observation system at time $t$.

**Corollary 8** *If Step 2 of Exp3-DOM uses the Greedy Set Cover algorithm to compute the dominating sets $R_t$, then the regret of Exp-DOM with doubling trick satisfies*

$$\max_{k \in V} \mathbb{E}\big[L_{A,T} - L_{k,T}\big] = \mathcal{O}\left( \ln(K) \sqrt{\ln(KT) \sum_{t=1}^{T} \alpha(G_t)} + \ln(K)\ln(KT) \right)$$

*where, for each t, $\alpha(G_t)$ is the independence number of the graph $G_t$ induced by observation system $\{S_{i,t}\}_{i \in V}$.*

Comparing Corollary 8 to Corollary 5 delivers the announced characherization in the general adversarial and directed setting. Moreover, a quick comparison between Corollary 2 and Corollary 8 reveals that a symmetric observation system overcomes the advantage of working in an informed setting: The bound we obtained for the uninformed symmetric setting (Corollary 2) is sharper by logarithmic factors than the one we derived for the informed —but more general, i.e., directed— setting (Corollary 8).

## 5 Conclusions and work in progress

We have investigated online prediction problems in partial information regimes that interpolate between the classical bandit and expert settings. We have shown a number of results characterizing prediction performance in terms of: the structure of the observation system, the amount of information available before prediction, the nature (adversarial or fully random) of the process generating the observation system. Our results are substantial improvements over the paper [14] that initiated this interesting line of research. Our improvements are diverse, and range from considering both informed and uninformed settings to delivering more refined graph-theoretic characterizations, from providing more efficient algorithmic solutions to relying on simpler (and often more general) analytical tools.

Some research directions we are currently pursuing are the following: (1) We are currently investigating the extent to which our results could be applied to the case when the observation system $\{S_{i,t}\}_{i \in V}$ may depend on the loss $\ell_{I_t,t}$ of player's action $I_t$. Note that this would prevent a direct construction of an unbiased estimator for unobserved losses, which many worst-case bandit algorithms (including ours —see the appendix) hinge upon. (2) The upper bound contained in Corollary 4 and expressed in terms of $\mathtt{mas}(\cdot)$ is almost certainly suboptimal, even in the uninformed setting, and we are trying to see if more adequate graph complexity measures can be used instead. (3) Our lower bound in Corollary 5 heavily relies on the corresponding lower bound in [14] which, in turn, refers to a constant graph sequence. We would like to provide a more complete charecterization applying to sequences of adversarially-generated graphs $G_1, G_2, \ldots, G_T$ in terms of sequences of their corresponding independence numbers $\alpha(G_1), \alpha(G_2), \ldots, \alpha(G_T)$ (or variants thereof), in both the uninformed and the informed settings. (4) All our upper bounds rely on parameters to be tuned as a function of sequences of observation system quantities (e.g., the sequence of independence numbers). We are trying to see if an adaptive learning rate strategy à la [4], based on the observable quantities $Q_t$, could give similar results without such a prior knowledge.

**Acknowledgments**

NA was supported in part by an ERC advanced grant, by a USA-Israeli BSF grant, and by the Israeli I-CORE program. NCB acknowledges partial support by MIUR (project ARS TechnoMedia, PRIN 2010-2011, grant no. 2010N5K7EB_003). YM was supported in part by a grant from the Israel Science Foundation, a grant from the United States-Israel Binational Science Foundation (BSF), a grant by Israel Ministry of Science and Technology and the Israeli Centers of Research Excellence (I-CORE) program (Center No. 4/11).

## References


[1] N. Alon and J. H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[2] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, 2009.

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comput. Syst. Sci.*, 64(1):48–75, 2002.

[5] Y. Caro. New results on the independence number. In *Tech. Report, Tel-Aviv University*, 1979.

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

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

[8] Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *J. Comput. Syst. Sci.*, 78(5):1404–1422, 2012.

[9] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

[10] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Euro-COLT*, pages 23–37. Springer-Verlag, 1995. Also, JCSS 55(1): 119-139 (1997).

[11] A. M. Frieze. On the independence number of random graphs. *Discrete Mathematics*, 81:171–175, 1990.

[12] A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

[13] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

[14] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *25th Annual Conference on Neural Information Processing Systems (NIPS 2011)*, 2011.

[15] Alan Said, Ernesto W De Luca, and Sahin Albayrak. How social relationships affect user similarities. In *Proceedings of the International Conference on Intelligent User Interfaces Workshop on Social Recommender Systems, Hong Kong*, 2010.

[16] V. G. Vovk. Aggregating strategies. In *COLT*, pages 371–386, 1990.

[17] V. K. Wey. A lower bound on the stability number of a simple graph. In *Bell Lab. Tech. Memo No. 81-11217-9*, 1981.

# A Technical lemmas and proofs

This section contains the proofs of all technical results occurring in the main text, along with ancillary graph-theoretic lemmas. Throughout this appendix, $\mathbb{E}_t[\cdot]$ is a shorthand for $\mathbb{E}\big[\cdot \mid I_1,\dots,I_{t-1}\big]$.

**Proof of Theorem 1**
Following the proof of Exp3 [3], we have

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i\in V}\frac{w_{i,t+1}}{W_t}\\
&= \sum_{i\in V}\frac{w_{i,t}\,\exp(-\eta\,\widehat{\ell}_{i,t})}{W_t}\\
&= \sum_{i\in V} p_{i,t}\,\exp(-\eta\,\widehat{\ell}_{i,t})\\
&\le \sum_{i\in V} p_{i,t}\left(1-\eta\widehat{\ell}_{i,t}+\frac{1}{2}\,\eta^2(\widehat{\ell}_{i,t})^2\right) \qquad \text{using } e^{-x}\le 1-x+x^2/2 \text{ for all } x\ge 0\\
&\le 1-\eta\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t}+\frac{\eta^2}{2}\sum_{i\in V} p_{i,t}(\widehat{\ell}_{i,t})^2\;.
\end{aligned}$$

Taking logs, using $\ln(1-x)\le -x$ for all $x\ge 0$, and summing over $t=1,\dots,T$ yields

$$\ln\frac{W_{T+1}}{W_1}\le -\eta\sum_{t=1}^T\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t}+\frac{\eta^2}{2}\sum_{t=1}^T\sum_{i\in V} p_{i,t}(\widehat{\ell}_{i,t})^2\;.$$

Moreover, for any fixed comparison action $k$, we also have

$$\ln\frac{W_{T+1}}{W_1}\ge \ln\frac{w_{k,T+1}}{W_1} = -\eta\sum_{t=1}^T\widehat{\ell}_{k,t}-\ln K\;.$$

Putting together and rearranging gives

$$\sum_{t=1}^T\sum_{i\in V} p_{i,t}\widehat{\ell}_{i,t}\le\sum_{t=1}^T\widehat{\ell}_{k,t}+\frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^T\sum_{i\in V} p_{i,t}(\widehat{\ell}_{i,t})^2\;. \tag{4}$$

Note that, for all $i\in V$,

$$\mathbb{E}_t[\widehat{\ell}_{i,t}] = \sum_{j\,:\,i\in S_{j,t}} p_{j,t}\,\frac{\ell_{i,t}}{q_{i,t}} = \sum_{j\,:\,j\xrightarrow{t}i} p_{j,t}\,\frac{\ell_{i,t}}{q_{i,t}} = \frac{\ell_{i,t}}{q_{i,t}}\sum_{j\,:\,j\xrightarrow{t}i} p_{j,t} = \ell_{i,t}\;.$$

Moreover,

$$\mathbb{E}_t\big[(\widehat{\ell}_{i,t})^2\big] = \sum_{j\,:\,i\in S_{j,t}} p_{j,t}\,\frac{\ell_{i,t}^2}{q_{i,t}^2} = \frac{\ell_{i,t}^2}{q_{i,t}^2}\sum_{j\,:\,j\xrightarrow{t}i} p_{j,t} \le \frac{1}{q_{i,t}^2}\sum_{j\,:\,j\xrightarrow{t}i} p_{j,t} = \frac{1}{q_{i,t}}\;.$$

Hence, taking expectations $\mathbb{E}_t$ on both sides of (4), and recalling the definition of $Q_t$, we can write

$$\sum_{t=1}^T\sum_{i\in V} p_{i,t}\,\ell_{i,t}\le\sum_{t=1}^T\ell_{k,t}+\frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^T Q_t\;. \tag{5}$$

Finally, taking expectations to remove conditioning gives

$$\mathbb{E}\big[L_{A,T}-L_{k,T}\big]\le\frac{\ln K}{\eta}+\frac{\eta}{2}\sum_{t=1}^T\mathbb{E}[Q_t]\;,$$

as claimed. $\square$

**Proof of Corollary 3**
Fix round $t$, and let $G=(V,D)$ be the Erdős-Renyi random graph generated at time $t$, $N_i^-$ be the in-neighborhood of node $i$, i.e., the set of nodes $j$ such that $(j,i)\in D$, and denote by $d_i^-$ the indegree of $i$.

**Claim 1** *Let $p_1, \ldots, p_K$ be an arbitrary probability distribution defined over $V$, $f : V \to V$ be an arbitrary permutation of $V$, and $\mathbb{E}_f$ denote the expectation w.r.t. permutation $f$ when $f$ is drawn uniformly at random. Then, for any $i \in V$, we have*

$$\mathbb{E}_f \left[ \frac{p_{f(i)}}{p_{f(i)} + \sum_{j\,:\,f(j) \in N^-_{f(i)}} p_{f(j)}} \right] = \frac{1}{1 + d_i^-} \ .$$

**Proof.** Consider selecting a subset $S \subset V$ of $1 + d_i^-$ nodes. We shall consider the contribution to the expectation when $S = N^-_{f(i)} \cup \{f(i)\}$. Since there are $K(K-1)\cdots(K-d_i^-+1)$ terms (out of $K!$) contributing to the expectation, we can write

$$\begin{aligned}
\mathbb{E}_f \left[ \frac{p_{f(i)}}{p_{f(i)} + \sum_{j\,:\,f(j) \in N^-_{f(i)}} p_{f(j)}} \right] &= \frac{1}{\binom{K}{d_i^-}} \sum_{S \subset V, |S| = d_i^-} \frac{1}{1+d_i^-} \sum_{i \in S} \frac{p_i}{p_i + \sum_{j \in S, j \neq i} p_j} \\
&= \frac{1}{\binom{K}{d_i^-}} \sum_{S \subset V, |S| = d_i^-} \frac{1}{1+d_i^-} \\
&= \frac{1}{1+d_i^-} \ .
\end{aligned}$$

□

**Claim 2** *Let $p_1, \ldots, p_K$ be an arbitrary probability distribution defined over $V$, and $\mathbb{E}$ denote the expectation w.r.t. the Erdős-Renyi random draw of arcs at time $t$. Then, for any fixed $i \in V$, we have*

$$\mathbb{E} \left[ \frac{p_i}{p_i + \sum_{j\,:\,j \xrightarrow{t} i} p_j} \right] = \frac{1}{rK} \left( 1 - (1-r)^K \right) \ .$$

**Proof.** For the given $i \in V$ and time $t$, consider the Bernoulli random variables $X_j$, $j \in V \setminus \{i\}$, and denote by $\mathbb{E}_{j\,:\,j \neq i}$ the expectation w.r.t. all of them. We symmetrize $\mathbb{E}\left[\frac{p_i}{p_i + \sum_{j\,:\,j \xrightarrow{t} i} p_j}\right]$ by means of a random permutation $f$, as in Claim 1. We can write

$$\begin{aligned}
\mathbb{E} \left[ \frac{p_i}{p_i + \sum_{j\,:\,j \xrightarrow{t} i} p_j} \right] &= \mathbb{E}_{j\,:\,j \neq i} \left[ \frac{p_i}{p_i + \sum_{j\,:\,j \neq i} X_j p_j} \right] \\
&= \mathbb{E}_{j\,:\,j \neq i} \mathbb{E}_f \left[ \frac{p_{f(i)}}{p_{f(i)} + \sum_{j\,:\,j \neq i} X_{f(j)} p_{f(j)}} \right] \qquad \text{(by symmetry)} \\
&= \mathbb{E}_{j\,:\,j \neq i} \left[ \frac{1}{1 + \sum_{j\,:\,j \neq i} X_j} \right] \qquad \text{(from Claim 1)} \\
&= \sum_{i=0}^{K-1} \binom{K-1}{i} r^i (1-r)^{K-1-i} \frac{1}{i+1} \\
&= \frac{1}{rK} \sum_{i=0}^{K-1} \binom{K}{i+1} r^{i+1} (1-r)^{K-1-i} \\
&= \frac{1}{rK} \left( 1 - (1-r)^K \right) \ .
\end{aligned}$$

□

At this point, we follow the proof of Theorem 1 up until (5). We take an expectation $\mathbb{E}_{G_1,\ldots,G_T}$ w.r.t. the randomness in generating the sequence of graphs $G_1, \ldots, G_T$. This yields

$$\sum_{t=1}^{T} \mathbb{E}_{G_1,\ldots,G_T} \left[ \sum_{i \in V} p_{i,t}\, \ell_{i,t} \right] \leq \sum_{t=1}^{T} \ell_{k,t} + \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathbb{E}_{G_1,\ldots,G_T} \left[ Q_t \right] \ .$$

We use Claim 2 to upper bound $\mathbb{E}_{G_1,\ldots,G_T}[Q_t]$ by $\frac{1}{r}\left(1-(1-r)^K\right)$, and take the outer expectation to remove conditioning, as in the proof of Theorem 1. This concludes the proof. $\square$

The following lemma can be seen as a generalization of Lemma 3 in [14].

**Lemma 9** *Let $G = (V, D)$ be a directed graph with vertex set $V = \{1, \ldots, K\}$, and arc set $D$. Let $N_i^-$ be the in-neighborhood of node $i$, i.e., the set of nodes $j$ such that $(j, i) \in D$. Then*

$$\sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} \le \mathtt{mas}(G) \, .$$

**Proof.** We will show that there is a subset of vertices $V'$ such that the induced graph is acyclic and $|V'| \ge \sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j\in N_i^-} p_j}$.

We prove the lemma by growing set $V'$ starting off from $V' = \emptyset$. Let

$$\Phi_0 = \sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} \, ,$$

and $i_1$ be the vertex which minimizes $p_i + \sum_{j\in N_i^-} p_j$ over $i \in V$. We are going to delete $i_1$ from the graph, along with all its incoming neighbors (set $N_{i_1}^-$), and all edges which are incident (both departing and incoming) to these nodes, and then iterating on the remaining graph. Let us denote the in-neighborhoods of the shrunken graph from the first step by $N_{i,1}^-$.

The contribution of all the deleted vertices to $\Phi_0$ is

$$\sum_{r \in N_{i_1}^- \cup \{i_1\}} \frac{p_r}{p_r + \sum_{j \in N_r^-} p_j} \le \sum_{r \in N_{i_1}^- \cup \{i_1\}} \frac{p_r}{p_{i_1} + \sum_{j \in N_{i_1}^-} p_j} = 1 \, ,$$

where the inequality follows from the minimality of $i_1$.

Let $V' \leftarrow V' \cup \{i_1\}$, and $V_1 = V - (N_{i_1}^- \cup \{i_1\})$. Then from the first step we have

$$\Phi_1 = \sum_{i \in V_1} \frac{p_i}{p_i + \sum_{j \in N_{i,1}^-} p_j} \ge \sum_{i \in V_1} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} \ge \Phi_0 - 1 \, .$$

We apply the very same argument to $\Phi_1$ with node $i_2$ (minimizing $p_i + \sum_{j\in N_{i,1}^-} p_j$ over $i \in V_1$), to $\Phi_2$ with node $i_3$, ..., to $\Phi_{s-1}$ with node $i_s$, up until $\Phi_s = 0$, i.e., up until no nodes are left in the shrunken graph. This gives $\Phi_0 \le s = |V'|$, where $V' = \{i_1, i_2, \ldots, i_s\}$. Moreover, since in each step $r = 1, \ldots, s$ we remore all remaining arcs incoming to $i_r$, the graph induced by set $V'$ cannot contain cycles. $\square$

**Proof of Corollary 4**
The claim follows from a direct combination of Theorem 1 with Lemma 9. $\square$

**Proof of Fact 6**
Using standard properties of geometric sums, one can immediately see that

$$\sum_{i=1}^{K} \frac{p_i}{\sum_{j=i}^{K} p_j} = \sum_{i=1}^{K-1} \frac{2^{-i}}{2^{-i+1}} + \frac{2^{-K+1}}{2^{-K+1}} = \frac{K-1}{2} + 1 = \frac{K+1}{2} \, ,$$

hence the claimed result. $\square$

The following graph-theoretic lemma turns out to be fairly useful for analyzing directed settings. It is a directed-graph counterpart to a well-known result [5, 17] holding for undirected graphs.

**Lemma 10** *Let $G = (V, D)$ be a directed graph, with $V = \{1, \ldots, K\}$. Let $d_i^-$ be the indegree of node $i$, and $\alpha = \alpha(G)$ be the independence number of $G$. Then*

$$\sum_{i=1}^{K} \frac{1}{1 + d_i^-} \le 2\alpha \, \ln\left(1 + \frac{K}{\alpha}\right) \, .$$

**Proof.** We will proceed by induction, starting off from the original $K$-node graph $G = G_K$ with indegrees $\{d_i^-\}_{i=1}^K = \{d_{i,K}^-\}_{i=1}^K$, and independence number $\alpha = \alpha_K$, and then progressively shrink $G$ by eliminating nodes and incident (both departing and incoming) arcs, thereby obtaining a sequence of smaller and smaller graphs $G_K, G_{K-1}, G_{K-2}, \ldots$, and associated indegrees $\{d_{i,K}^-\}_{i=1}^K$, $\{d_{i,K-1}^-\}_{i=1}^{K-1}$, $\{d_{i,K-2}^-\}_{i=1}^{K-2}$, $\ldots$, and independence numbers $\alpha_K, \alpha_{K-1}, \alpha_{K-2}, \ldots$. Specifically, in step $s$ we sort nodes $i = 1, \ldots, s$ of $G_s$ in nonincreasing value of $d_{i,s}^-$, and obtain $G_{s-1}$ from $G_s$ by eliminating node 1 (i.e., one having the largest indegree among the nodes of $G_s$), along with its incident arcs. On all such graphs, we will use the classical Turan's theorem (e.g., [1]) stating that any *undirected* graph with $n_s$ nodes and $m_s$ edges has an independent set of size at least $\frac{n_s}{\frac{2m_s}{n_s}+1}$. This implies that if $G_s = (V_s, D_s)$, then $\alpha_s$ satisfies[10]

$$\frac{|D_s|}{|V_s|} \geq \frac{|V_s|}{2\alpha_s} - \frac{1}{2} \,. \tag{6}$$

We then start from $G_K$. We can write

$$d_{1,K}^- = \max_{i=1\ldots K} d_{i,K}^- \geq \frac{1}{K}\sum_{i=1}^K d_{i,K}^- = \frac{|D_K|}{|V_K|} \geq \frac{|V_K|}{2\alpha_K} - \frac{1}{2} \,.$$

Hence,

$$\begin{aligned}
\sum_{i=1}^K \frac{1}{1+d_{i,K}^-} &= \frac{1}{1+d_{1,K}^-} + \sum_{i=2}^K \frac{1}{1+d_{i,K}^-} \\
&\leq \frac{2\alpha_K}{\alpha_K + K} + \sum_{i=2}^K \frac{1}{1+d_{i,K}^-} \\
&\leq \frac{2\alpha_K}{\alpha_K + K} + \sum_{i=1}^{K-1} \frac{1}{1+d_{i,K-1}^-},
\end{aligned}$$

where the last inequality follows from $d_{i+1,K}^- \geq d_{i,K-1}^-$, $i = 1, \ldots K-1$, due to the arc elimination turning $G_K$ into $G_{K-1}$. Recursively applying the very same argument to $G_{K-1}$ (i.e., to the sum $\sum_{i=1}^{K-1} \frac{1}{1+d_{i,K-1}^-}$), and then iterating all the way to $G_1$ yields the upper bound

$$\sum_{i=1}^K \frac{1}{1+d_{i,K}^-} \leq \sum_{i=1}^K \frac{2\alpha_i}{\alpha_i + i} \,.$$

Combining with $\alpha_i \leq \alpha_K = \alpha$, and $\sum_{i=1}^K \frac{1}{\alpha+i} \leq \ln\left(1+\frac{K}{\alpha}\right)$ concludes the proof. □

The next lemma relates the size $|R_t|$ of the dominating set $R_t$ computed by the Greedy Set Cover algorithm of [9] operating on the time-$t$ observation system $\{S_{i,t}\}_{i\in V}$ to the independence number $\alpha(G_t)$ and the domination number $\gamma(G_t)$ of $G_t$.

**Lemma 11** *Let $\{S_i\}_{i\in V}$ be an observation system, and $G = (V, D)$ be the induced directed graph, with vertex set $V = \{1, \ldots, K\}$, independence number $\alpha = \alpha(G)$, and domination number $\gamma = \gamma(G)$. Then the dominating set $R$ constructed by the Greedy Set Cover algorithm (see Section 2) satisfies*

$$|R| \leq \min\bigl\{\gamma(1+\ln K), \lceil 2\alpha \ln K\rceil + 1\bigr\} \,.$$

**Proof.** As recalled in Section 2, the Greedy Set Cover algorithm of [9] achieves $|R| \leq \gamma(1+\ln K)$. In order to prove the other bound, consider the sequence of graphs $G = G_1, G_2, \ldots$, where each $G_{s+1} = (V_{s+1}, D_{s+1})$ is obtained by removing from $G_s$ the vertex $i_s$ selected by the Greedy Set

[10] Notice that $|D_s|$ is at least as large as the number of edges of the undirected version of $G_s$ which the independence number $\alpha_s$ actually refers to.

Cover algorithm, together with all the vertices in $G_s$ that are dominated by $i_s$, and all arcs incident to these vertices. By definition of the algorithm, the outdegree $d_s^+$ of $i_s$ in $G_s$ is largest in $G_s$. Hence,

$$d_s^+ \geq \frac{|D_s|}{|V_s|} \geq \frac{|V_s|}{2\alpha_s} - \frac{1}{2} \geq \frac{|V_s|}{2\alpha} - \frac{1}{2}$$

by Turan's theorem (e.g., [1]), where $\alpha_s$ is the independence number of $G_s$ and $\alpha \geq \alpha_s$. This shows that

$$|V_{s+1}| = |V_s| - d_s^+ - 1 \leq |V_s|\left(1 - \frac{1}{2\alpha}\right) \leq |V_s| e^{-1/(2\alpha)} .$$

Iterating, we obtain $|V_s| \leq K\, e^{-s/(2\alpha)}$. Choosing $s = \lceil 2\alpha \ln K \rceil + 1$ gives $|V_s| < 1$, thereby covering all nodes. Hence the dominating set $R = \{i_1, \dots, i_s\}$ so constructed satisfies $|R| \leq \lceil 2\alpha \ln K \rceil + 1$. □

**Lemma 12** *If $a, b \geq 0$, and $a + b \geq B > A > 0$, then*

$$\frac{a}{a+b-A} \leq \frac{a}{a+b} + \frac{A}{B-A} .$$

**Proof.**

$$\frac{a}{a+b-A} - \frac{a}{a+b} = \frac{aA}{(a+b)(a+b-A)} \leq \frac{A}{a+b-A} \leq \frac{A}{B-A} .$$

□

We now lift Lemma 10 to a more general statement.

**Lemma 13** *Let $G = (V, D)$ be a directed graph, with vertex set $V = \{1, \dots, K\}$, and arc set $D$. Let $N_i^-$ be the in-neighborhood of node $i$, i.e., the set of nodes $j$ such that $(j, i) \in D$. Let $\alpha$ be the independence number of $G$, $R \subseteq V$ be a dominating set for $G$ of size $r = |R|$, and $p_1, \dots, p_K$ be a probability distribution defined over $V$, such that $p_i \geq \beta > 0$, for $i \in R$. Then*

$$\sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} \leq 2\alpha \ln\left(1 + \frac{\lceil \frac{K^2}{r\beta} \rceil + K}{\alpha}\right) + 2r .$$

**Proof.** The idea is to appropriately discretize the probability values $p_i$, and then upper bound the discretized counterpart of $\sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j}$ by reducing to an expression that can be handled by Lemma 10. In order to make this discretization effective, we need to single out the terms $\frac{p_i}{p_i + \sum_{j \in N_i^-} p_j}$ corresponding to nodes $i \in R$. We first write

$$\begin{aligned}
\sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} &= \sum_{i \in R} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} + \sum_{i \notin R} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} \\
&\leq r + \sum_{i \notin R} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j} , \qquad (7)
\end{aligned}$$

and then focus on (7).

Let us discretize the unit interval[11] $(0, 1]$ into subintervals $(\frac{j-1}{M}, \frac{j}{M}]$, $j = 1, \dots, M$, where $M = \lceil \frac{K^2}{r\beta} \rceil$. Let $\widehat{p}_i = j/M$ be the discretized version of $p_i$, being $j$ the unique integer such that

$$\widehat{p}_i - 1/M < p_i \leq \widehat{p}_i .$$

[11] The zero value won't be of our concern here, because if $p_i = 0$, the corresponding term in (7) can be disregarded.

Let us focus on a single node $i \notin R$ with indegree $d_i^- = |N_i^-|$, and introduce the shorthand notation $P_i = \sum_{j \in N_i^-} p_j$, and $\widehat{P}_i = \sum_{j \in N_i^-} \widehat{p}_j$. We have that $\widehat{P}_i \geq P_i \geq \beta$, since $i$ is dominated by some node $j \in R \cap N_i^-$ such that $p_j \geq \beta$. Moreover, $P_i > \widehat{P}_i - \frac{d_i^-}{M} \geq \beta - \frac{d_i^-}{M} > 0$, and $\widehat{p}_i + \widehat{P}_i \geq \beta$. Hence, for any fixed node $i \notin R$, we can write

$$
\begin{aligned}
\frac{p_i}{p_i + P_i} &\leq \frac{\widehat{p}_i}{\widehat{p}_i + P_i} \\
&< \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i - \frac{d_i^-}{M}} \\
&\leq \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{d_i^-/M}{\beta - d_i^-/M} \\
&= \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{d_i^-}{\beta M - d_i^-} \\
&< \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + \frac{r}{K - r},
\end{aligned}
$$

where in the second-last inequality we used Lemma 12 with $a = \widehat{p}_i$, $b = \widehat{P}_i$, $A = d_i^-/M$, and $B = \beta > d_i^-/M$. Recalling (7), and summing over $i$ then gives

$$
\sum_{i=1}^{K} \frac{p_i}{p_i + P_i} \leq r + \sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + r = \sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} + 2r \; . \tag{8}
$$

Therefore, we continue by bounding from above the right-hand side of (8). We first observe that

$$
\sum_{i \notin R} \frac{\widehat{p}_i}{\widehat{p}_i + \widehat{P}_i} = \sum_{i \notin R} \frac{\widehat{s}_i}{\widehat{s}_i + \widehat{S}_i}, \qquad \widehat{S}_i = \sum_{j \in N_i^-} \widehat{s}_j \; , \tag{9}
$$

where $\widehat{s}_i = M\widehat{p}_i$, $i = 1, \ldots, K$, are integers. Based on the original graph $G$, we construct a new graph $\widehat{G}$ made up of connected cliques. In particular:

- Each node $i$ of $G$ is replaced in $\widehat{G}$ by a clique $C_i$ of size $\widehat{s}_i$; nodes within $C_i$ are connected by length-two cycles.
- If arc $(i, j)$ is in $G$, then for *each* node of $C_i$ draw an arc towards *each* node of $C_j$.

We would like to apply Lemma 10 to $\widehat{G}$. Notice that, by the above construction:

- The independence number of $\widehat{G}$ is the same as that of $G$;
- The indegree $\widehat{d_k^-}$ of each node $k$ in clique $C_i$ satisfies $\widehat{d_k^-} = \widehat{s}_i - 1 + \widehat{S}_i$.
- The total number of nodes of $\widehat{G}$ is
$$
\sum_{i=1}^{K} \widehat{s}_i = M \sum_{i=1}^{K} \widehat{p}_i < M \sum_{i=1}^{K} \left( p_i + \frac{1}{M} \right) = M + K \; .
$$

Hence, we are in a position to apply Lemma 10 to $\widehat{G}$ with indegrees $\widehat{d_k^-}$, revealing that

$$
\sum_{i \notin R} \frac{\widehat{s}_i}{\widehat{s}_i + \widehat{S}_i} = \sum_{i \notin R} \sum_{k \in C_i} \frac{1}{1 + \widehat{d_k^-}} \leq \sum_{i=1}^{K} \sum_{k \in C_i} \frac{1}{1 + \widehat{d_k^-}} \leq 2\alpha \ln\left(1 + \frac{M + K}{\alpha}\right) \; .
$$

Putting together as in (8) and (9), and recalling the value of $M$ gives the claimed result. □

**Proof of Theorem 7**

We start to bound the contribution to the overall regret of an instance indexed by $b$. When clear from the context, we remove the superscript $b$ from $\gamma^{(b)}$, $w_{i,t}^{(b)}$, $p_{i,t}^{(b)}$, and other related quantities. For any $t \in T^{(b)}$ we have

$$
\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i \in V} \frac{w_{i,t+1}}{W_t} \\
&= \sum_{i \in V} \frac{w_{i,t}}{W_t} \exp\bigl(-(\gamma/2^b)\, \widehat{\ell}_{i,t}\bigr) \\
&= \sum_{i \in R_t} \frac{p_{i,t} - \gamma/|R_t|}{1-\gamma} \exp\bigl(-(\gamma/2^b)\, \widehat{\ell}_{i,t}\bigr) + \sum_{i \notin R_t} \frac{p_{i,t}}{1-\gamma} \exp\bigl(-(\gamma/2^b)\, \widehat{\ell}_{i,t}\bigr) \\
&\le \sum_{i \in R_t} \frac{p_{i,t} - \gamma/|R_t|}{1-\gamma} \left(1 - \frac{\gamma}{2^b}\widehat{\ell}_{i,t} + \frac{1}{2}\left(\frac{\gamma}{2^b}\widehat{\ell}_{i,t}\right)^2\right) + \sum_{i \notin R_t} \frac{p_{i,t}}{1-\gamma} \left(1 - \frac{\gamma}{2^b}\widehat{\ell}_{i,t} + \frac{1}{2}\left(\frac{\gamma}{2^b}\widehat{\ell}_{i,t}\right)^2\right) \\
&\quad \text{(using } e^{-x} \le 1 - x + x^2/2 \text{ for all } x \ge 0\text{)} \\
&\le 1 - \frac{\gamma/2^b}{1-\gamma} \sum_{i \in V} p_{i,t}\widehat{\ell}_{i,t} + \frac{\gamma^2/2^b}{1-\gamma} \sum_{i \in R_t} \frac{\widehat{\ell}_{i,t}}{|R_t|} + \frac{1}{2}\frac{(\gamma/2^b)^2}{1-\gamma} \sum_{i \in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2 .
\end{aligned}
$$

Taking logs, upper bounding, and summing over $t \in T^{(b)}$ yields

$$
\ln \frac{W_{|T^{(b)}|+1}}{W_1} \le -\frac{\gamma/2^b}{1-\gamma} \sum_{t \in T^{(b)}} \sum_{i \in V} p_{i,t}\widehat{\ell}_{i,t} + \frac{\gamma^2/2^b}{1-\gamma} \sum_{t \in T^{(b)}} \sum_{i \in R_t} \frac{\widehat{\ell}_{i,t}}{|R_t|} + \frac{1}{2}\frac{(\gamma/2^b)^2}{1-\gamma} \sum_{t \in T^{(b)}} \sum_{i \in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2 .
$$

Moreover, for any fixed comparison action $k$, we also have

$$
\ln \frac{W_{|T^{(b)}|+1}}{W_1} \ge \ln \frac{w_{k,|T^{(b)}|+1}}{W_1} = -\frac{\gamma}{2^b} \sum_{t \in T^{(b)}} \widehat{\ell}_{k,t} - \ln K .
$$

Putting together, rearranging, and using $1 - \gamma \le 1$ gives

$$
\sum_{t \in T^{(b)}} \sum_{i \in V} p_{i,t}\widehat{\ell}_{i,t} \le \sum_{t \in T^{(b)}} \widehat{\ell}_{k,t} + \frac{2^b \ln K}{\gamma} + \gamma \sum_{t \in T^{(b)}} \sum_{i \in R_t} \frac{\widehat{\ell}_{i,t}}{|R_t|} + \frac{\gamma}{2^{b+1}} \sum_{t \in T^{(b)}} \sum_{i \in V} p_{i,t}\bigl(\widehat{\ell}_{i,t}\bigr)^2 .
$$

Reintroducing the notation $\gamma^{(b)}$ and summing over $b = 0, 1, \ldots, \lfloor \log_2 K \rfloor$ gives

$$
\sum_{t=1}^{T} \left( \sum_{i \in V} p_{i,t}^{(b_t)} \widehat{\ell}_{i,t}^{(b_t)} - \widehat{\ell}_{k,t} \right) \le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \frac{2^b \ln K}{\gamma^{(b)}} + \sum_{t=1}^{T} \sum_{i \in R_t} \frac{\gamma^{(b_t)} \widehat{\ell}_{i,t}^{(b_t)}}{|R_t|} + \sum_{t=1}^{T} \frac{\gamma^{(b_t)}}{2^{b_t+1}} \sum_{i \in V} p_{i,t}^{(b_t)} \bigl(\widehat{\ell}_{i,t}^{(b_t)}\bigr)^2 . \tag{10}
$$

Now, similarly to the proof of Theorem 1, we have that, for any $i$ and $t$, $\mathbb{E}_t\bigl[\widehat{\ell}_{i,t}^{(b_t)}\bigr] = \ell_{i,t}$ and $\mathbb{E}_t\bigl[(\widehat{\ell}_{i,t}^{(b_t)})^2\bigr] \le \frac{1}{q_{i,t}^{(b_t)}}$ . Hence, taking expectations $\mathbb{E}_t$ on both sides of (10) and recalling the definition of $Q_t^{(b)}$ gives

$$
\sum_{t=1}^{T} \left( \sum_{i \in V} p_{i,t}^{(b_t)} \ell_{i,t} - \ell_{k,t} \right) \le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \frac{2^b \ln K}{\gamma^{(b)}} + \sum_{t=1}^{T} \sum_{i \in R_t} \frac{\gamma^{(b_t)} \ell_{i,t}}{|R_t|} + \sum_{t=1}^{T} \frac{\gamma^{(b_t)}}{2^{b_t+1}} Q_t^{(b_t)} . \tag{11}
$$

Moreover,

$$
\sum_{t=1}^{T} \sum_{i \in R_t} \frac{\gamma^{(b_t)} \ell_{i,t}}{|R_t|} \le \sum_{t=1}^{T} \sum_{i \in R_t} \frac{\gamma^{(b_t)}}{|R_t|} = \sum_{t=1}^{T} \gamma^{(b_t)} = \sum_{b=0}^{\lfloor \log_2 K \rfloor} \gamma^{(b)} |T^{(b)}|
$$

and

$$
\sum_{t=1}^{T} \frac{\gamma^{(b_t)}}{2^{b_t+1}} Q_t^{(b_t)} = \sum_{b=0}^{\lfloor \log_2 K \rfloor} \frac{\gamma^{(b)}}{2^{b+1}} \sum_{t \in T^{(b)}} Q_t^{(b)} .
$$

Hence, plugging back into (11), taking outer expectations on both sides and recalling that $T^{(b)}$ is random (since the adversary adaptively decides which steps $t$ fall into $T^{(b)}$), we get

$$
\begin{aligned}
\mathbb{E}\big[L_{A,T} - L_{k,T}\big] &\le \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\frac{2^b \ln K}{\gamma^{(b)}} + \gamma^{(b)}|T^{(b)}| + \frac{\gamma^{(b)}}{2^{b+1}} \sum_{t \in T^{(b)}} Q_t^{(b)}\right] \\
&= \sum_{b=0}^{\lfloor \log_2 K \rfloor} \left(\frac{2^b \ln K}{\gamma^{(b)}} + \gamma^{(b)} \mathbb{E}\left[\sum_{t \in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}}\right)\right]\right) . \qquad (12)
\end{aligned}
$$

This establishes (2).

In order to prove inequality (3), we need to tune each $\gamma^{(b)}$ separately. However, a good choice of $\gamma^{(b)}$ depends on the unknown random quantity

$$
\overline{Q}^{(b)} = \sum_{t \in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}}\right) .
$$

To overcome this problem, we slightly modify Exp3-DOM by applying a doubling trick[12] to guess $\overline{Q}^{(b)}$ for each $b$. Specifically, for each $b = 0, 1, \ldots, \lfloor \log_2 K \rfloor$, we use a sequence $\gamma_r^{(b)} = \sqrt{(2^b \ln K)/2^r}$, for $r = 0, 1, \ldots$. We initially run the algorithm with $\gamma_0^{(b)}$. Whenever the algorithm is running with $\gamma_r^{(b)}$ and observes that $\sum_s \overline{Q}_s^{(b)} > 2^r$, where the sum is over all $s$ so far in $T^{(b)}$,[13] then we restart the algorithm with $\gamma_{r+1}^{(b)}$. Because the contribution of instance $b$ to (12) is

$$
\frac{2^b \ln K}{\gamma^{(b)}} + \gamma^{(b)} \sum_{t \in T^{(b)}} \left(1 + \frac{Q_t^{(b)}}{2^{b+1}}\right) ,
$$

the regret we pay when using any $\gamma_r^{(b)}$ is at most $2\sqrt{(2^b \ln K) 2^r}$. The largest $r$ we need is $\lceil \log_2 \overline{Q}^{(b)} \rceil$ and

$$
\sum_{r=0}^{\lceil \log_2 \overline{Q}^{(b)} \rceil} 2^{r/2} < 5\sqrt{\overline{Q}^{(b)}} .
$$

Since we pay regret at most 1 for each restart, we get

$$
\mathbb{E}\big[L_{A,T} - L_{k,T}\big] \le c \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\sqrt{(\ln K)\left(2^b|T^{(b)}| + \frac{1}{2}\sum_{t \in T^{(b)}} Q_t^{(b)}\right)} + \lceil \log_2 \overline{Q}^{(b)} \rceil\right] .
$$

for some positive constant $c$. Taking into account that

$$
\begin{aligned}
\sum_{b=0}^{\lfloor \log_2 K \rfloor} 2^b |T^{(b)}| &\le 2 \sum_{t=1}^{T} |R_t| \\
\sum_{b=0}^{\lfloor \log_2 K \rfloor} \sum_{t \in T^{(b)}} Q_t^{(b)} &= \sum_{t=1}^{T} Q_t^{(b_t)} \\
\sum_{b=0}^{\lfloor \log_2 K \rfloor} \lceil \log_2 \overline{Q}^{(b)} \rceil &= \mathcal{O}\big((\ln K)\ln(KT)\big) ,
\end{aligned}
$$

[12] The pseudo-code for the variant of Exp3-DOM using such a doubling trick is not displayed in this extended abstract.

[13] Notice that $\sum_s \overline{Q}_s^{(b)}$ is an observable quantity.

we obtain

$$\begin{aligned}
\mathbb{E}\big[L_{A,T} - L_{k,T}\big] &\le c \sum_{b=0}^{\lfloor \log_2 K \rfloor} \mathbb{E}\left[\sqrt{(\ln K)\left(2^b |T^{(b)}| + \frac{1}{2}\sum_{t \in T^{(b)}} Q_t^{(b)}\right)}\right] + \mathcal{O}\big((\ln K)\ln(KT)\big) \\
&\le c \,\lfloor \log_2 K \rfloor \mathbb{E}\left[\sqrt{\frac{\ln K}{\lfloor \log_2 K \rfloor}\sum_{t=1}^{T}\left(2|R_t| + \frac{1}{2} Q_t^{(b_t)}\right)}\right] + \mathcal{O}\big((\ln K)\ln(KT)\big) \\
&= \mathcal{O}\left((\ln K)\,\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\left(4|R_t| + Q_t^{(b_t)}\right)}\right] + (\ln K)\ln(KT)\right)
\end{aligned}$$

as desired. □

**Proof of Corollary 8**

We start off from the upper bound (3) in the statement of Theorem 7. We want to bound the quantities $|R_t|$ and $Q_t^{(b_t)}$ occurring therein at any step $t$ in which a restart does not occur —the regret for the time steps when a restart occurs is already accounted for by the term $\mathcal{O}\big((\ln K)\ln(KT)\big)$ in (3). Now, Lemma 11 gives

$$|R_t| = \mathcal{O}\big(\alpha(G_t)\ln K\big) \ .$$

If $\gamma_t = \gamma_t^{(b_t)}$ for any time $t$ when a restart does not occur, it is not hard to see that $\gamma_t = \Omega\big(\sqrt{(\ln K)/(KT)}\big)$. Moreover, Lemma 13 states that

$$Q_t = \mathcal{O}\big(\alpha(G_t)\ln(K^2/\gamma_t) + |R_t|\big) = \mathcal{O}\big(\alpha(G_t)\ln(K/\gamma_t)\big) \ .$$

Hence,

$$Q_t = \mathcal{O}\big(\alpha(G_t)\ln(KT)\big).$$

Putting together as in (3) gives the desired result. □