Online learning with noisy side observations

Tomáš Kocák
SequeL team, INRIA Lille
Villeneuve d'Ascq, France

Gergely Neu
Universitat Pompeu Fabra
Barcelona, Spain

Michal Valko
SequeL team, INRIA Lille
Villeneuve d'Ascq, France

Abstract

We propose a new partial-observability model for online learning problems where the learner, besides its own loss, also observes some *noisy* feedback about the other actions, depending on the underlying structure of the problem. We represent this structure by a weighted directed graph, where the edge weights are related to the quality of the feedback shared by the connected nodes. Our main contribution is an efficient algorithm that guarantees a regret of $\widetilde{\mathcal{O}}(\sqrt{\alpha^* T})$ after T rounds, where α^* is a novel graph property that we call the *effective independence number*. Our algorithm is completely parameter-free and does not require knowledge (or even estimation) of α^*. For the special case of binary edge weights, our setting reduces to the partial-observability models of Mannor and Shamir (2011) and Alon et al. (2013) and our algorithm recovers the near-optimal regret bounds.

1 Introduction

The general framework of online learning considers sequential decision-making problems where a *learner* repeatedly chooses actions so as to minimize the sum of losses assigned by the *environment* in response to the learner's actions. After making each decision, the learner observes some feedback about the losses assigned by the environment. Traditionally, the literature considers two types of feedback: *full-information* feedback (Cesa-Bianchi and Lugosi, 2006), where the learner observes the losses associated with *all* potential decisions and *bandit* feedback (Auer et al., 2002) where the learner only observes the loss of its own decisions. More recently, Mannor and Shamir (2011) proposed a partial-feedback scheme that models situations that lie between the two extremes: in their model, the learner observes losses associated with some additional actions besides its own loss. While this framework is often more realistic than either of the two extremes, it fails to address one important practical concern: in reality, one can rarely expect *perfect* side-observations to be available. In the current paper, we propose a similar model that can incorporate *imperfect* side-observations corrupted by various levels of noise, depending on the problem structure.

As an illustration to our setting, consider the problem of controlling solar panels so as to maximize their power production. In this problem, the learner has to repeatedly decide about the orientation of the panels so as to find alignments with strong sunshine. Besides the amount of the energy being actually produced in the current alignment, the learner can also possibly base its decisions on measurements of sensors installed on the solar panel. However, the observations generated by these sensors can be of variable quality depending on visibility conditions, the quality of the sensors and the alignment of the panels. Overall, this problem can be seen as a bandit problem with noisy side-observations fitting into our framework, where actions correspond to alignments and the noisy side observations give information about similar alignments.

Intuitively, in the case when the noise level of side observations does not change with time, a possible strategy one can think of is to use only the observations from the "most reliable" sources and ignore the rest. Having made the distinction between "reliable" and "unreliable", the learner could model the observation structure in the framework of Mannor and Shamir (2011); Alon et al. (2013), by treating every "reliable" observation as *perfect*. This approach raises two concerns. First, determining the cutoff for unreliable observations that allows the "most efficient" use of information is a highly nontrivial design choice. As we show later, knowing the *perfect cutoff* would help

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

us to improve performance over the pure bandit setting without side observations. Second, one has to address the *bias* arising from handling every reliable observation as perfect. While one can think of many obvious ways to handle this bias by appropriate weighting observations, none of these solutions are directly compatible with the model of Mannor and Shamir (2011); Alon et al. (2013). Our main contribution in this paper is an algorithm that is able to deal with both issues *without the knowledge of the optimal cutoff.*

The main tool we use for modeling uncertain observations is a *weighted directed graph* encoding the quality of side-observations. In this graph, the weight of the arc $i \to j$ measures the quality of the side observation obtained from action j when selecting action i. All weights are assumed to lie in the interval $[0, 1]$, with a weight of 1 corresponding to a perfectly accurate side observation, and a weight of 0 corresponding to a side observation of useless noise. Our model generalizes the previously considered models of Mannor and Shamir (2011) and Alon et al. (2013): their respective settings are captured by considering undirected and directed graphs with binary weights in our framework. In these special cases, the *independence number*[1] α of the observation graph plays a key role in characterizing the complexity of learning: the minimax regret after T rounds is known to be $\widetilde{\Theta}(\sqrt{\alpha T})$. In this paper, we define a similar quantity for weighted graphs: the *effective independence number* α^* and propose a learning algorithm that enjoys a regret bound of $\widetilde{\mathcal{O}}(\sqrt{\alpha^* T})$ without any conditions made on the loss sequence. The effective independence number α^* is closely related to the cutoff threshold for noisy observations. Intuitively, it is linked to the independence number of a graph that only considers reliable observations. In practical scenarios, neither the cutoff nor α^* is ever known to the learner. In any case, the most interesting situations for our setting are the cases when we can bound α^* by a small quantity.

While we are mainly inspired by situations where the weights of the graph are fixed and known in advance, we treat a more general setting where the observation structure can arbitrarily change over time and the weights are revealed to the learner only after it has made its decision. Our algorithm is fully adaptive in the sense that it does not require any prior knowledge of the sequence of observation graphs or the time horizon. To achieve this result, we combine the *implicit exploration* strategy introduced by Kocák et al. (2014) with a loss estimation technique that effectively suppresses the observation noise. For the special case of binary weights, the effective independence number and the independence number coincide; otherwise α^* is bounded by the number of actions N. Thus, the regret bound of our algorithm is of near-optimal order for binary graphs, and is always within logarithmic factors of the minimax regret of order $\sqrt{NT}$ for the standard multi-armed bandit problem without side observations. As we will show later in the paper, there are several interesting cases for which the effective independence number can be bounded in a nontrivial way.

Independently of the work presented in this paper, Wu et al. (2015) considered an essentially identical partial-observability model for online learning: there, side observations are modeled as zero-mean Guassian random variables with *variance* depending on the chosen action. It is easy to see that their model and ours can capture exactly the same type of problems: a side observation with zero variance in their model corresponds to a perfect observation with weight 1 in our model, while useless noise is equivalently represented by infinite-variance or zero-weight observations. The results of Wu et al. (2015) are, however, of a completely different flavor than the ones presented in the current paper; the primary difference being that Wu et al. assume that the losses are i.i.d. Gaussian random variables, while our results hold without any assumptions made on the sequence of losses. The main contributions of Wu et al. are (i) a general problem-dependent lower bound on the regret and (ii) algorithms that work under the assumption that all the useful (i.e., finite-variance) side-observations have the same variance. This latter assumption does not use the full strength of the framework where the variance of side observations can vary for different actions. Notably, the regret bounds presented in our paper match (up to logarithmic factors) the lower bounds of Wu et al. (2015) for the special cases that they consider. That said, their lower bounds and our upper bounds are not directly comparable for more general observability graphs.

Besides the works mentioned above, several other partial-observability models have been considered in the literature. The most general of these settings is the *partial-monitoring* framework considered by Bartók et al. (2011, 2014). Unlike our model, this framework is most useful for identifying and handling feedback structures that are *more restrictive* than bandit feedback. In contrast, our framework deals with feedback structures that are strictly more expressive than plain bandit feedback. Similarly to Bartók et al., the recent work of Alon et al. (2015) also considers a generalization of the partial-observability models of Mannor and Shamir (2011) and Alon et al. (2013) that may be more

[1] The independence number of a graph G is defined as the largest set of points in the graph such that no two points within this set are connected.

restrictive than bandit feedback. Another well-studied setting in machine learning is where the observations are corrupted by noise irrespective of the decisions of the learner (see, e.g., Cesa-Bianchi et al., 2010). Such settings do not pose an exploration-exploitation dilemma to the learner and is thus are not relevant to our goals.[2]

2 Background

Let us now give the formal definition of our learning problem. We consider a sequential decision-making problem where a *learner* and an *environment* interact in the following way (see also Figure 1). In every round $t \in [T] = \{1, 2, \dots, T\}$, the environment selects a weighted graph G_t with N nodes and a loss function $\ell_t : [N] \to [0, 1]$ where $\ell_{t,i}$ is the loss associated with arm i. The weight of each arc $i \to j$ in G_t is denoted as $s_{t,(i,j)}$ and assumed to lie in $[0, 1]$. Following the environment's move, the learner selects an *action* (or *arm*) $I_t \in [N]$ and incurs the loss ℓ_{t,I_t}. Finally, the learner also observes G_t and the feedback

$$c_{t,i} = s_{t,(I_t,i)} \cdot \ell_{t,i} + \big(1 - s_{t,(I_t,i)}\big) \cdot \xi_{t,i}$$

for every arm i, where $\xi_{t,i}$ is the *observation noise*. We assume that each $\xi_{t,i}$ is zero-mean, satisfies $|\xi_{t,i}| \leq R$ for some known constant $R \geq 0$, and is generated independently of all other noise terms and the history of the process[3]. The interaction history between the learner and the environment up to the end of round t is captured by the sigma-algebra $\mathcal{F}_t$. In this work, we consider *adaptive* (or *non-oblivious*) environments that are allowed to choose ℓ_t and G_t in full knowledge of the history $\mathcal{F}_{t-1}$. We also assume that all graphs G_t are such that $s_{t,(i,i)} = 1$ for all i, that is, the learner always observes its own loss ℓ_{t,I_t} without corruption.

The goal of the learner is to choose its actions so as to ensure that its cumulative loss grows as slowly as possible. As traditional in the online learning literature (Cesa-Bianchi and Lugosi, 2006), we measure the performance of the learner in terms of the (total expected) *regret* defined as the gap between the expected loss of the player and the expected loss of the best fixed-arm policy:

$$R_T = \max_{i \in [N]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \sum_{t=1}^{T} \ell_{t,i}\right].$$

[2]In fact, it can be shown by the techniques of Devroye et al. (2013) that in the setting of online learning with finite actions and observations corrupted by the same level of i.i.d. noise, the simplest possible strategy of *following the leader* gives near-optimal guarantees.

[3]We are mainly interested in the setting where $R = \Theta(1)$, that is, we are neither in the easy case where R is close to zero or the hard one where it may be as large as $\Omega(\sqrt{T})$

Parameters:
set of arms $[N]$, number of rounds T.
For all $t = 1, 2, \dots, T$ **repeat**

1. The environment picks a loss function $\ell_t : [N] \to [0, 1]$ and a directed weighted graph G_t with edge weights in $[0, 1]$.
2. Based on its previous observations (and possibly some source of randomness), the learner picks an action $I_t \in [N]$.
3. The learner suffers loss ℓ_{t,I_t}.
4. The learner observes G_t and the feedback
$$c_{t,i} = s_{t,(I_t,i)} \cdot \ell_{t,i} + \big(1 - s_{t,(I_t,i)}\big) \cdot \xi_{t,i}$$
for every arm $i \in [N]$.

Figure 1: The protocol of online learning with noisy observations.

In this paper, we are interested in constructing algorithms for the learner that guarantees a tight upper bound on the regret. Before proposing our algorithm, a few comments are in order. First, notice that our framework technically contains the settings of Mannor and Shamir (2011) and Alon et al. (2013) as special cases where the edge weights are chosen from $\{0, 1\}$: in this situation, our framework suggests that the learner either gets *perfect* side-observations or just zero-mean noise, which can be safely ignored by the learner. Also notice that since we assume $s_{t,(i,i)} = 1$ for all i, our problem is not harder for the learner than the standard multi-armed bandit problem. Indeed, thanks to this property, the learner could simply ignore all side-observations and run a bandit algorithm such as Exp3 of Auer et al. (2002) that guarantees a regret bound of $\mathcal{O}(\sqrt{NT \log N})$.

3 Algorithms and main result

This section presents our main contribution: a learning algorithm with strong theoretical performance guarantees for the setting described in the previous section. As the intuitions underlying our algorithm are rather intricate, we will proceed gradually: we first identify the main challenges of constructing learning algorithms for our setting, then offer a solution that overcomes these difficulties in an efficient manner.

A central concept in our performance guarantees is a new graph property that we call *effective independence number*, defined as follows:

Definition 1. *Let G be a weighted directed graph with N nodes and edge weights bounded in $[0, 1]$. For all*

$\varepsilon \in [0,1]$, let $G(\varepsilon)$ be the (unweighted) directed graph where arc $i \to j$ is present if and only if $s_{i,j} \geq \varepsilon$ in G. Letting $\alpha(\varepsilon)$ be the independence number of $G(\varepsilon)$, the effective independence number of G is defined as

$$\alpha^* = \min_{\varepsilon \in [0,1]} \frac{\alpha(\varepsilon)}{\varepsilon^2}.$$

Roughly speaking, the effective independence number is a measure of connectivity of weighted graphs. A detailed discussion of the effective independence number is deferred to Section 4. In what follows, we describe two learning algorithms that guarantee a regret bound depending on the effective independence numbers (α_t^*) of the observation graphs (G_t) as $\widetilde{\mathcal{O}}\left(\sqrt{\sum_t \alpha_t^*}\right)$.

For presenting our ideas (and our eventual algorithm), we take as template the seminal Exp3 algorithm of Auer et al. (2002), as presented by Bubeck and Cesa-Bianchi (2012) (see Algorithm 1). The main idea of this algorithm is maintaining an estimate $\widehat{\ell}_{t,i}$ of the losses $\ell_{t,i}$ for every t and i and choosing arm i with probability proportional to $\exp\bigl(-\eta_t \sum_{s=1}^{t-1} \widehat{\ell}_{s,i}\bigr)$ in round t, where $\eta_t > 0$ is a parameter of the algorithm often called the *learning rate*. The main challenge in constructing a learning algorithm for our setting is designing appropriate estimates for the losses. In particular, it is obvious that the learner should not rely on observations with high amount of noise in the same way as it relies on observations with almost no noise. One natural way to address this issue is explicitly distinguishing between "reliable" and "unreliable" side observations, and using only reliable sources for estimating losses. We first show that while this intuitive loss-estimation method does lead to strong performance guarantees, it requires a very careful choice of the cutoff parameter distinguishing reliable and unreliable sources. In Section 3.2, we propose our main algorithm that overcomes this issue and guarantees equally strong performance guarantees without having to explicitly distinguish between reliable and unreliable sources.

3.1 A naïve algorithm: Exp3-IXt

We first consider an algorithm that bases its decisions on the following estimates of each $\ell_{t,i}$:

$$\widehat{\ell}_{t,i}^{(\mathrm{B})} = \frac{c_{t,i}}{\sum_{j=1}^N p_{t,j} s_{t,(j,i)} + \gamma_t}. \tag{1}$$

where B stands for "basic". Here, $\gamma_t \geq 0$ is a so-called *implicit exploration* (or, in short, IX) parameter first used by Kocák et al. (2014) for decreasing the variance of importance-weighted estimates. Notice that setting

Algorithm 1 Algorithm template: Exp3 (Auer et al., 2002)

Initialization: $\widehat{L}_{0,i} = 0$ for all $i \in [N]$.
for $t = 1$ **to** T **do**
  Set η_t and γ_t.
  Construct the probability distribution $\boldsymbol{p}_t$ with.

$$p_{t,i} = \frac{\exp\bigl(-\eta_t \widehat{L}_{t-1,i}\bigr)}{\sum_{j=1}^N \exp\bigl(-\eta_t \widehat{L}_{t-1,j}\bigr)}.$$

  Play random arm I_t according to $\boldsymbol{p}_t$.
  Incur loss ℓ_{t,I_t}.
  Observe $c_{t,i} = s_{t,(I_t,i)}\ell_{t,i} + (1 - s_{t,(I_t,i)})\xi_{t,i}$ for all $i \in [N]$.
  Observe graph G_t.
  Construct loss estimates $\widehat{\ell}_{t,i}$.
  Set $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \widehat{\ell}_{t,i}$.
end for

$\gamma_t = 0$, makes estimates above unbiased since

$$\mathbb{E}\left[c_{t,i} \,\middle|\, \mathcal{F}_{t-1}\right] = \left(\sum_{j=1}^N p_{t,j} s_{t,(j,i)}\right) \cdot \ell_{t,i},$$

where we used our assumption that $\mathbb{E}\left[\xi_{t,i}\right] = 0$. Using these estimates in our algorithmic template Exp3 (see Algorithm 1), one would expect to get reasonable performance guarantees. Unfortunately however, we were not able to prove a performance guarantee for the resulting algorithm.

A close examination reveals that the reason for the poor performance of the above algorithm is the large variance of the estimates (1) which is caused by including observations from "unreliable sources" with small weights. One intuitive idea is to explicitly draw the line between reliable and unreliable sources by cutting connections with weights under a certain threshold. This effect is realized by the estimates

$$\widehat{\ell}_{t,i}^{(\mathrm{T})} = \frac{c_{t,i} \mathbb{I}_{\{s_{t,(I_t,i)} \geq \varepsilon_t\}}}{\sum_{j=1}^N p_{t,j} s_{t,(j,i)} \mathbb{I}_{\{s_{t,(j,i)} \geq \varepsilon_t\}} + \gamma_t}, \tag{2}$$

where $\varepsilon_t \in [0,1]$ is a threshold value and T stands for "thresholded". We call the algorithm resulting from using the above estimates in Algorithm 1 Exp3-IXt, standing for "Exp3 with Implicit eXploration and Truncated side-observation weights". Thanks to the thresholding operation, the variance of the loss estimates can be nicely controlled and it becomes possible to prove a strong performance guarantee for Exp3-IXt. In particular, we prove the following result about the regret of Exp3-IXt:

Theorem 1. *For all t, let α_t^* be the effective independence number of G_t. Then, there exists a setting of (η_t) and (γ_t) for which the regret of* Exp3-IXt *is bounded as*

$$R_T = \widetilde{\mathcal{O}}\left((1+R)\sqrt{\sum_{t=1}^{T}\frac{\alpha(G_t(\varepsilon_t))}{\varepsilon_t^2}}\right).$$

The theorem is proved in the Appendix. Note that if we choose $\varepsilon_t = \arg\min_{\varepsilon\in[0,1]}\frac{\alpha(G_t(\varepsilon))}{\varepsilon^2}$ for all t, the above bound essentially becomes $\widetilde{\mathcal{O}}(\sqrt{\alpha^*_{\text{avg}}T})$ where $\alpha^*_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T}\alpha_t^*$ is the average effective independence number of the sequence of graphs played by the environment. Note however that tuning ε_t can be a very challenging task in practice, since computing independence numbers in general is known to be NP-hard. Even worse, computing the *effective* independence number of a weighted graph can require computing up to N^2 independence numbers. In the next section, we propose an adaptive algorithm that does not need to tune this parameter and still manages to guarantee the same regret bound.

3.2 An adaptive algorithm: Exp3-WIX

This section presents our main algorithm that obtains strong regret bounds without having to estimate any effective independence numbers. The key element of this algorithm is using loss estimates of the form

$$\widehat{\ell}_{t,i} = \frac{s_{t,(I_t,i)}\cdot c_{t,i}}{\sum_{j=1}^{N} p_{t,j}s^2_{t,(j,i)} + \gamma_t}, \tag{3}$$

where $\gamma_t \geq 0$ is again the so-called implicit exploration parameter already introduced in the previous section. Notice that the difference from the estimates (1) is that the observation $c_{t,i}$ is multiplied by the weight of useful information in $c_{t,i}$ and the denominator is modified accordingly, so that the estimates are unbiased when setting $\gamma_t = 0$ since

$$\mathbb{E}\left[s_{t,(I_t,i)}\cdot c_{t,i}\middle|\,\mathcal{F}_{t-1}\right] = \left(\sum_{j=1}^{N} p_{t,j}s^2_{t,(j,i)}\right)\cdot \ell_{t,i}.$$

The role of this scaling is pulling the noise term $\xi_{t,i}$ toward zero for actions i with small weights $s_{I_t,i}$, and thus achieving a similar variance-reducing effect as the truncations employed by Exp3-IXt.

Armed with the loss estimates (3), we are ready to define our algorithm: Exp3 (presented as Algorithm 1) with Weighted observations and Implicit eXploration, or, in short, Exp3-WIX. Overall, Exp3-WIX has two set of parameters to tune: the sequence of learning rates $(\eta_t)_t$ and the sequence of IX parameters $(\gamma_t)_t$. Our main theorem below states the performance guarantees of Exp3-WIX with an adaptive learning-rate sequence that does not need any prior knowledge about the number of rounds or the nature of the side-observation graphs. The key quantity for computing the parameters η_t and γ_t is

$$Q_t = \sum_{i=1}^{N}\frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j}s^2_{t,(j,i)} + \gamma_t},$$

defined for all t.

Theorem 2. *For all t, let α_t^* be the effective independence number of G_t. Then, setting*

$$\eta_t = \sqrt{\frac{\log N}{2(1+R+R^2)(N+\sum_{s=1}^{t-1}Q_s)}}$$

and $\gamma_t = R\eta_t$, the regret of Exp3-WIX *is bounded as*

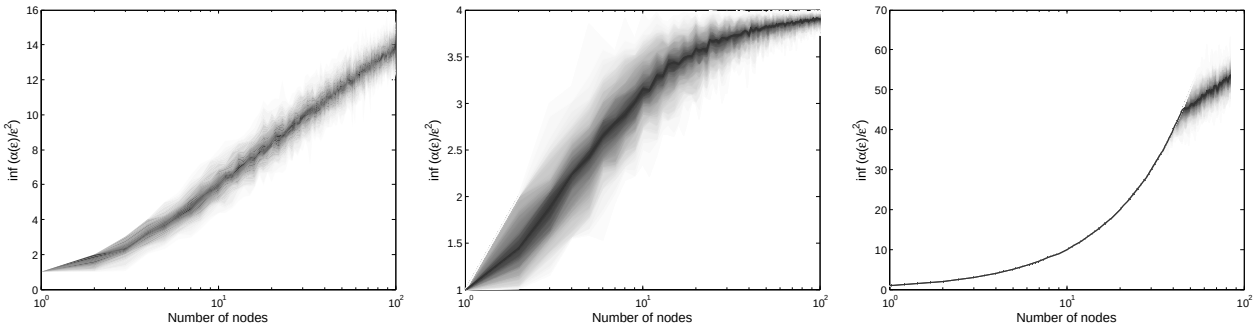
$$R_T = \widetilde{\mathcal{O}}\left((1+R)\sqrt{N+\sum_{t=1}^{T}\alpha_t^*}\right).$$

The theorem is proved in Section 5. In plain words, Theorem 2 guarantees that the regret of Exp3-WIX grows as $\widetilde{\mathcal{O}}(\sqrt{\alpha^*_{\text{avg}}T})$. Notice that in order to obtain this regret bound, Exp3-WIX never needs to compute the effective independence number of any of the observation graphs. This saves us from a significant computational overhead as compared to the naïve algorithm Exp3-IXt that needed to set a truncation parameter to discard unreliable observations.

4 The effective independence number

The previous section has established that the performance guarantees of our algorithms can be expressed in terms of the effective independence number of the observation graphs. In this section, we provide some basic insights about the nature of this quantity and describe some graph structures with small effective independence numbers.

The first observation we make is that the effective independence number is always well-defined, as the function $\alpha(\varepsilon)/\varepsilon^2$ can be easily shown to be piecewise decreasing and lower semicontinuous with at most N discontinuities. Thanks to these properties, this expression takes its minimum within the closed interval $[0,1]$. Second, we note that the effective independence number of any weighted graph is trivially bounded by the number N of the nodes in the graph. This follows from the fact that $\alpha^* \leq \alpha(1)/1 \leq N$. This essentially guarantees that incorporating side-observations can never be harmful to the performance of the learner:



(a) $U(0,1)$ weights (b) $U(\frac{1}{2},1)$ weights (c) $U(0,\frac{1}{2})$ weights

Figure 2: Dependence of α^* on the size of the graph with random weights, 100 graphs for each size.

the regret of Exp3-WIX is always within logarithmic factors of the minimax regret of order $\sqrt{NT}$ for the standard multi-armed bandit problem without side observations.

It is also easy to see that the effective independence number exactly matches the independence number if all edge weights are binary. This in particular implies that for such graphs, the regret of Exp3-WIX grows at the minimax rate established by Alon et al. (2013) up to logarithmic factors, matching the performance guarantees of the algorithms of Alon et al. (2013) and Kocák et al. (2014). Another interesting case is when all weights are either zero or equal to a fixed constant ε, also assuming $s_{i,i} = \varepsilon$. In this case, the effective independence number becomes $\frac{\alpha}{\varepsilon^2}$, where α is the independence number of the underlying unweighted graph. This case was studied in the recent paper of Wu et al. (2015), who show (in their Corollary 4) that the *minimax* regret in this case is of $\Theta(\sqrt{\alpha T}/\varepsilon)$—implying that our performance bounds for this case are again near-optimal[4]. Also observe that whenever all weights are bounded by some constant $c > 0$ from below, the effective independence number becomes upper-bounded by $1/c^2$, *irrespective of the number of actions*. That is, our algorithm can achieve an *exponential* performance gain over bandit algorithms in terms of N by leveraging such feedback structures.

Let us now describe a class of weighted graphs with bounded effective independence numbers. Consider a geometric graph whose nodes represent vertices of a uniform $k \times k$ grid on $[0,1]^2$. The weight of edge (i,j) is given as $1/(1+d_{i,j}^2)$, where $d_{i,j}$ is the Euclidean distance of the respective vertices represented by i and j. This graph can be used to model a sensor network where the measurement accuracy of measurements degrades with the distance. Thus, reading the measurements from one sensor will give information about the measurements of nearby sensors as well. Intuitively, increasing the number of sensors (i.e., refining the grid) should only improve the information-sharing between sensors up to a certain level. It is natural to expect a reasonable graph property quantifying the information-sharing efficiency to capture this intuition. We have numerically evaluated the effective independence number of a number of graphs from the above family to test if it satisfies the above criterion. We have found that the effective independence numbers remain bounded by a *constant* (roughly 30) even when refining the grid infinitely, confirming that the effective independence number captures the above phenomenon.

Finally, we conducted some numerical simulations to evaluate the average effective independence numbers of certain types of weighted random graphs. In particular, we considered random graphs with i.i.d. weights distributed uniformly on $[0,1]$, $[\frac{1}{2},1]$ and $[0,\frac{1}{2}]$. The distributions of the effective independence numbers are illustrated as scatter plots for different graph sizes on Figure 2. First, observe that the average α^* of $U(0,1)$-weighted graphs shows a logarithmic trend in terms of N. The results concerning $U(\frac{1}{2},1)$-weighted graphs are not surprising given that we have already established that graphs with bounded weights have finite effective independence numbers. For $U(0,\frac{1}{2})$-weighted graphs, we see that α^* grows linearly up until a certain threshold, when it starts to follow a logarithmic trend. The intuition behind this linear behavior for small graphs is the following. First, observe that the optimal value of ε is greater than $1/\sqrt{N}$. That is, until N is large enough so that a critical mass of edges are above this quantity, the optimal value of $\alpha(\varepsilon)/\varepsilon^2$ remains N. Once N is beyond this critical value, α^* starts following a logarithmic trend.

[4]While we prove our bounds for the case where $s_{i,i} = 1$ for all i, it is easy to extend our results to the case where all such weights equal a constant in $[0,1]$.

5 Analysis

Let us now turn to proving Theorem 2. Our analysis is slightly more general than necessary for proving Theorem 2, in that they allow any sequence of learning rates and IX parameters. To avoid clutter, we will omit the t indices from $s_{t,(\cdot,\cdot)}$. In principle, our analysis combines (more-or-less) standard tools for analyzing Exp3 with adaptive learning rates and ideas from Alon et al. (2013) and Kocák et al. (2014), while also heavily exploiting the structure of our loss estimates (3). In particular, these estimates allow us to bound the expected regret of Exp3-WIX in terms of the quantities (Q_t).

Lemma 1. *Let $(\eta_t)_t$ and $(\gamma_t)_t$ be two $(\mathcal{F}_t)$-measurable non-increasing sequences satisfying $\gamma_t \geq \eta_t R$ for all t. Then, the expected regret of* Exp3-WIX *is bounded as*

$$R_T \leq \mathbb{E}\left[\frac{\log N}{\eta_T} + \sum_{t=1}^{T}\left(\gamma_t + (1+R^2)\eta_t\right) Q_t\right].$$

The full proof of the lemma is delegated to the Appendix. Below, we provide a brief sketch covering the key parts of the proof.

Proof sketch. By straightforward adaptation of the techniques of Auer et al. (2002); Bubeck and Cesa-Bianchi (2012); Györfi and Ottucsák (2007); Kocák et al. (2014), we can prove the bound

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\right] \leq \mathbb{E}\left[\widehat{L}_{T,j} + \frac{\log N}{\eta_T}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\eta_t \sum_{i=1}^{N} p_{t,i}\left(\widehat{\ell}_{t,i}\right)^2\right].$$

for any fixed j Thus, we are left with the problem of relating the left-hand side to the total expected loss of the learner and to upper-bounding the right-hand side. As the first step, observe that

$$\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] = \sum_{i=1}^{N} p_{t,i}\frac{\sum_{j=1}^{N} p_{t,j}s_{j,i}^2\ell_{t,i}}{\sum_{j=1}^{N} p_{t,j}s_{j,i}^2 + \gamma_t} \geq \sum_{i=1}^{N} p_{t,i}\ell_{t,i} - \gamma_t Q_t,$$

where we used $\mathbb{E}\left[\xi_{t,i}|\mathcal{F}_{t-1}\right] = 0$ in the first step and $s_{j,i} \leq 1$ in the second. The first term on the right-hand side can be bounded by $L_{T,j}$ by observing that $\mathbb{E}\left[\widehat{\ell}_{t,j}\middle|\mathcal{F}_{t-1}\right] \leq \ell_{t,j}$ holds for all fixed j by the definition of the loss estimates (3). Finally, the last term is bounded as

$$\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\left(\widehat{\ell}_{t,i}\right)^2\middle|\mathcal{F}_{t-1}\right] \leq \sum_{i=1}^{N} p_{t,i}\frac{\sum_{j=1}^{N} p_{t,j}s_{j,i}^2(1+R^2)}{\left(\sum_{j=1}^{N} p_{t,j}s_{j,i}^2 + \gamma_t\right)^2} \leq \sum_{i=1}^{N} p_{t,i}\frac{1+R^2}{\left(\sum_{j=1}^{N} p_{t,j}s_{j,i}^2 + \gamma_t\right)} = (1+R^2)Q_t.$$

The statement of the lemma follows from putting everything together. □

Observe that since we assume $s_{i,i} = 1$ for all i, Q_t can be trivially bounded by N. As a result, it is straightforward to show that the regret of Exp3-WIX is of order $\sqrt{TN\log N}$. The remaining challenge is thus bounding Q_t in a nontrivial way, capturing the structure of the observation graph G_t. The following lemma provides such a bound in terms of the effective independence number of G_t.

Lemma 2. *Let α_t^* be the effective independence number of G_t. Then, for any positive γ_t,*

$$Q_t \leq 2\alpha^*\left(1 + \log\left(1 + \frac{N^2/\gamma_t + N^2 + N}{\alpha^*}\right)\right).$$

The proof of this statement builds on results by Alon et al. (2013) and Kocák et al. (2014). Below, we give a short sketch of the full proof that is given in the Appendix.

Proof sketch. Let us define $\varepsilon_* = \arg\min_{\varepsilon\in[0,1]} \alpha(\varepsilon)/\varepsilon^2$ and observe that

$$\frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j}s_{j,i}^2 + \gamma_t} \leq \frac{1}{\varepsilon^2}\cdot\frac{p_{t,i}}{p_{t,i} + \sum_{j\neq i} p_{t,j}\mathbb{I}_{\{s_{j,i}\geq\varepsilon\}} + \gamma_t}$$

holds for all $\varepsilon \in [0,1]$, and in particular for ε_*. Applying a variant of Lemma 1 in Kocák et al. (2014) to the binary graph $G_t(\varepsilon_*)$, we obtain

$$Q_t \leq \frac{2}{\varepsilon_*^2}\cdot\left(\alpha_t(\varepsilon_*)\log\left(1 + \frac{\varepsilon_*^2N^2/\gamma_t + N + 1}{\alpha_t(\varepsilon_*)}\right) + \frac{2}{\varepsilon_*^2}\right).$$

The statement of the lemma follows from using the trivial bound $\alpha_t(\varepsilon_*) \geq 1$. □

Now, every ingredient is ready for proving Theorem 2. In particular, plugging in the choice of the parameters η_t and γ_t into the bound of Lemma 1 and applying Lemma 3.5 of **?**, we obtain

$$R_T \leq 2\sqrt{2(1+R+R^2)\left(N + \sum_{t=1}^{T} Q_t\right)\log N}.$$

Then, the statement of the theorem follows from combining the above with the bound of Lemma 2.

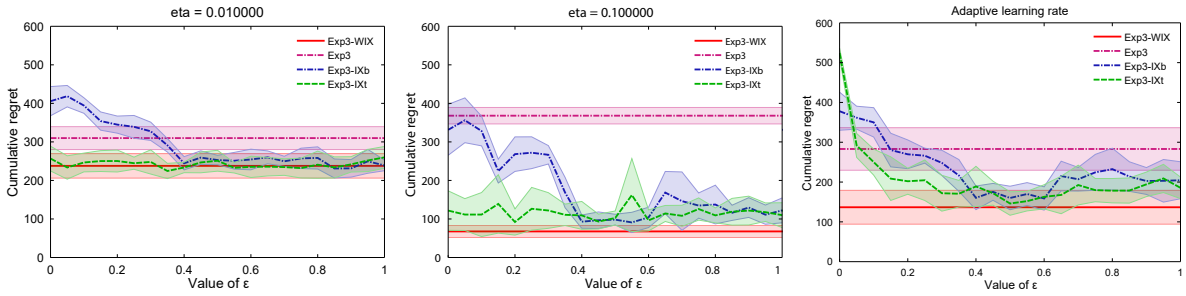


Figure 3: Comparison of total regrets of the algorithms at time T for static and adaptive learning rates.

6 Experiments

In this section, we empirically compare Exp3-WIX to some of its natural competitors: Exp3-IXt, vanilla Exp3 that ignores all side observations and a straightforward variation of the Exp3-IX algorithm of Kocák et al. (2014). This latter algorithm, referred to as Exp3-IXb (with "b" standing for "basic"), uses a threshold ε to decide which observations are too noisy to use and which are the ones to be retained: All the edges with weights smaller than a parameter ε are deleted and the rest of the weights are set to 1. The algorithm then plays basic Exp3-IX for the resulting binary graph. That is, the difference between Exp3-IXt and Exp3-IXb is that the latter does not adjust for the bias arising from using unreliable side observations. Note that Exp3-IXb comes without any formal performance guarantee.

For the purpose of the experiments, we assumed to have 25 actions forming 5×5 grid embedded in a plane. The distance of neighbors in the grid was set to be 1. Using this structure, we defined the weight connecting two nodes as $\min\left\{3/d^2, 1\right\}$, and d is the Euclidean distance between actions in the grid. This choice is motivated by the fact that the intensity of many physical phenomena decays proportionally to the inverse square of the distance (e.g., gravitational force, electromagnetic phenomena).

A simple idea for constructing synthetic loss sequences is letting the instantaneous loss of each action evolve as a random walk with small Gaussian increments (with appropriate truncations when the loss goes beyond the $[0, 1]$ interval). In our experiments, we took this idea one step further: We constructed 20 independent random walks for each action and alternated them, that is, we used one random walk each to define every twentieth loss. Using this procedure, we generated a single loss sequence of $T = 5,000$ steps to test the algorithms. For a fair comparison, we ran each algorithm for their respective theoretically motivated adaptive learning rates, and also for a number of static learning rates between 0.001 and 1. For static learning rates, we observed the best performance of Exp3 for learning rates around 0.01, all the other algorithms did well for learning rates around 0.1. Due to the lack of space, we included plots only for these two learning rates.

We ran Exp3-IXb and Exp3-IXt for several values of ε from 0 to 1. In all experiments, we set the implicit exploration parameters to zero. This is well-justified in the case of undirected graphs, as shown by the analysis of Alon et al. (2013). Figure 3 shows the performance of the algorithms for $\eta = 0.01$, $\eta = 0.1$ and the adaptive learning rates for each algorithm as a function of the threshold parameter ε. Each curve on this graph is the average of the total regrets measured in 10 independent runs with error bars proportional to the empirical standard deviation.

Our experiments confirm that guessing the right value for the threshold parameter is indeed a very difficult problem: while Exp3-WIX performs consistently well for all parameter settings, Exp3-IXt and Exp3-IXb only perform reasonably well for moderate values of ε that are not supported by theory. In fact, the value of ε optimizing $\alpha(\varepsilon)/\varepsilon^2$ is 1, which is shown to perform poorly in the experiments. Perhaps surprisingly, Exp3-IXb performs well despite the obvious bias in its loss estimates. The performance of Exp3 is significantly worse than Exp3-WIX, confirming the benefit of side-observations, however noisy they are.

7 Conclusions and open problems

The main contribution of our work is introducing a new partial-observability model for adversarial online learning and proposing an efficient learning algorithm with rigorous performance guarantees for this setting. Our regret bounds depend on a newly introduced graph property that we call the effective independence number. While the recent results of Wu et al. (2015) suggest that our bounds are minimax optimal in some special cases of our framework, it is not yet known whether the effective independence number is the exact quantity that characterizes the minimax regret in general—we leave this exciting question open for future investigation.

References

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Annual Conference on Learning Theory*.

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2013). From Bandits to Experts: A Tale of Domination and Independence. In *Neural Information Processing Systems*.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2-3):235–256.

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

Bartók, G., Pál, D., and Szepesvári, C. (2011). Minimax regret of finite partial-monitoring games in stochastic environments. *COLT*, 2011:133–154.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5:1–122.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York, NY.

Cesa-Bianchi, N., Shalev-Shwartz, S., and Shamir, O. (2010). Online learning of noisy data with kernels. In *COLT*, pages 218–231.

Devroye, L., Lugosi, G., and Neu, G. (2013). Prediction by random-walk perturbation. *Proceedings of the Twenty-Sixth Conference on Learning Theory (COLT 2013)*.

Györfi, L. and Ottucsák, b. (2007). Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory*, 53(5):866–1872.

Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems 27*.

Mannor, S. and Shamir, O. (2011). From Bandits to Experts: On the Value of Side-Observations. In *Neural Information Processing Systems*.

Wu, Y., György, A., and Szepesvári, C. (2015). Online learning with gaussian payoffs and side observations. To appear in Advances in Neural Information Processing Systems 28. `http://www.szit.bme.hu/~gya/publications/WuGySz15nips.pdf`.

A Proof of Lemma 1

The first part of the analysis is similar to the analysis of the basic EXP3 algorithm besides, we are using an adaptive learning rate η_t to obtain anytime regret bound, and therefore, we do not need to know the stopping time T of the algorithm. We start by introducing some notation. Let

$$W_t = \frac{1}{N}\sum_{i=1}^{N} e^{-\eta_t \widehat{L}_{t-1,i}} \qquad \text{and} \qquad W'_t = \frac{1}{N}\sum_{i=1}^{N} e^{-\eta_{t-1} \widehat{L}_{t-1,i}}.$$

Following the proof of Lemma 1 of Györfi and Ottucsák (2007), we track the evolution of $\log(W'_{t+1}/W_t)$ to control the regret. We have

$$\begin{aligned}
\frac{1}{\eta_t}\log\frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t}\log\sum_{i=1}^{N}\frac{\frac{1}{N}e^{-\eta_t \widehat{L}_{t,i}}}{W_t} = \frac{1}{\eta_t}\log\sum_{i=1}^{N}\frac{\frac{1}{N}e^{-\eta_t \widehat{L}_{t-1,i}}e^{-\eta_t \widehat{\ell}_{t,i}}}{W_t} \\
&= \frac{1}{\eta_t}\log\sum_{i=1}^{N} p_{t,i} e^{-\eta_t \widehat{\ell}_{t,i}} \le \frac{1}{\eta_t}\log\sum_{i=1}^{N} p_{t,i}\left(1-\eta_t\widehat{\ell}_{t,i}+(\eta_t\widehat{\ell}_{t,i})^2\right) \qquad (4) \\
&= \frac{1}{\eta_t}\log\left(1-\eta_t\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}+\eta_t^2\sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2\right),
\end{aligned}$$

where in (4), we used the inequality $\exp(-x) \le 1 - x + x^2$ that holds for $x \ge -1$. The use of this inequality is made possible by the definitions of η_t and γ_t that guarantee $\eta_t\widehat{\ell}_{t,i} \ge -1$ for all i. Further, we use the inequality $\log(1-x) \le -x$, which holds for all x, to upper bound last term

$$\begin{aligned}
\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i} &\le \left[\frac{\log W_t}{\eta_t} - \frac{\log W'_{t+1}}{\eta_t}\right] + \sum_{i=1}^{N}\eta_t p_{t,i}(\widehat{\ell}_{t,i})^2 \\
&= \left[\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right) + \left(\frac{\log W_{t+1}}{\eta_{t+1}} - \frac{\log W'_{t+1}}{\eta_t}\right)\right] + \sum_{i=1}^{N}\eta_t p_{t,i}(\widehat{\ell}_{t,i})^2. \qquad (5)
\end{aligned}$$

The second term in brackets on the right hand side is upper bounded by zero, since

$$W_{t+1} = \sum_{i=1}^{N}\frac{1}{N}e^{-\eta_{t+1}\widehat{L}_{t,i}} = \sum_{i=1}^{N}\frac{1}{N}\left(e^{-\eta_t\widehat{L}_{t,i}}\right)^{\frac{\eta_{t+1}}{\eta_t}} \le \left(\sum_{i=1}^{N}\frac{1}{N}e^{-\eta_t\widehat{L}_{t,i}}\right)^{\frac{\eta_{t+1}}{\eta_t}} = (W'_{t+1})^{\frac{\eta_{t+1}}{\eta_t}},$$

Using Jensen's inequality to the concave function $x^{\frac{\eta_{t+1}}{\eta_t}}$ for $x \in \mathbb{R}$. The function is concave since $\eta_{t+1} \le \eta_t$ by definition. Taking logarithms in the above inequality, we get

$$\frac{\log W_{t+1}}{\eta_{t+1}} - \frac{\log W'_{t+1}}{\eta_t} \le 0.$$

Using this inequality, we can simplify (5)

$$\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i} \le \eta_t\sum_{i=1}^{N} p_{t,i}\left(\widehat{\ell}_{t,i}\right)^2 + \left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right).$$

Taking *conditional* expectations with respect to the σ-algebra $\mathcal{F}_{t-1}$, generated by the history up to time $t-1$, and summing up both sides over the time, we get

$$\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] \le \sum_{t=1}^{T}\mathbb{E}\left[\eta_t\sum_{i=1}^{N} p_{t,i}\left(\widehat{\ell}_{t,i}\right)^2\middle|\mathcal{F}_{t-1}\right] + \sum_{t=1}^{T}\mathbb{E}\left[\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\middle|\mathcal{F}_{t-1}\right]. \qquad (6)$$

For the following part of the analysis, we use a slightly more general form of our loss estimates:

$$\widehat{\ell}_{t,i} = \sum_{i=1}^{N} p_{t,i} \frac{s_{I_t,i}^{\delta} c_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t} = \sum_{i=1}^{N} p_{t,i} \frac{s_{I_t,i}^{1+\delta} \ell_{t,i} + s_{I_t,i}^{\delta}(1 - s_{I_t,i})\xi_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t}.$$

Later we show that $\delta = 1$ is optimal, which recovers the loss estimates (3). The next step is to bound the three expectations involved in Equation (6). For the first expectation we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\frac{s_{I_t,i}^{\delta} c_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t}\middle|\mathcal{F}_{t-1}\right] = \sum_{i=1}^{N} p_{t,i}\frac{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta}\ell_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t} \\
&\geq \sum_{i=1}^{N} p_{t,i}\ell_{t,i} - \gamma_t \sum_{i=1}^{N} \frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t} = \sum_{i=1}^{N} p_{t,i}\ell_{t,i} - \gamma_t Q_t(\delta),
\end{aligned}$$

where

$$Q_t(\delta) = \sum_{i=1}^{N} \frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t}.$$

For the second expectation we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2\middle|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{N} p_{t,i}\frac{\mathbb{E}\left[s_{I_t,i}^{2+2\delta}\middle|\mathcal{F}_{t-1}\right]\ell_{t,i}^2 + \mathbb{E}\left[s_{I_t,i}^{2\delta}(1-s_{I_t,i})^2\middle|\mathcal{F}_{t-1}\right]\mathbb{E}\left[\xi_{t,i}^2\middle|\mathcal{F}_{t-1}\right]}{\left(\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t\right)^2} \\
&\leq \sum_{i=1}^{N} p_{t,i}\frac{\sum_{j=1}^{N} p_{t,j} s_{j,i}^{2+2\delta} + \sum_{j=1}^{N} p_{t,j} s_{j,i}^{2\delta} R^2}{\left(\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t\right)^2} \\
&\leq \sum_{i=1}^{N} p_{t,i}\frac{1+R^2}{\left(\sum_{j=1}^{N} p_{t,j} s_{j,i}^{1+\delta} + \gamma_t\right)} = (1+R^2)Q_t(\delta)
\end{aligned}$$

Where the last inequality holds for $\delta \geq 1$. For the third expectation we have

$$-\mathbb{E}\left[\frac{\log W_{T+1}}{\eta_{T+1}}\right] \leq \min_{k\in[N]}\left(-\mathbb{E}\left[\frac{\log \frac{1}{N} e^{-\eta_T \widehat{L}_{T,k}}}{\eta_T}\right]\right) = \mathbb{E}\left[\frac{\log N}{\eta_T}\right] + \min_{k\in[N]}\left(\mathbb{E}\left[\widehat{L}_{T,k}\right]\right).$$

To conclude, observe that $Q_t(1) \leq Q_t(d)$ holds almost surely and thus we can set $\delta = 1$. Then the statement of the lemma follows from combining all of the bounds above. □

B Proof of Lemma 2

As done in the analysis of Alon et al. (2013); Kocák et al. (2014), we use following two lemmas to bound Q_t.

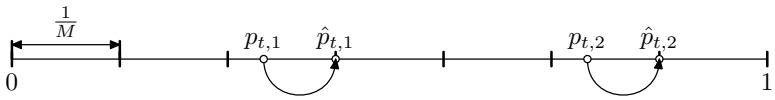
Lemma 3. (cf. Lemma 10 of Alon et al. (2013)) *Let G be a directed graph, with $V = \{1, \ldots, N\}$. Let d_i^- be the indegree of the node i and $\alpha = \alpha(G)$ be the independence number of G. Then*

$$\sum_{i=1}^{N} \frac{1}{1 + d_i^-} \leq 2\alpha \log\left(1 + \frac{N}{\alpha}\right).$$

Lemma 4. (cf. Lemma 12 of Alon et al. (2013)) *If $a, b \geq 0$ and $a + b \geq B > A > 0$, then*

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

Before using the previous lemmas, we need to discretize the values of $p_{t,i}$. Let $\hat{p}_{t,i}$ be the discretized version of $p_{t,i}$ which satisfies $\hat{p}_{t,i} = k/M$ for some integer k, $M = \lceil \varepsilon^2 N^2/\gamma_t \rceil$, and $\hat{p}_{t,i} - 1 < p_{t,i} \leq \hat{p}_{t,i}$.



Using Lemma 4 for $a = \hat{p}_{t,i}$, $b = \sum_{j\neq i} \hat{p}_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\} + \gamma/\varepsilon$, $B = \gamma/\varepsilon$, and $A = N/M$ we get

$$\begin{aligned}
\frac{p_{t,i}}{p_{t,i} + \sum_{j\neq i} p_{t,j} s_{j,i}^2 + \gamma_t} &\leq \frac{p_{t,i}}{\varepsilon^2 p_{t,i} + \sum_{j\neq i} \varepsilon^2 p_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\} + \gamma_t} \\
&= \frac{1}{\varepsilon^2} \frac{p_{t,i}}{p_{t,i} + \sum_{j\neq i} p_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\} + \gamma_t/\varepsilon^2} \\
&\leq \frac{1}{\varepsilon^2} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j\neq i} \hat{p}_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\} + \gamma_t/\varepsilon^2 - N/M} \\
&\leq \frac{1}{\varepsilon^2} \left(\frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j\neq i} \hat{p}_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\}} + \frac{N/M}{\gamma_t/\varepsilon^2 - N/M} \right) \\
&\leq \frac{1}{\varepsilon^2} \left(\frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j\neq i} \hat{p}_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\}} + \frac{2}{N} \right).
\end{aligned}$$

From this point, one can follow the proof of Lemma 1 in Kocák et al. (2014) to prove

$$\begin{aligned}
Q_t &\leq \frac{1}{\varepsilon^2} \sum_{i=1}^{N} \frac{\hat{p}_{t,i}}{\hat{p}_{t,i} + \sum_{j\neq i} \hat{p}_{t,j}\mathbb{I}\{s_{j,i} \geq \varepsilon\}} + \frac{2}{\varepsilon^2} \leq \frac{2}{\varepsilon^2}\alpha_t(\varepsilon) \log\left(1 + \frac{M+N}{\alpha_t(\varepsilon)}\right) + \frac{2}{\varepsilon^2} \\
&\leq \frac{2}{\varepsilon^2}\alpha_t(\varepsilon) \log\left(1 + \frac{N^2/\gamma_t + N/\varepsilon^2 + 1/\varepsilon^2}{\alpha_t(\varepsilon)/\varepsilon^2}\right) + \frac{2}{\varepsilon^2},
\end{aligned}$$

This bound holds for every $\varepsilon \in [0,1]$, therefore, it holds also for ε_*. Finally, using $1/\varepsilon^2 \leq N$ and $\alpha(\varepsilon) \geq 1$, we can recover the statement of the lemma. □

C The proof of Theorem 1

The proof of Theorem 1 roughly follows the proof of Theorem 2 with one key difference. For simplicity, let us define

$$Q'_t = \sum_{i=1}^{N} \frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i} \geq \varepsilon_t\}} + \gamma_t}$$

and consider an oblivious adversary that chooses the whole sequence of observations graphs deterministically before the first round. Our starting point is the bound of Equation (6), which also holds for Exp3-IXt. First, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i} \middle| \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{i=1}^{N} p_{t,i} \frac{c_{t,i}\mathbb{I}_{\{s_{I_t,i} \geq \varepsilon_t\}}}{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i} \geq \varepsilon_t\}} + \gamma_t} \middle| \mathcal{F}_{t-1}\right] = \sum_{i=1}^{N} p_{t,i} \frac{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i} \geq \varepsilon_t\}} \ell_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i} \geq \varepsilon_t\}} + \gamma_t} \\
&\geq \sum_{i=1}^{N} p_{t,i}\ell_{t,i} - \gamma_t \sum_{i=1}^{N} \frac{p_{t,i}}{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i} \geq \varepsilon_t\}} + \gamma_t} = \sum_{i=1}^{N} p_{t,i}\ell_{t,i} - \gamma_t Q'_t.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2 \middle| \mathcal{F}_{t-1}\right] &= \sum_{i=1}^{N} p_{t,i} \frac{\mathbb{E}\left[s_{I_t,i}^2 \mathbb{I}_{\{s_{I_t,i}\geq\varepsilon_t\}} \middle| \mathcal{F}_{t-1}\right] \ell_{t,i}^2 + \mathbb{E}\left[\left(1 - s_{I_t,i}\mathbb{I}_{\{s_{I_t,i}\geq\varepsilon_t\}}\right)^2 \middle| \mathcal{F}_{t-1}\right] \mathbb{E}\left[\xi_{t,i}^2 \middle| \mathcal{F}_{t-1}\right]}{\left(\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i}\geq\varepsilon_t\}} + \gamma_t\right)^2} \\
&\leq \sum_{i=1}^{N} p_{t,i} \frac{\sum_{j=1}^{N} p_{t,j} s_{j,i}^2 + R^2}{\left(\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i}\geq\varepsilon_t\}} + \gamma_t\right)^2} \leq \frac{1}{\varepsilon_t} \sum_{i=1}^{N} p_{t,i} \frac{1+R^2}{\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i}\geq\varepsilon_t\}} + \gamma_t} \\
&= \frac{(1+R^2)}{\varepsilon_t} Q_t',
\end{aligned}$$

where the last inequality uses that $\sum_{j=1}^{N} p_{t,j} s_{j,i} \mathbb{I}_{\{s_{j,i}\geq\varepsilon_t\}} + \gamma_t \geq \varepsilon_t$. Now, following the proof of Lemma 2, we can prove

$$Q_t' \leq 2\frac{\alpha(G_t(\varepsilon_t))}{\varepsilon_t}\left(1 + \log\left(1 + \frac{N^2/\gamma_t + N + 1}{\alpha}\right)\right).$$

For finishing the proof, let us set $\eta_t = \eta \geq 0$ and $\gamma_t = \gamma \geq 0$ for all t. Putting all of the above results together, we get

$$\begin{aligned}
R_T &\leq \frac{\log N}{\eta} + \gamma \sum_{t=1}^{T} Q_t' + \eta(1+R^2) \sum_{t=1}^{T} \frac{Q_t'}{\varepsilon_t} \\
&\leq \frac{\log N}{\eta} + \gamma C_1 \sum_{t=1}^{T} \frac{\alpha(G_t(\varepsilon_t))}{\varepsilon_t} + \eta C_2 (1+R^2) \sum_{t=1}^{T} \frac{\alpha(G_t(\varepsilon_t))}{\varepsilon_t^2},
\end{aligned}$$

where C_1 and C_2 are $\mathcal{O}\left(\log\left(N/\gamma\right)\right)$. Optimizing the choice of η and γ concludes the proof of Theorem 1. □