Online learning with Erdős–Rényi side-observation graphs

Tomáš Kocák
SequeL team
Inria Lille - Nord Europe

Gergely Neu
Universitat Pompeu Fabra
Barcelona, Spain

Michal Valko
SequeL team
Inria Lille - Nord Europe

Abstract

We consider adversarial multi-armed bandit problems where the learner is allowed to observe losses of a number of arms beside the arm that it actually chose. We study the case where all non-chosen arms reveal their loss with an *unknown* probability r_t, independently of each other and the action of the learner. Moreover, we allow r_t to change in every round t, which rules out the possibility of estimating r_t by a well-concentrated sample average. We propose an algorithm which operates under the assumption that r_t is large enough to warrant at least one side observation with high probability. We show that after T rounds in a bandit problem with N arms, the expected regret of our algorithm is of order $\mathcal{O}\left(\sqrt{\sum_{t=1}^{T}(1/r_t)\log N}\right)$, given that $r_t \geq \log T/(2N-2)$ for all t. All our bounds are within logarithmic factors of the best achievable performance of any algorithm that is even allowed to know exact values of r_t.

1 INTRODUCTION

In sequential learning, a learner is repeatedly asked to choose an action for which it obtains a loss and receives a feedback from the environment (Cesa-Bianchi and Lugosi, 2006). We typically study two feedback settings: the learner either observes the losses for all the potential actions (full information) or it observes only the loss of the action it chose. This latter feedback scheme is known as the *bandit* setting (cf. Auer et al., 2002a). In this paper, instead of considering these two limit cases, we study a more refined feedback model, known as *bandit with side observations* (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014, 2016), that generalizes both of them. Typical examples for learning with full information and bandit feedback are sequential trading on a stock market (where all stock prices are fully observable after each trading period), and electronic advertising (where the learner can only observe the clicks on actually shown ads), respectively. However, advertising in a social network offers a more intricate user feedback than captured by the basic bandit model: when proposing an item to a user in a social network, the advertiser can often learn about the preferences of the user's connections as well. Naturally, the advertiser would want to improve its recommendation strategy by incorporating these side observations.

Besides advertising and recommender systems, side observations can also arise in sensor networks, where the action of the learner amounts to probing a particular sensor. In this setting, each sensor can reveal readings of some other sensors that are in its range. When our goal is to sequentially select a sensor maximizing a property of interest, a good learning strategy should be able to leverage these side readings.

In this paper, we follow the formalism of Mannor and Shamir (2011) who model side observations with a graph structure over the actions: two actions mutually reveal their losses if they are connected by an edge in the graph in question. In a realistic scenario this graph is *time dependent* and *unknown* to the learner (e.g., the advertiser or the algorithm scheduling sensor readings). All previous algorithms for the studied setting (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014, 2016) require the environment to reveal a substantial part of a graph, at least after the side observations have been revealed. Specifically, these algorithms require the knowledge of the *second neighborhood* (the set of neighbors of the neighbors) of the chosen action in order to update their internal loss estimates. On the other hand, they are able to handle arbitrary graph structures, potentially chosen by an adversary and prove performance guarantees using graph properties based on cliques or independence sets.

The main contribution of our work is a learning algorithm that, unlike previous solutions, does *not require the knowledge of the exact graph* underlying the observations, beyond knowing from which nodes the side observations

came from. Relaxing this assumption, however, has to come with a price: As the very recent results of Cohen et al. (2016) show, achieving nontrivial advantages from side observations may be impossible without perfectly known side-observation graphs when an adversary is allowed to pick *both* the losses and the side-observation graphs. On the positive side, Cohen et al. offer efficient algorithms achieving strong improvements over the standard regret guarantees under the assumption that the losses are generated in an i.i.d. fashion and the graphs may be generated adversarially. Complementing these results, we consider the case of adversarial losses and make the assumption that the side-observation graph in round t is generated from an *Erdős–Rényi* model with an *unknown* and *time-dependent* parameter r_t. The main challenge for the learner is then the necessity to exploit the side observations despite not knowing the sequence (r_t). It is easy to see that this model can be equivalently understood as each non-chosen arm revealing its loss with probability r_t, independently of all other observations. That said, we still find it useful to think of the side observations as being generated from an Erdős–Rényi model, as it allows direct comparisons with the related literature. In particular, the case of learning with Erdős–Rényi side-observation graphs was considered before by Alon et al. (2013): Given *full access* to the underlying graph structure, their algorithm `Exp3-SET` can be shown to guarantee a regret bound of $\mathcal{O}\big(\sqrt{\sum_t (1/r_t)(1-(1-r_t)^N)\log N}\big)$. While the assumption of having full access to the graph be dropped relatively easily in this particular case, exact knowledge of r_t seems to be crucial for constructing reliable loss estimates and use them to guide the choice of action in each round.

It turns out that the problem of estimating r_t while striving to perform efficiently is in fact a major difficulty in our setting. Indeed, as we allow r_t to change arbitrarily between each round, we cannot rely on any past observations to construct well-concentrated estimates of these parameters. That is, the main challenge is estimating r_t from only a handful of samples. The core technical tool underlying our approach is a direct estimation procedure for the losses that does not estimate r_t explicitly.

Armed with this estimation procedure, we propose a learning algorithm called `Exp3-Res` that guarantees a regret of $\mathcal{O}(\sqrt{\sum_t (1/r_t)\log N})$, provided that $r_t \geq \log T/(2N-2)$ holds for all rounds t. This assumption essentially corresponds to requiring that, with high probability, at least 1 side observation is produced in every round, or, in other words, the side-observation graphs encountered are all *non-empty*. Notice that for the assumed range of r_t's, our regret bound improves upon the standard regret bound of `Exp3`, which is of $\mathcal{O}(\sqrt{NT\log N})$. It is easy to see that when r_t becomes smaller than $1/N$, side observations become unreliable and the bound of `Exp3` cannot be improved. That is, if our assumption cannot be verified a priori, then ignoring all side observations and using the `Exp3` algorithm of Auer et al. (2002a) instead can yield a better performance. On the other hand, given that our assumption holds, our bounds cannot be significantly improved as suggested by the lower-bound of $\Omega(\sqrt{T/r})$ proved for a static r by Alon et al. 2013.

Many other partial-information settings have been studied in previous work. One of the simplest of these settings is the label-efficient prediction game considered by Cesa-Bianchi et al. (2005), where the learner can observe either losses of all the actions or none of them, not even the loss of the chosen action. This observation can be queried by the learner at most an $\varepsilon < 1$ fraction of the total number of rounds, which means no losses are observed in the remaining rounds. An even more restricted information setting, label efficient bandit feedback was considered by Allenberg et al. (2006), where the learner can only query the loss of the chosen action, instead of all losses (see also Audibert and Bubeck, 2010). Algorithms for these two settings have regret of $\widetilde{\mathcal{O}}(\sqrt{T/\varepsilon})$ and $\widetilde{\mathcal{O}}(\sqrt{NT/\varepsilon})$, respectively. While these bounds may appear very similar to ours, notice that our setting offers a more intricate (and, for some problems, more realistic) feedback scheme, which also turns out to be much more challenging to exploit. In another related setting, Seldin et al. (2014) consider M side observations that the learner can proactively choose in each round without limitations. Seldin et al. deliver an algorithm with regret of $\widetilde{\mathcal{O}}(\sqrt{(N/M)T})$, also proving that choosing M observations uniformly at random is minimax optimal; given this sampling scheme, it is not even necessary to observe the loss of the chosen action. Their result is comparable to ours and the result by Alon et al. (2013) for Erdős–Rényi observation graphs with parameter $r = M/N$. However, Seldin et al. also assume that M is known, which obviates the need for estimating r. We provide a more technical discussion on the related work in Section 6.

In our paper, we assume that, just like the observation probabilities, the losses are *adversarial*, that is, they can change at each time step without restrictions. Learning with side observations and stochastic losses was studied by Caron et al. (2012) and Buccapatnam et al. (2014). While this is an easier setting that the adversarial one, the authors assumed, in both cases, that the graphs have to be known in advance. Recently, Carpentier and Valko (2016) studied another stochastic setting where the graph is also not known in advance, however their setting considers different feedback and loss structure (influence maximization) which differs from the side-observation setting.

Furthermore, Alon et al. (2015) considered a strictly more difficult setting than ours, where the loss of the chosen action may not be a part of the received feedback.

2 PROBLEM DEFINITION

We now formalize our learning problem. We consider a sequential interaction scheme between a learner and an environment, where the following steps are repeated in every round $t = 1, 2, \ldots, T$:

1. The environment chooses $r_t \in [0, 1]$ and a loss function over the arms, with $\ell_{t,i}$ being the loss associated with arm $i \in [N] \stackrel{\text{def}}{=} \{1, 2, \ldots, N\}$ at time t.
2. Based on its previous observations (and possibly some randomness), the learner draws an arm $I_t \in [N]$.
3. The learner suffers loss ℓ_{t,I_t}.
4. For all $i \neq I_t$, $O_{t,i}$ is independently drawn from a Bernoulli distribution with mean r_t. Furthermore, O_{t,I_t} is set as 1.
5. For all $i \in [N]$ such that $O_{t,i} = 1$, the learner observes the loss $\ell_{t,i}$.

The goal of the learner is to minimize its total expected losses, or, equivalently, to minimize the *total expected regret* (or, in short, regret) defined as

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

We will denote the interaction history between the learner and the environment up to the beginning of round t by $\mathcal{F}_{t-1}$. We also define $p_{t,i} = \mathbb{P}\left[I_t = i \middle| \mathcal{F}_{t-1}\right]$.

The main challenge in our setting is leveraging side observations *without knowing* r_t. Had we had access to the exact value of r_t, we would be able to define the following estimate of $\ell_{t,i}$:

$$\widehat{\ell}^{\star}_{t,i} = \frac{O_{t,i}\ell_{t,i}}{p_{t,i} + (1 - p_{t,i})r_t} \tag{1}$$

It is easy to see that the loss estimates defined this way are unbiased in the sense that $\mathbb{E}\left[\widehat{\ell}_{t,i} \middle| \mathcal{F}_{t-1}\right] = \ell_{t,i}$ for all t and i. It is also straightforward to show that an appropriately tuned instance of the Exp3 algorithm of Auer et al. (2002a) fed with these loss estimates is guaranteed to achieve a regret of $\mathcal{O}(\sqrt{\sum_t (1/r_t) \log N})$ (see also Seldin et al. 2014).

Then, one might consider a simple algorithm that devotes a number of observations to obtain an estimate $\widehat{r}_t$ of r_t and plug this estimate into (1). However, notice that since r_t is allowed to change arbitrarily over time, we can only work with a severely limited sample budget for estimating r_t: only $N - 1$ independent observations! Thus, we can obtain only very loose confidence intervals around r_t which translate to even more useless confidence intervals around $\widehat{\ell}^{\star}_{t,i}$.

Below, we describe a simple trick for obtaining loss estimates that have similar properties to the ones defined in (1) without requiring exact knowledge or even explicit estimation of r_t. Our procedure is based on the geometric resampling method of Neu and Bartók (2013). To get an intuition of the method, let us assume that we have access to the independent geometrically distributed random variable $G^{\star}_{t,i}$ with parameter $o_{t,i} = p_{t,i} + (1 - p_{t,i})r_t$. Then, replacing $1/o_{t,i}$ by $G^{\star}_{t,i}$ in the definition of $\widehat{\ell}^{\star}_t$ and ensuring that $G^{\star}_{t,i}$ is independent of $O_{t,i}$, we can obtain an unbiased loss estimate essentially equivalent to $\widehat{\ell}^{\star}_t$.

The challenge posed by this approach is that in our setting, we do not have exact sample access to the geometric random variable $G^{\star}_{t,i}$. In the next section, we describe our algorithm that is based on replacing $G^{\star}_{t,i}$ in the above definition by an appropriate surrogate.

3 ALGORITHM

Our algorithm is called `Exp3-Res` and displayed as Algorithm 1. It is based on the `Exp3` algorithm of Auer et al. (2002a) and crucially relies on the construction of a surrogate $G_{t,i}$ of $G^{\star}_{t,i}$. Throughout this section, we will assume that $r_t \geq \frac{\log T}{2N-2}$, which implies that the probability of having no side observations in round t is of order $1/\sqrt{T}$.

The algorithm is initialized by setting $w_{1,i} = 1/N$ for all $i \in [N]$, and then performing the updates

$$w_{t+1,i} = \frac{1}{N} \exp\left(-\eta_{t+1}\widehat{L}_{t,i}\right) \tag{2}$$

after each round t, where $\eta_{t+1} > 0$ is a parameter of the algorithm called the *learning rate* in round t and $\widehat{L}_{t,i}$ is cumulative sum of the loss estimates $\widehat{\ell}_{s,i}$ up to (and including) time t. In round t, the learner draws its action I_t such that $I_t = i$ holds with probability $p_{t,i} \propto w_{t,i}$. To simplify some of the notation below, we introduce the shorthand notations $\mathbb{P}_t\left[\cdot\right] = \mathbb{P}\left[\cdot \middle| \mathcal{F}_{t-1}\right]$ and $\mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot \middle| \mathcal{F}_{t-1}\right]$.

For any fixed t, i, we now describe an efficiently computable surrogate $G_{t,i}$ for the geometrically distributed random variable $G^{\star}_{t,i}$ with parameter $o_{t,i}$ that will be used for constructing our loss estimates. In particular, our strategy will be to construct several independent copies $\left\{O'_{t,i}(k)\right\}$ of $O_{t,i}$ and choosing $G_{t,i}$ as the index k of the first copy with $O'_{t,i}(k) = 1$. It is easy to see that with infinitely many copies, we could exactly recover $G^{\star}_{t,i}$; our actual surrogate is going to be weaker thanks to the smaller sample size. For clarity of notation, we will omit most explicit references to t and i, with the understanding that all calculations need to be independently executed for all pairs t, i.

Let us now describe our mechanism for constructing the copies $\{O'(k)\}$. Since we need independence of $G_{t,i}$ and $O_{t,i}$ for our estimates, we use only side observations from

actions $[N] \setminus \{I_t, i\}$. First, let's define σ as a uniform random permutation of $[N] \setminus \{I_t, i\}$. For all $k \in [N-2]$, we define $R(k) = O_{t,\sigma(k)}$. Note that due to the construction, $\{R(k)\}_{k=1}^{N-2}$ are pairwise independent Bernoulli random variables with parameter r_t, independent of $O_{t,i}$. Furthermore, knowing $p_{t,i}$ we can define $P(1), \ldots, P(N-2)$ as pairwise independent Bernoulli random variables with parameter $p_{t,i}$. Using $P(k)$ and $R(k)$ we define the random variable $O'(k)$ as

$$O'(k) = P(k) + (1 - P(k))R(k)$$

for all $k \in [N-2]$. Using independence of all previously defined random variables, it is easy to check that the variables $\{O'(k)\}_{k=1}^{N-2}$ are pairwise independent Bernoulli random variables with expectation $o_{t,i} = p_{t,i} + (1-p_{t,i})r_t$. Now we are ready to define $G_{t,i}$ as

$$G_{t,i} = \min\left\{k \in [N-2] : O(k)' = 1\right\} \cup \{N-1\}. \quad (3)$$

The following lemma states some properties of $G_{t,i}$.

Lemma 1. *For any value of g we have*

$$\mathbb{E}\left[G_{t,i}\right] = \frac{1}{o_{t,i}} - \frac{1}{o_{t,i}}(1-o_{t,i})^{N-1}$$

$$\mathbb{E}\left[G_{t,i}^2\right] = \frac{2-o_{t,i}}{o_{t,i}^2} + \frac{1}{o_{t,i}^2}(1-o_{t,i})^{N-2} \times \\ \times \left(o_{t,i}^2 + o_{t,i} - 2 + 2o_{t,i}(N-2)(o_{t,i}-1)\right)$$

Proof. The proof follows directly from using the definition of $G_{t,i}$ and simplifying the sums

$$\mathbb{E}\left[G_{t,i}\right] = \sum_{k=1}^{N-2}\left[k o_{t,i}(1-o_{t,i})^{k-1}\right] + \\ + (N-1)\left(1-o_{t,i}\right)^{N-2},$$

$$\mathbb{E}\left[G_{t,i}^2\right] = \sum_{k=1}^{N-2}\left[k^2 o_{t,i}(1-o_{t,i})^{k-1}\right] + \\ + (N-1)^2\left(1-o_{t,i}\right)^{N-2}.$$

□

Using Lemma 1, it is easy to see that $G_{t,i}$ follows a truncated geometric law in the sense that

$$\mathbb{P}\left[G_{t,i} = m\right] = \mathbb{P}\left[\min\left\{G_{t,i}^\star, N-1\right\} = m\right]$$

holds for all $m \in [N-1]$. Using all this notation, we construct an estimate of $\ell_{t,i}$ as

$$\widehat{\ell}_{t,i} = G_{t,i} O_{t,i} \ell_{t,i}. \quad (4)$$

The rationale underlying this definition of $G_{t,i}$ is rather delicate. First, note that $p_{t,i}$ is deterministic given the history $\mathcal{F}_{t-1}$ and therefore, does not depend on $O_{t,i}$. Second,

Algorithm 1 `Exp3-Res`

Input:
Set of actions $[N]$.
Initialization:
$\widehat{L}_{0,i} \leftarrow 0$ for $i \in [N]$.
Run:
for $t = 1$ **to** T **do**
  $\eta_t \leftarrow \sqrt{\log N \Big/ \left(N^2 + \sum_{s=1}^{t-1}\sum_{i=1}^{N} p_{s,i}(\widehat{\ell}_{s,i})^2\right)}$.
  $w_{t,i} \leftarrow (1/N)\exp(-\eta_t \widehat{L}_{t-1,i})$ for $i \in [N]$.
  $W_t \leftarrow \sum_{i=1}^{N} w_{t,i}$.
  $p_{t,i} \leftarrow w_{t,i}/W_t$.
  Choose $I_t \sim p_t = (p_{t,1}, \ldots, p_{t,N})$.
  Receive the observation set O_t.
  Receive the pairs $\{i, \ell_{t,i}\}$ for all i s.t. $O_{t,i} = 1$.
  Compute $G_{t,i}$ for all $i \in [N]$ using (3).
  $\widehat{\ell}_{t,i} \leftarrow \ell_{t,i} O_{t,i} G_{t,i}$ for all $i \in [N]$.
  $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \widehat{\ell}_{t,i}$ for all $i \in [N]$.
end for

$O_{t,i}$ is also independent of $O_{t,j}$ for $j \notin \{i, I_t\}$. As a result, $G_{t,i}$ is independent of $O_{t,i}$, and we can use the identity $\mathbb{E}_t\left[G_{t,i}O_{t,i}\right] = \mathbb{E}_t\left[G_{t,i}\right]\mathbb{E}_t\left[O_{t,i}\right]$. The next lemma relates the loss estimates (4) to the true losses, relying on the observations above and the assumption $r_t \geq \frac{\log T}{2N-2}$.

Lemma 2. *Assume $r_t \geq \frac{\log T}{2N-2}$. Then, for all t and i,*

$$0 \leq \ell_{t,i} - \mathbb{E}_t\left[\widehat{\ell}_{t,i}\right] \leq \frac{1}{\sqrt{T}}.$$

Proof. Fix an arbitrary t and i. Using Lemma 1 along with $\mathbb{E}_t\left[O_{t,i}\right] = o_{t,i}$ and the independence of $G_{t,i}$ and $O_{t,i}$, we get

$$\mathbb{E}_t\left[\widehat{\ell}_{t,i}\right] = \mathbb{E}_t\left[G_{t,i}O_{t,i}\ell_{t,i}\right] = \ell_{t,i} - \ell_{t,i}(1-o_{t,i})^{N-1},$$

which immediately implies the lower bound on $\ell_{t,i} - \mathbb{E}_t\left[\widehat{\ell}_{t,i}\right]$. For proving the upper bound, observe that

$$\ell_{t,i}(1-o_{t,i})^{N-1} \leq (1-r_t)^{N-1} \leq e^{-r_t(N-1)} \leq \frac{1}{\sqrt{T}}$$

holds by our assumption on r_t, where we used the elementary inequality $1 - x \leq e^x$ that holds for all $x \in \mathbb{R}$. □

The next theorem states our main result concerning `Exp3-Res` with an adaptive learning rate.

Theorem 1. *Assume that $r_t \geq \frac{\log T}{2N-2}$ holds for all t and set*

$$\eta_t = \sqrt{\frac{\log N}{N^2 + \sum_{s=1}^{t-1}\sum_{i=1}^{N} p_{s,i}(\widehat{\ell}_{s,i})^2}}.$$

Then, the expected regret of `Exp3-Res` *satisfies*

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

4 PROOF OF THEOREM 1

In this section, we present details of the proof of Theorem 1 but first, we state an auxiliary lemma.

Lemma 3 (Lemma 3.5 of Auer et al., 2002b)**.** *Let* b_1, b_2, $\ldots$, b_T *be non-negative real numbers. Then*

$$\sum_{t=1}^{T} \frac{b_t}{\sqrt{\sum_{s=1}^{t} b_t}} \leq 2\sqrt{\sum_{t=1}^{T} b_t}.$$

Proof. The proof is based on the inequality $x/2 \leq 1 - \sqrt{1-x}$ for $x \leq 1$. Setting $x = b_t / \sum_{s=1}^{t} b_s$ and multiplying both sides of the inequality by $\sqrt{\sum_{s=1}^{t} b_s}$ we get

$$\frac{b_t}{\sqrt{\sum_{s=1}^{t} b_t}} \leq \sqrt{\sum_{s=1}^{t} b_s} - \sqrt{\sum_{s=1}^{t} b_s - b_t}.$$

The proof is concluded by summing over t. □

The first part of the analysis follows the proof of Lemma 1 by Györfi and Ottucsák (2007). Defining $W'_t = \frac{1}{N}\sum_{i=1}^{N} e^{-\eta_{t-1}\widehat{L}_{t-1,i}}$, we get

$$\begin{aligned}
\frac{1}{\eta_t}\log\frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t}\log\sum_{i=1}^{N}\frac{\frac{1}{N}e^{-\eta_t\widehat{L}_{t-1,i}}e^{-\eta_t\widehat{\ell}_{t,i}}}{W_t} \\
&= \frac{1}{\eta_t}\log\sum_{i=1}^{N} p_{t,i}e^{-\eta_t\widehat{\ell}_{t,i}} \\
&\leq \frac{1}{\eta_t}\log\sum_{i=1}^{N} p_{t,i}\left(1 - \eta_t\widehat{\ell}_{t,i} + (\eta_t\widehat{\ell}_{t,i})^2\right) \qquad (5) \\
&= \frac{1}{\eta_t}\log\left(1 - \eta_t\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i} + \eta_t^2\sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2\right),
\end{aligned}$$

where in (5), we used the inequality $\exp(-x) \leq 1 - x + x^2$ that holds for $x \geq -1$. Further, we used the inequality $\log(1-x) \leq -x$, which holds for all $x \leq 1$, to upper bound last term.

Using $\eta_{t+1} \leq \eta_t$ and Jensen's inequality, we get

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

which, together with the last inequality, gives us

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

for every $t \in [T]$. Taking expectations and summing over time, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t}{2}\sum_{i=1}^{N} p_{t,i}\left(\widehat{\ell}_{t,i}\right)^2\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right)\right].
\end{aligned}$$

The goal of the second part of the analysis is to construct bounds for each of the three expectations in the previous inequality. For the term on the left-hand side, we use Lemma 2 to get the lower-bound

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\widehat{\ell}_{t,i}\right] \geq \sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\ell_{t,i} + \sqrt{T}.$$

Note that is the only step in the analysis where the actual magnitude (and not just the sign) of the bias of the loss estimates shows up. Anything bigger than $\sqrt{T}$ would degrade our final regret bound.

We are left with bounding the two terms on the right-hand side. To simplify some notation below, let us define $b_t = \sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2$. By our definition of η_t and the help of Lemma 3, we can bound the first term on the right hand side as

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t b_t}{2}\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{b_t\sqrt{\log N}}{2\sqrt{N^2 + \sum_{s=1}^{t-1} b_s}}\right] \\
&\leq \mathbb{E}\left[\sqrt{\left(N^2 + \sum_{t=1}^{T} b_t\right)\log N}\right] \\
&\leq \sqrt{\left(N^2 + \sum_{t=1}^{T}\mathbb{E}\left[b_t\right]\right)\log N},
\end{aligned}$$

where we also used the fact that $N^2 \geq b_t$ and Jensen's inequality in the last line. We continue by bounding $\mathbb{E}[b_t]$:

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i=1}^{N} p_{t,i}(\widehat{\ell}_{t,i})^2\right] &= \sum_{i=1}^{N} p_{t,i}\ell_{t,i}^2\mathbb{E}_t\left[O_{t,i}G_{t,i}^2\right] \\
&\leq \sum_{i=1}^{N} p_{t,i}o_{t,i}\frac{2 - o_{t,i}}{o_{t,i}^2} \leq \frac{2}{r_t}, \qquad (6)
\end{aligned}$$

where we used $o_{t,i} \geq r_t$ together with the second part of Lemma 1 which gives us

$$\mathbb{E}_t\left[G_{t,i}^2\right] = \frac{2-o_{t,i}}{o_{t,i}^2} + \frac{1}{o_{t,i}^2}(1-o_{t,i})^{N-2}\times$$
$$\times\left(o_{t,i}^2 + o_{t,i} - 2 + 2o_{t,i}(N-2)(o-1)\right)$$
$$\leq \frac{2-o_{t,i}}{o_{t,i}^2},$$

since both $o_{t,i}^2 + o_{t,i} - 2$ and $2o_{t,i}(N-2)(o-1)$ are non-positive. Thus, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta_t b_t}{2}\right] \leq \sqrt{\left(\sum_{t=1}^{T}\frac{1}{r_t} + N^2\right)\log N}. \qquad (7)$$

Finally, using $W_1 = 1$, the sum in the last expectation telescopes to

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right)\right] = \mathbb{E}\left[-\frac{\log W_{T+1}}{\eta_{T+1}}\right].$$

Using the definition of W_t, we get that

$$\mathbb{E}\left[-\frac{\log W_{T+1}}{\eta_{T+1}}\right] \leq \mathbb{E}\left[-\frac{\log w_{T+1,j}}{\eta_{T+1}}\right]$$
$$\leq \mathbb{E}\left[\frac{\log N}{\eta_{T+1}}\right] + \mathbb{E}\left[\widehat{L}_{T,j}\right]$$

holds for any arm $j \in [N]$. Now note that the first term can be bounded by using the defintion of η_{T+1} with the help of (6) and Jensen's inequality. Using $\mathbb{E}_t\left[\widehat{\ell}_{t,i}\right] \leq \ell_{t,i}$ from Lemma 2 and combining everything together, we obtain the regret bound

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} p_{t,i}\ell_{t,i}\right] - \min_{j\in[N]}\mathbb{E}\left[\sum_{t\in T_k}\ell_{t,j}\right]$$
$$\leq 2\sqrt{\left(N^2 + \sum_{t=1}^{T}\frac{1}{r_t}\right)\log N} + \sqrt{T}.$$

5 EXPERIMENTS

In this section, we study the empirical performance of Exp3-Res compared to three other algorithms:

- Exp3 – a basic adversarial multi-armed bandit algorithm which uses only loss observations of chosen arms and discards all side observations.
- Oracle – full-information algorithm with access to losses of every action in every time step, regardless of the value of r_t. Our particular choice is Hedge (Littlestone and Warmuth, 1994; Freund and Schapire, 1997).
- Exp3-R – a variant of the Exp3-Res algorithm with access to the sequence $(r_t)_t^T$, using (1) to construct unbiased loss estimate instead of using geometric resampling.

The most interesting parameter of our experiment is the sequence (r_t), since it controls amount of side observation presented to the learner. In order to show that Exp3-Res can effectively make use of the additional information provided by the environment, we designed several sequences (r_t) with different amounts of side observation provided to the learner. In the case of small r_t-s, the problem is almost as difficult as the multi-armed bandit problem. On the other hand, in the case of large r_t-s, the problem is almost as easy as the full-information problem. Therefore, we expect that the performance of Exp3-Res will interpolate between the performance of the Exp3-R and Oracle algorithms depending on the values of the r_t-s. In the next section, we validate this claim empirically.

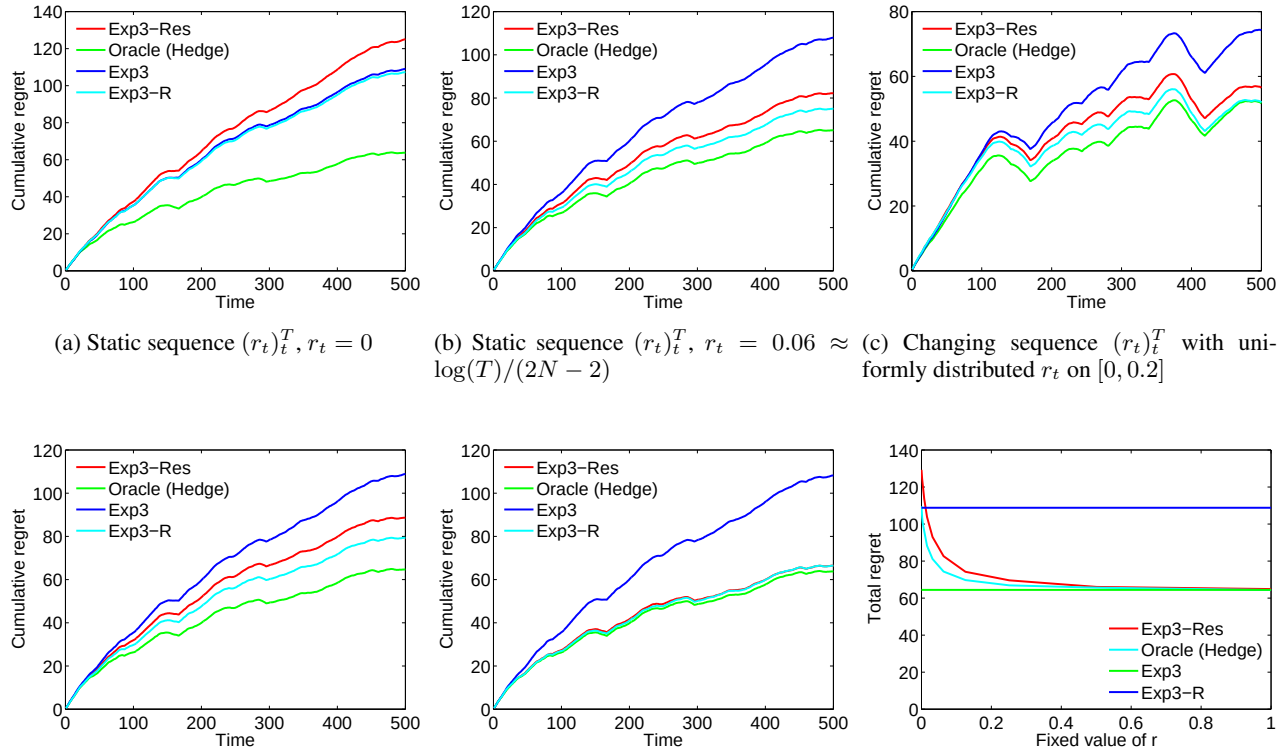
5.1 EXPERIMENT DETAILS

To ensure sufficient challenge for the algorithms, we have generated a sequence of losses as a random walk for each arm with independent increments uniformly distributed on $[-0.1, 0.1]$ while enforcing the random walks to stay within $[0, 1]$ by setting the value of a random walk to 0 or 1, respectively, if the random walk gets outside the boundaries. The loss sequence is fixed through all of the experiments to demonstrate the impact of the sequence $(r_t)_t^T$ on the regret of algorithms. We have observed qualitatively similar behavior for other loss sequences.

We fix the number of arms in all of the experiments as 50, and the time horizon as 500. Every curve represents an average of 100 runs.

5.2 RESULT OF THE EXPERIMENTS

We performed experiments on many different loss sequences and sequences of r_t-s. Since the results are essentially the same for all the different sequences, we included in the present paper just the results for one loss sequence with different sequences of r_t-s. In the case of $r_t \geq \log(T)/(2N-2)$, the case of high probability of having some side observation, the performance of the algorithm Exp3-Res proposed in the present paper is comparable to the performance of the idealistic Exp3-R which knows exact value of r_t in every time step. Moreover, if the average r_t is close to 1, the performance of the proposed algorithm is close to the performance of Oracle which observes all the losses. If the average r_t is close to zero, the performance of the algorithm is a little bit worse that the performance of basic Exp3. This is also supported by the theory, since our algorithm is not able to construct reliable estimates in the case of small r_t-s.



(a) Static sequence $(r_t)_t^T$, $r_t = 0$

(b) Static sequence $(r_t)_t^T$, $r_t = 0.06 \approx \log(T)/(2N-2)$

(c) Changing sequence $(r_t)_t^T$ with uniformly distributed r_t on $[0, 0.2]$

(d) Sequence $(r_t)_t^T$ generated as a random walk on $[0, 0.1]$

(e) Sequence $(r_t)_t^T$ generated as a random walk on $[0, 1]$

(f) Total regret for different values of static $(r_t)_t^T$

Figure 1: Comparison of algorithm for different amount of side information sequences (different sequences $(r_t)_t^T$)

6 CONCLUSION & FUTURE WORK

In this paper, we considered multi-armed bandit problems with stochastic side observations modeled by Erdős–Rényi graphs. Our contribution is a computationally efficient algorithm that operates under the assumption $r_t \geq \log T/(2N-2)$, which essentially guarantees that at least one piece of side observation is generated in every round, with high probability. In this case, our algorithm guarantees a regret bound of $\mathcal{O}\left(\sqrt{\log N \sum_{t=1}^{T} \frac{1}{r_t}}\right)$ (Theorem 1). In this section, we discuss several open questions regarding this result.

The most obvious question is whether it is possible to remove our assumptions on the values of r_t. We can only give a definite answer in the simple case when all r_t's are identical: In this case, one can think of simply computing the empirical frequency $\widehat{r}_t$ of all previous side observations in round t to estimate the constant r, plug the result into (1), and then use the resulting loss estimates in an exponential-weighting scheme. It is relatively straightforward (but also rather tedious) to show that the resulting algorithm satisfies a regret bound of $\widetilde{\mathcal{O}}\left(\sqrt{T/r}\right)$ for all possible values of r, thanks to the fact that $\hat{r}_t$ quickly concentrates around the true value of r. Notice however that this approach clearly breaks down if the r_t's change over time.

In the case of changing r_t's, the number of observations we can use to estimate r_t is severely limited, so much that we cannot expect any direct estimate of r_t to concentrate around the true value. Our algorithm proposed in Section 3 gets around this problem by directly estimating the importance weights $1/o_{t,i}$ instead of r_t, which enables us to construct reliable loss estimates, although only at the price of our assumption on the range of r_t. While we acknowledge that this assumption can be difficult to confirm a priori in practice, we remark that we find it quite surprising that *any algorithm whatsoever* can take advantage of such limited observations, even under such a restriction. We also point out that for values of r_t that are consistently below our bound, it is not possible to substantially improve the regret bounds of `Exp3` which are of $\widetilde{\mathcal{O}}\left(\sqrt{TN}\right)$, as shown by the lower bounds of Alon et al. (2013). We expect that in several practical applications, one can verify whether the r_t's satisfy our assumption or not, and decide to use `Exp3-Res` or `Exp3` accordingly. In fact, our experiments suggest that our algorithm performs well even if neither of these two assumptions are verified: we have seen that the empirical performance of `Exp3-Res` is only slightly worse than that

of Exp3 even when the values of r_t are very small (Section 5). Still, finding out whether our restriction on r_t can be relaxed in general is a very important and interesting question left for future study.

An important corollary of our results is that, under some assumptions, it is possible to leverage side observations in a non-trivial way without having access to the second neighborhoods in the side-observation graphs as defined by Mannor and Shamir (2011). This complements the recent results of Cohen et al. (2016), who show that non-stochastic side-observations may provide non-trivial advantage over bandit feedback when the losses are stochastic even when the side-observation graphs are unobserved, but learning with unobserved feedback graphs can be as hard as learning with bandit feedback when both the losses and the graphs are generated by an adversary. A natural question that our work leads to is whether it is possible to efficiently leverage side-observations under significantly weaker assumptions on the observation model.

Acknowledgements The research presented in this paper was supported by CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, French National Research Agency project ExTra-Learn (n.ANR-14-CE24-0010-01), and by UPFellows Fellowship (Marie Curie COFUND program n° 600387).

References

Allenberg, C., Auer, P., Györfi, L., and Ottucsák, Gy. (2006). Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Algorithmic Learning Theory*, pages 229–243.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *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*.

Audibert, J.-Y. and Bubeck, S. (2010). Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11:2785–2836.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002a). The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auer, P., Cesa-Bianchi, N., and Gentile, C. (2002b). Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75.

Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. (2014). Stochastic bandits with side observations on networks. In *International Conference on Measurement and Modeling of Computer Systems*.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. (2012). Leveraging side observations in stochastic bandits. In *Uncertainty in Artificial Intelligence*.

Carpentier, A. and Valko, M. (2016). Revealing graph bandits for maximizing local influence. In *International Conference on Artificial Intelligence and Statistics*, pages 10–18.

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

Cesa-Bianchi, N., Lugosi, G., and Stoltz, G. (2005). Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162.

Cohen, A., Hazan, T., and Koren, T. (2016). Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning (to appear)*.

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

Györfi, L. and Ottucsák, Gy. (2007). Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory*, 53(5):1866–1872.

Kocák, T., Neu, G., and Valko, M. (2016). Online learning with noisy side observations. In *International Conference on Artificial Intelligence and Statistics*, pages 1186–1194.

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

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

Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations. In *Neural Information Processing Systems*.

Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In *Algorithmic Learning Theory*.

Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori, Y. (2014). Prediction with limited advice and multi-armed bandits with paid observations. In *International Conference on Machine Learning*.