Well-informed bandits: Learning with stochastic side-observations

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

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 a fixed but *unknown* probability r, independently of each other and the action of the learner. We propose two algorithms that work for different ranges of r. We show that after T rounds in a bandit problem with N arms, the expected regret of our first algorithm is $O(\sqrt{(T/r)\log N})$ whenever $r \geq (\log T)/(2N)$, while our second algorithm achieves a regret of $O(\sqrt{(T/r)\log(N+T)})$ for smaller values of r. We also give a quick estimation procedure that decides the range of r. All our bounds are within logarithmic factors of the best achievable performance of any algorithm that is even allowed to know r.

1 Introduction

In sequential learning settings, 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), that generalizes both of them. Typical examples for learning with full information and bandit feedback are sequential trading on a stock market, and electronic advertising, 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 the previous algorithms for the studied setting (Mannor and Shamir, 2011; Alon et al., 2013; Kocák et al., 2014) require the environment to reveal a substantial part of a graph, at least after an action has been chosen. 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 expressed using graph properties based on cliques or independence sets.

The most compelling contribution of our work is a generic algorithm that does *not require the knowledge of the graph* structure beyond knowing from which nodes the side observations came from. In this paper, we focus on side observations that are *stochastic*. In particular, we assume that each non-chosen arm reveals its loss with an *unknown* probability r, inde-

Preliminary work. Under review by AISTATS 2015. Do not distribute.

pendently of all other observations. In other words, no matter what action we choose, we receive the loss of all other actions, each with probability r. We also receive the loss of the chosen action with certainty. Such a correspondence between the action and the feedback can be modeled exactly by Erdős-Rényi random graphs with parameter r. Indeed, we can equivalently imagine that each time step a random graph on N vertices is sampled but not revealed to the learner, such that each edge has probability r of appearing. After the learner chooses an action, the losses of that action and all the neighbors of that graph from time t are revealed. The Exp3-SET algorithm by Alon et al. (2013) is able to attain the regret of $O(\sqrt{(T/r)\log(N)(1-(1-r)^N})$ for this setting. However, a major drawback of that method is that it *requires the knowledge of both r and the underlying graph*. While the latter assumption can be dropped relatively easily, exact knowledge of r seems to be crucial for constructing reliable loss estimates and use them to guide the choice of action in each round.

Clearly, the problem of estimating r while striving to perform efficiently is in fact very challenging, and it is the main issue we tackle in the present paper. Indeed, we need to estimate r and define reliable loss estimates based on a single stream of interactions. The core technical tool underlying our approach is a an r-dependent estimation procedure for loss estimation that does not estimate r explicitly. The most challenging setting for this procedure is when r becomes so small that one often encounters an empty graph of side observations. To deal with this problem, we group several rounds into *episodes* that aggregate samples for improving the accuracy of the procedure. We therefore present two algorithms called DuplExp3 for large values of r and DuplExp3($\underline{r}$) that also works for the smaller ones. We show that both algorithms achieve a regret of $\widetilde{O}(\sqrt{(T/r)}$. An obvious question to answer is, in which of the two settings do we find ourselves in. We answer this question by an efficient test procedure that also guarantees that the regret accumulated by the learner never exceeds the regret bound of the standard Exp3 algorithm. Summing up, we deliver a complete solution for this setting, no matter the value of r. This claim is supported by the fact that the lower bound for the setting is of the order $\sqrt{(T/r)}$ (Alon et al., 2013).

In a related setting, Seldin et al. (2014) consider M side observations that the learner can proactively choose in each round without limitations. They deliver an algorithm with regret of $\widetilde{O}(\sqrt{(N/M)T})$, also proving that choosing M observations uniformly at random is minimax optimal. In that sense, their result is for $r = M/N$ comparable to the known results for Erdős-Rényi action structures. However, they also assume that M is known, which makes their work distinct from ours.

In our paper, we assume that the losses are *adversarial*, that is, they can change at each time step without restrictions. The easier problem of learning with side observations and stochastic losses was studied by Caron et al. (2012) and Buccapatnam et al. (2014).

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 a loss function over the arms, with $\ell_{t,i}$ being the loss associated with arm $i \in [N] \overset{\text{def}}{=} \{1, 2, \ldots, N\}$.
2. Based on its previous observations (and possibly some external 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. 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 $\mathcal{F}_{t-1} = \sigma(I_{t-1}, \ldots, I_1)$. 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 prior knowledge of* r. Indeed, had we had access to the exact value of r, we would be able to define the following estimate of $\ell_{t,i}$:

$$\hat{\ell}^*_{t,i} = \frac{O_{t,i}\ell_{t,i}}{p_{t,i} + (1 - p_{t,i})r}. \tag{1}$$

It is easy to see that the loss estimates defined this way are unbiased in the sense that $\mathbb{E}\left[\hat{\ell}_{t,i} \middle| \mathcal{F}_{t-1}\right] = \ell_{t,i}$ for all t and i. It is 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 $O(\sqrt{(T/r)\log N})$ (see also Seldin et al. 2014). Then, one might consider a simple algorithm that first devotes a number of rounds to obtain an estimate $\hat{r}$ of r and plug these estimates into Equation 1. However, this approach is hindered by the fact that standard confidence intervals around r do not translate into reasonable confidence intervals around $\hat{\ell}^*_{t,i}$. See Section 7 for a more detailed discussion.

Below, we describe a simple trick for obtaining loss estimates that have similar properties to the ones defined in Equation (1) without requiring exact knowledge or even explicit estimation of r. 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 two independent geometrically distributed random variables M^*_t and $K_{t,i}$ with respective parameters r and $p_{t,i}$. Then, it is easy to see that $G^*_{t,i} = \min\{K_{t,i}, M^*_t\}$ is also geometrically distributed with parameter $o_{t,i} = p_{t,i} + (1 - p_{t,i})r$. Replacing $1/o_{t,i}$ by $G^*_{t,i}$ in the definition of $\hat{\ell}^*_t$ and ensuring that $G^*_{t,i}$ is independent of $O_{t,i}$, it is possible to prove a bound $O(\sqrt{(T/r)\log N})$ on the regret by using ideas from Neu and Bartók (2013). The challenge of this approach is that in our setting, we do not have exact sample access to the geometric random variable M^*_t. In the next sections, we describe our algorithms that are based on replacing M^*_t in the above definition by appropriate surrogates.

3 An algorithm for large values of r

We first consider the case when $r \geq \frac{\log T}{2N}$, which implies that the probability of having no side observations in round t is roughly bounded by $1/\sqrt{T}$. Our algorithm for this setting, called DuplExp3, is based on the Exp3 algorithm of Auer et al. (2002a). As the name suggests, DuplExp3 is a combination of two carefully interwoven Exp3 sub-algorithms, one overseeing the rounds when t is even and the other one the rest. This structure is used to make sure that the surrogate of G^*_t and O_t are independent, so all expectations concerning $\hat{\ell}_t$ can be controlled. The algorithm is initialized by setting $w_{1,i} = w_{2,i} = 1/N$ for all $i \in [N]$, and then performing the updates

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

after each round t, where $\widehat{L}_{t,i}$ will be defined shortly. In round t, the learner draws its action I_t such that $I_t = i$ holds with probability $p_{t,i} \propto w_{t,i}$. Note that, for technical reasons discussed below, the distribution of I_t is fully specified at the end of round $t-2$ (and not at the end of round $t-1$ as in vanilla Exp3). Moreover, $\widehat{L}_{t,i}$ is defined as a cumulative sum of loss estimates $\hat{\ell}_{s,i}$ with $s \equiv t \,(\mathrm{mod}\ 2)$ up to (and including) time t. To simplify some of the notation below, we introduce the shorthand notations $\mathbb{P}_t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_{t-2}]$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{t-2}]$.

We now describe an efficiently computable surrogate for M^*_t that will be used for constructing our loss estimates. First, fix a $t \geq 2$ and define $O'_{t-1,i} = O_{t-1,i}$ for all $i < I_{t-1}$ and $O'_{t-1,i} = O_{t-1,i+1}$ for all $i \geq I_{t-1}$. Then, define

$$M_t = \min\left\{i \in [N-1] : O'_{t-1,i} = 1\right\} \cup \{N\}.$$

It is easy to see that M_t follows a truncated geometric law in the sense that

$$\mathbb{P}[M_t = m] = \mathbb{P}[\min\{M^*_t, N\} = m]$$

holds for all $m \in [N]$. Finally, we define $K_{t,i}$ as a geometric random variable with parameter $p_{t,i} = \mathbb{P}[I_t = i|\mathcal{F}_{t-2}]$ and

$$G_{t,i} = \min\{K_{t,i}, M_t\}.$$

Using all this notation, we construct an estimate of $\ell_{t,i}$ as

$$\hat{\ell}_{t,i} = G_{t,i}O_{t,i}\ell_{t,i}. \tag{3}$$

Note that $G_{t,i}$ can be constructed for all i in parallel by sampling at most N independent copies of I_t.

The rationale underlying this definition of $G_{t,i}$ is rather delicate. First, note that p_t does not depend on O_{t-1}, thanks to the update rule (2). Second, $G_{t,i}$ is independent of $O_{t,i}$ given $\mathcal{F}_{t-2}$, so we can use the product rule of expectations to evaluate $\mathbb{E}\left[\hat{\ell}_{t,i}\right]$. The next lemma shows two key properties of the loss estimates (3), relying on the observations above.

Lemma 1. *Assume $r \geq \frac{\log T}{2N}$. Then, $\mathbb{E}_t\left[\hat{\ell}_{t,i}\right] \leq \ell_{t,i}$ and*

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

Proof. First, note that the law of $G_{t,i}$ is geometric with parameter $o_{t,i}$ truncated at N, as it is defined as the minimum of a geometric and a truncated geometric random variable. Using Lemma 1 of Neu and Bartók (2013) we get

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

Using this fact along with $\mathbb{E}_t[O_{t,i}] = o_{t,i}$ and the independence of $G_{t,i}$ and $O_{t,i}$, we get

$$\mathbb{E}_t[G_{t,i}O_{t,i}\ell_{t,i}] \leq \ell_{t,i},$$

Algorithm 1 DuplExp3

```
Input:
Set of actions $[N]$
Initialization:
$\widehat{L}_{-1,i} \leftarrow 0$ for $i \in [N]$
$\widehat{L}_{0,i} \leftarrow 0$ for $i \in [N]$
$O_{0,i} \leftarrow 0$ for $i \in [N]$
Run:
for $t = 1$ to $T$ do
    $\tau_t = \{s \in [T] : s \equiv t \,(\mathrm{mod}\ 2);\ s < t\}$
    $\eta_t = \sqrt{\log(N) / \left(N^2 + \sum_{s\in\tau_t} \sum_{i=1}^N p_{s,i} (\hat{\ell}_{s,i})^2\right)}$
    $w_{t,i} \leftarrow (1/N) \exp(-\eta_t \widehat{L}_{t-2,i})$ for $i \in [N]$
    $W_t \leftarrow \sum_{i=1}^N w_{t,i}$
    $p_{t,i} \leftarrow w_{t,i}/W_t$
    Choose $I_t \sim \mathbf{p}_t = (p_{t,1}, \ldots, p_{t,N})$
    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 Equation (3)
    $\hat{\ell}_{t,i} \leftarrow \ell_{t,i} O_{t,i} G_{t,i}$ for all $i \in [N]$
    $\widehat{L}_{t,i} = \widehat{L}_{t-2,i} + \hat{\ell}_{t,i}$ for all $i \in [N]$
end for
```

proving the first statement. For the second statement, we have

$$\mathbb{E}_t \left[\sum_{k=1}^N p_{t,i} \hat{\ell}_{t,i} \right] = \sum_{k=1}^N p_{t,i} \ell_{t,i} \left(1 - (1 - o_{t,i})^N\right)$$
$$= \sum_{k=1}^N p_{t,i} \ell_{t,i} - \sum_{k=1}^N p_{t,i} \ell_{t,i} (1 - o_{t,i})^N .$$

The proof is concluded by observing that

$$\sum_{i=1}^N p_{t,i} \ell_{t,i} (1 - o_{t,i})^N \le \sum_{i=1}^N p_{t,i} (1 - r)^N$$
$$= (1 - r)^N \le e^{-rN} \le \frac{1}{\sqrt{T}}$$

holds by our assumption on r. □

The next theorem states our main result concerning DuplExp3.

Theorem 1. *Assume that* $r \ge \frac{\log T}{2N}$ *and set*

$$\eta_t = \sqrt{\frac{\log(N)}{N^2 + \sum_{s\in\tau_t} \sum_{i=1}^N p_{s,i} (\hat{\ell}_{s,i})^2}}$$

for $\tau_t = \{s \in [T] : s \equiv t \,(\mathit{mod}\ 2);\ s < t\}$. *Then, the expected regret of* DuplExp3 *satisfies*

$$R_T \le 4 \sqrt{\left(\frac{T}{r} + N^2\right) \log N} + \sqrt{T}.$$

Proof. For the simplicity of the analysis, we assume that T is even and consider two sub-algorithms: one which uses the losses from the even rounds and estimates r from the odd rounds and one that acts conversely. This way, we can study regret bounds for both sub-algorithms separately. Let

$$T_k = \{t \in [T] : t \equiv k \,(\mathrm{mod}\ 2)\} \text{ for } k = 0, 1$$

be the set of rounds in which we play according to the sub-algorithm k. To obtain a regret bound for each individual sub-algorithm, we follow the proof of Lemma 1 in Györfi and Ottucsák (2007) to get

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

Taking expectations and summing over rounds in T_k, we get

$$\mathbb{E} \left[\sum_{t\in T_k} \sum_{i=1}^N p_{t,i} \hat{\ell}_{t,i} \right] \le \mathbb{E} \left[\sum_{t\in T_k} \frac{\eta_t}{2} \sum_{i=1}^N p_{t,i} (\hat{\ell}_{t,i})^2 \right]$$
$$+ \mathbb{E} \left[\sum_{t\in T_k} \left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+2}}{\eta_{t+2}} \right) \right].$$

Thus, we are left with bounding the three expectations above. For the first one, we use the second statement of Lemma 1 to get

$$\mathbb{E} \left[\sum_{t\in T_k} \sum_{i=1}^N p_{t,i} \hat{\ell}_{t,i} \right] \ge \sum_{t\in T_k} \sum_{i=1}^N p_{t,i} \ell_{t,i} - \frac{\sqrt{T}}{2}$$
$$= \mathbb{E} \left[\sum_{t\in T_k} \ell_{t,I_t} \right] - \frac{\sqrt{T}}{2}.$$

To simplify some notation below, let us define $b_t = \sum_{i=1}^N p_{t,i} (\hat{\ell}_{t,i})^2$. By our definition of η_t and the help of Lemma 3.5 of Auer et al. (2002b), we can bound the second expectation as

$$\mathbb{E} \left[\sum_{t\in T_k} \frac{\eta_t b_t}{2} \right] \le \mathbb{E} \left[\sqrt{\log(N) \left(N^2 + \sum_{t\in T_k} b_t\right)} \right]$$
$$\le \sqrt{\log(N) \left(N^2 + \sum_{t\in T_k} \mathbb{E}\left[b_t\right]\right)},$$

where we also used Jensen's inequality in the last line. We continue by bounding

$$\mathbb{E}_t \left[\sum_{i=1}^N p_{t,i} (\hat{\ell}_{t,i})^2 \right] = \sum_{i=1}^N p_{t,i} \ell_{t,i}^2 \mathbb{E}_t \left[O_{t,i} G_{t,i}^2 \right]$$
$$\le \sum_{i=1}^N p_{t,i} o_{t,i} \frac{2 - o_{t,i}}{o_{t,i}^2} \le \sum_{i=1}^N p_{t,i} \frac{2}{o_{t,i}}$$
$$\le \sum_{i=1}^N p_{t,i} \frac{2}{r} = \frac{2}{r},$$

where we used $\mathbb{E}_t\left[G_{t,i}^2\right] \leq \mathbb{E}_t\left[\left(G_{t,i}^*\right)^2\right] = \frac{2-o_{t,i}}{o_{t,i}^2}$ and $o_{t,i} \geq r$. Thus, we obtain

$$\mathbb{E}\left[\sum_{t\in T_k} \frac{\eta_t b_t}{2}\right] \leq \sqrt{\left(\frac{T}{r} + N^2\right)\log(N)}. \qquad (4)$$

Finally, the sum in the last expectation telescopes to

$$\mathbb{E}\left[\sum_{t\in T_k}\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+2}}{\eta_{t+2}}\right)\right] = \mathbb{E}\left[-\frac{\log W_{\overline{T}_k+2}}{\eta_{\overline{T}_k+2}}\right],$$

where $\overline{T}_k = \max(T_k)$. Using definition of W_t we get the following upper-bound for the last expression which holds for any arm $j \in [N]$:

$$\mathbb{E}\left[-\frac{\log W_{\overline{T}_k+2}}{\eta_{\overline{T}_k+2}}\right] \leq \mathbb{E}\left[\frac{\log N}{\eta_{\overline{T}_k+2}}\right] + \mathbb{E}\left[\widehat{L}_{\overline{T}_k,j}\right]$$

Now note that the first term can be bounded with the help of Equation (4). Using $\mathbb{E}_t\left[\hat{\ell}_{t,i}\right] \leq \ell_{t,i}$ from Lemma 1 and combining everything together, we obtain the regret bound for sub-algorithm k

$$\begin{aligned} R_{T_k} &= \mathbb{E}\left[\sum_{t\in T_k} \ell_{t,I_t}\right] + \max_{j\in[N]} \mathbb{E}\left[-\sum_{t\in T_k} \ell_{t,j}\right] \\ &\leq 2\sqrt{\frac{T}{r}\log(N) + N^2\log(N)} + \frac{\sqrt{T}}{2}. \end{aligned}$$

The final regret bound is simply the sum of bounds for both sub-algorithms. □

4 Generalizing to smaller values of r

The main shortcoming of the algorithm presented in the previous section is that it relies on the assumption $r \geq \frac{\log T}{2N}$ which might not always hold. In this section we provide a more general algorithm, DuplExp3($\underline{r}$), that works without this assumption. However, the algorithm still needs a reasonable lower bound $\underline{r}$ such that $\underline{r} \leq r$. In Section 5, we give a procedure that provides such a lower bound that also satisfies $\mathbb{E}\left[\underline{r}\right] \geq r/\left(\log\log T + 1\right)$ at the price of an additive $O(\log T)$ term in the regret.

Intuitively, the trouble caused by small values of r is that there might not be enough samples in a single round to guarantee a reasonably low bias for our loss estimate. Consequently, DuplExp3($\underline{r}$) is a simple modification of DuplExp3 that groups multiple rounds together to obtain more samples for reliable estimation of r. Precisely, the algorithm operates in episodes of length

$$A = \left\lceil \frac{\log T}{N\underline{r}} \right\rceil,$$

and the distribution of actions within episode j is fixed. The algorithm is initialized by setting $w_{1,i} = w_{2,i} = 1/N$ for all i, and performs the updates

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

Here, $\widehat{L}_{j,i}$ is the sum of loss estimates $\hat{\ell}_{s,i}$ for all $s \leq jA$ coming from the episode of the same parity as j, and $\hat{\ell}_{t,i}$ is computed as below. For all rounds t belonging to episode j, our algorithm draws action I_t according to the distribution

$$\mathbb{P}\left[I_t = i \middle| \mathcal{F}'_{j-2}\right] = p_{j,i} \propto w_{j,i},$$

where we also defined $\mathcal{F}'_j = \mathcal{F}_{jA}$.

Moving on, we now define a surrogate to M_j^* for episode $j \geq 2$. For all rounds t within episode $j-1$, we define $O'_{t,i}$ as in Section 3 and define O''_{j-1} as the concatenation of O_t's within episode $j-1$. More formally, for all $h \in [A(N-1)]$ we set

$$O''_{j-1,h} = O'_{(j-1)A+m,i}$$

where $m \in [A-1]$ and $i \in [N-1]$ are the unique integers such that $h = m(N-1) + i$. Using this notation, we define

$$M_j = \min\left\{i \in [A(N-1)] : O''_{j-1,i} = 1\right\} \cup \left\{A(N-1)\right\}.$$

By construction, M_j follows the law of a geometrically distributed random variable truncated at $A(N-1)$:

$$\mathbb{P}\left[M_j = m\right] = \mathbb{P}\left[\min\left\{M_j^*, A(N-1)\right\} = m\right].$$

Also defining $K_{j,i}$ as a geometrically distributed random variable with parameter $p_{j,i}$, we set

$$G_{j,i} = \min\left\{K_{j,i}, M_j\right\}$$

and finally define the estimate of $\ell_{t,i}$ as

$$\hat{\ell}_{t,i} = G_{j,i} O_{j,i} \ell_{t,i}$$

for all t within episode j. Furthermore, the sum of loss estimates within episode j is defined for all t, i as

$$\Delta\widehat{L}_{t,i} = \sum_{t=(j-1)A+1}^{jA} \hat{\ell}_{t,i}.$$

The number of episodes is defined as $J = \lceil T/A \rceil$.

We now state our main result concerning the performance of DuplExp3($\underline{r}$).

Theorem 2. *Assume that* $0 < \underline{r} \leq r$ *and set*

$$\eta_j = \sqrt{\frac{\log(N)}{A^2N^2 + \sum_{s\in\tau_j}\sum_{i=1}^N p_{s,i}(\Delta\widehat{L}_{s,i})^2}}$$

for $\tau_j = \{s \le T/A : s \equiv j\,(mod\ 2);\ s < j\}$*. Then, the expected regret of* DuplExp3($\underline{r}$) *satisfies*

$$R_T \le 4\sqrt{\left(T\left(\frac{1}{r}+(A-1)\right)+A^2N^2\right)\log(N)}+\sqrt{T}+A.$$

Before sketching the proof of this result, we note that when $\underline{r} > (\log T)/N$, DuplExp3($\underline{r}$) becomes identical to DuplExp3 and the above bound coincides with the bound of Theorem 1. A full proof is provided in the supplementary material.

Proof sketch. The theorem can be proven similarly as Theorem 1, with the main difference that the role of the loss estimates $\widehat{\ell}_t$ is now taken by $\widehat{L}_j$, and the time horizon T is replaced by J. For simplicity, let us assume that $J = T/A$ – which we can do at the expense of an additive term of A in the regret. Also, let us define $\mathbb{E}_j\left[\cdot\right] = \mathbb{E}\left[\cdot\middle|\,\mathcal{F}'_{j-1}\right]$ and $t_j = (j-1)A+1$.

The only challenge left is controlling $\mathbb{E}_j\left[\Delta\widehat{L}^2_{j,i}\right]$. This is done as

$$\begin{aligned}
&\mathbb{E}_j\left[\Delta\widehat{L}^2_{j,i}\right] = \mathbb{E}_j\left[\ell^2_{j,i}G^2_{j,i}\left(\sum_{t=(j-1)A+1}^{jA} O_{t,i}\right)^2\right] \\
&\le \frac{2-o_{j,i}}{o^2_{j,i}}\mathbb{E}\left[\sum_{t=(j-1)A+1}^{jA} O^2_{t,i} + \sum_{t,s=(j-1)A+1, s\neq t}^{jA} O_{t,i}O_{r,i}\right] \\
&\le \frac{2-o_{j,i}}{o^2_{j,i}}\left(Ao_{j,i} + A(A-1)o^2_{j,i}\right) \\
&\le \frac{2}{o_{j,i}}\left(A + A(A-1)o_{j,i}\right) \le 2A\left(\frac{1}{r}+(A-1)\right).
\end{aligned}$$

where in the last line we used that the expectation of $O^2_{t,i}$ is $o_{j,i}$ and $O_{t,i}$ and $O_{r,i}$ are conditionally independent for $t \neq r$. The statement follows from combining the above inequality with the steps of the proof of Theorem 1. □

5 Estimating r

We now turn to define a simple method for finding an appropriate lower bound $\underline{r}$ on r. We will crucially use the fact that, for a geometrically distributed random variable M with parameter r, we have

$$\mathbb{P}\left[M+1 \ge \frac{1}{r}\right] \ge \frac{1}{e}.$$

This implies that for k independent samples $M_1, \dots, M_k$,

$$\mathbb{P}\left[\max_j M_j + 1 \le \frac{1}{r}\right] \le \left(1-\frac{1}{e}\right)^k.$$

That is, setting $k = \lceil (e\log T)/(2) \rceil$, we can obtain a lower bound $\underline{r} = (\max_j M_j + 1)^{-1}$ that satisfies

$$\mathbb{P}\left[\underline{r} \le r\right] \ge 1 - \frac{1}{\sqrt{T}}. \tag{5}$$

In our specific setting, sampling M amounts to observing the gaps between consecutive side observations, when ordered in lexicographic order. To avoid being caught in a nasty loop, our algorithm also includes an initial sampling period that terminates if $r \le 1/N$ holds with high probability. Details of this estimation procedure are given in Algorithm 2.

Algorithm 2 Estimating $\underline{r}$

```
Input:
$k = \left\lceil \frac{e\log T}{2} \right\rceil$, $C = \left\lceil \frac{2\log T}{N} \right\rceil$.
Initialization:
$j \leftarrow 0$, $c \leftarrow 0$.
for $t = 1$ to $C$ do
    Draw $I_t \sim U([N])$.
    for $i = 1$ to $N$ do
        $c \leftarrow c + O_{t,i}\mathbb{I}_{\{i \neq I_t\}}$.
    end for
end for
if $c/(C(N-1)) \le 3/(2N)$ then
    return $\underline{r} = 0$.
else
    for $t = C+1$ to $T$ do
        Draw $I_t \sim U([N])$.
        for $i = 1$ to $N$ do
            $M_j \leftarrow M_j + \mathbb{I}_{\{i \neq I_t\}}$.
            $j \leftarrow j + O_{t,i}\mathbb{I}_{\{i \neq I_t\}}$.
            if $j = k$ then
                return $\underline{r} = (\max_n M_n + 1)^{-1}$
            else
                $M_j = 0$.
            end if
        end for
    end for
end if
```

The following lemma summarizes some important properties of Algorithm 2.

Lemma 2. *Let* $\underline{r}$ *be the output of Algorithm 2 and let* τ *be the index of the round when Algorithm 2 terminates. Then, for all values of* $r \in [0,1]$*, we have*

$$\mathbb{P}\left[\underline{r} \le r\right] \ge 1 - \frac{1}{\sqrt{T}}.$$

Furthermore, the following statements hold if $r \le 1/N$*:*

1. $\mathbb{P}\left[\underline{r} = 0\right] \ge 1 - \frac{1}{\sqrt{T}}$.
2. $\mathbb{E}\left[\tau\right] = \frac{4\log T}{N} + \sqrt{T} + 1$.

The following statements hold if $r \geq 2/N$:

1. $\mathbb{P}\left[\underline{r} = 0\right] \leq \frac{1}{\sqrt{T}}$.
2. $\mathbb{E}\left[\tau\right] \leq \log T \left(\frac{4}{N} + \frac{e}{2}\right) + 2$.

Finally, if $r \in [1/N, 2/N]$, then

$$\mathbb{E}\left[\tau\right] \leq \log T \left(\tfrac{4}{N} + e\right) + 2.$$

Proof. To avoid ambiguity, we set $\underline{r} = 0$ if Algorithm 2 does not terminate until the final round. The first statement obviously holds whenever $\underline{r} = 0$. Otherwise, the statement is implied by Equation (5).

To proceed, assume that $r \leq 1/N$. In this case, Hoeffding's inequality applied to $c/(C(N-1))$ implies

$$\mathbb{P}\left[\frac{c}{C(N-1)} \leq \frac{3}{2N}\right] \geq \mathbb{P}\left[\frac{c}{C(N-1)} \leq r + \frac{1}{2N}\right] \geq 1 - e^{-C(N-1)/(2N^2)} \geq 1 - e^{-C/(4N)},$$

where we used $N \geq 2$ in the last inequality. Combining the above result with the choice of C gives $\mathbb{P}\left[\underline{r} = 0\right] \geq 1 - 1/\sqrt{T}$ as claimed for this case. Then, the statement concerning τ follows from upper bounding the worst-case running time of Algorithm 2 by T:

$$\mathbb{E}\left[\tau\right] \leq C + T \cdot \mathbb{P}\left[\underline{r} > 0\right].$$

Now, assume that $r \geq 2/N$. The statement concerning $\mathbb{P}\left[\underline{r} = 0\right]$ can be proven by an identical argument as used for the case $r \leq 1/N$ above. For the statement concerning τ, observe that

$$\mathbb{E}\left[\tau\right] = C + \frac{k}{r(N-1)},$$

and use $r(N-1) \geq (2/N) \cdot (N/2) = 1$. The statement follows from plugging in the values of C and k.

Finally, assume that $r \in (1/N, 2/N)$. The bound on $\mathbb{E}\left[\tau\right]$ for this case is proven similarly as for the case $r \geq 2/N$. □

The following property will also be useful.

Lemma 3. *Assume that Algorithm 2 returns $\underline{r} > 0$. Then,*

$$\mathbb{E}\left[\frac{1}{\underline{r}}\right] \leq \frac{\log\log T + 1}{r}.$$

Proof. First, observe that $1/\underline{r} = \max_j M_j + 1$. For simplicity, we will use the notation $M^+ = \max_j M_j$. Notice that M^+ is nonnegative almost surely, thus its expectation can be expressed as $\mathbb{E}\left[M^+\right] = \sum_{m=1}^{\infty} \mathbb{P}\left[M^+ > m\right]$. Therefore, we have for any positive integer d that

$$\begin{aligned}\mathbb{E}\left[M^+\right] &= \sum_{m=1}^{\infty} \mathbb{P}\left[M^+ > m\right] \leq d + k \sum_{m=d+1}^{\infty} \mathbb{P}\left[M_1 > m\right] \\ &= d + k \sum_{m=d+1}^{\infty} (1-r)^m \\ &= d + k\frac{(1-r)^{d+1}}{r} \leq d + k\frac{e^{-(d+1)r}}{r},\end{aligned}$$

where the first inequality follows from the union bound and the last from $1 - r \leq e^{-r}$ that holds for all $r \in \mathbb{R}$. Setting $d = \left\lfloor \frac{\log k}{r} \right\rfloor$ gives

$$\mathbb{E}\left[M^+\right] \leq \frac{\log k + 1}{r}.$$

The statement of the lemma trivially follows when $k = 1$. Assuming $k > 1$, we have $k \leq e \log T$, implying $\log k \leq \log\log T + 1$ and thus the statement of the lemma for this case. □

6 A universal algorithm

In this section, we combine all previous results in the paper and provide a meta-algorithm that works without any prior knowledge on r whatsoever. This universal algorithm goes through the following steps:

1. Run Algorithm 2 to obtain $\underline{r}$.
2. If $\underline{r} = 0$, then run vanilla Exp3 with parameter $\eta = \sqrt{(2\log N)/(TN)}$ for the remaining time steps $t = \tau, \tau + 1, \ldots, T$.
3. If $\underline{r} > 0$, then run DuplExp3($\underline{r}$).

The following theorem provides a performance guarantee for this algorithm.

Theorem 3. *For $r \leq 1/N$, the expected regret of our algorithm satisfies*

$$R_T \leq \sqrt{2TN\log N} + \frac{4\log T}{N} + \sqrt{T} + 1.$$

For, $r \geq 2/N$, the regret can be bounded as

$$\begin{aligned}R_T \leq\ & 4\sqrt{\frac{T\log N}{r}\left(\frac{\log T\left(\log\log T + 1\right)}{N} + 1\right)} \\ &+ 4\frac{\log T\left(\log\log T + 1\right)\sqrt{\log N}}{r} \\ &+ 3\sqrt{T} + \log T\left(\frac{4}{N} + e\right) + 2.\end{aligned}$$

Otherwise, the expected regret is bounded by the maximum of the above two bounds.

Proof. Note that running Algorithm 2 adds an additive $\frac{4\log T}{N}+1$ factor to the regret, no matter what the value of r is.

First, let us consider the case when $r \le 1/N$. In this case, $\underline{r}=0$ holds with probability at least $1-\sqrt{T}$, and thus Exp3 is invoked for rounds following $\tau = C$. Hence, in these rounds, the expected regret of our algorithm is bounded by $\sqrt{2TN\log N}$ (see, e.g., Bubeck and Cesa-Bianchi, 2012). If $\underline{r}>0$ holds despite $r \le 1/N$, our algorithm suffers a regret of at most T. However, this only occurs with probability at most $1/\sqrt{T}$, resulting in an expected contribution of at most $\sqrt{T}$ to the regret.

In the case $r \ge 2/N$, $\underline{r}=0$ only occurs with probability at most $1/\sqrt{T}$, once again contributing at most $\sqrt{T}$ to the expected regret. Similarly, the probability of Algorithm 2 failing when $\underline{r}>0$ is at most $1/\sqrt{T}$, increasing the expected regret by another $\sqrt{T}$ term. On the other hand, when $\underline{r}>0$ and $\underline{r}\le r$, Theorem 2 guarantees a regret of at most

$$4\sqrt{T\log(N)\left(\frac{1}{r}+(A-1)\right)}+4AN\sqrt{\log(N)}+\sqrt{T}.$$

Now, using the definition of A and Lemma 3 to bound $\mathbb{E}[A]$, we obtain the first three terms of the bound stated in the theorem. The final terms are obtained by bounding the contribution of running Algorithm 2 to the regret with the help of Lemma 2.

Finally, for the case when $r\in(1/N,2/N)$, observe that the total expected regret for rounds $t>\tau$ is bounded by either $\sqrt{2TN\log N}$ when $\underline{r}=0$ or the bound of Theorem 2, when $\underline{r}>0$. In either case, running Algorithm 2 increases the regret by at most $\log T\left(\frac{4}{N}+\frac{1}{e}\right)+2$ on expectation. Combining these terms concludes the proof of the theorem. □

7 Conclusions & future work

In this paper, we have proposed an efficient algorithm for learning in multi-armed bandit problems with side observations. In the setting when the learner has access to a lower bound $\underline{r}$ on the observation probability r, this algorithm achieves a regret of $O(\sqrt{(T/\underline{r})((\log T)/N+1)\log N})$ (Theorem 2). This result also implies that the regret of our algorithm becomes $C(\sqrt{(T/r)\log N})$ for some positive constant C if $\log T \le N$. This assumption can be often verified for large-scale decision problems such as recommender systems, where T is the number of recommendations made to users and N is the number of items to choose from. Typically, N is in the order of thousands, which permits values of T far beyond machine precision. Thus, the finite-time performance guarantees of our algorithm are very similar to the guarantees of a hypothetical algorithm that knows the exact values of r. Another interesting question is whether our algorithm fails gracefully as the value of r approaches $1/N$. In this case, Theorem 2 guarantees a regret of $O(\sqrt{T(N+\log T)\log N})$. This guarantee, while asymptotically inferior to that of Exp3, can still be competitive in the regime where $\log T \le N$. Whether it is possible to remove the $\sqrt{\log T}$ factor from the asymptotic bounds remains an open question.

Another important corollary of our results is that, under some assumptions, it is possible to leverage side observations without having access to the second neighborhood in the side-observation graph as defined by Mannor and Shamir (2011). A natural question is then whether it is possible to relax our assumptions concerning the side-observation graph. There are several possible extensions that our approach is able to tackle. For instance, assume that the side-observation graphs are generated by a sequence of Erdős-Rényi models parametrized by $(r_t)_{t=1}^T$. Assuming that this sequence is such that $r_t \ne r_{t-1}$ occurs only K times and $r_t \ge (\log T)/N$ for all rounds, our algorithm can be shown to achieve a regret of $\widetilde{O}(\sqrt{(T/\rho)}+K)$, where ρ is the average of the r_t-s. However, it is yet unclear whether it is possible to improve the worst-case bound of $\widetilde{O}(\sqrt{TN})$ for arbitrary random graph models. Actually, the ultimate goal of achieving such improvements in the extreme case of adversarially generated observation graphs seems very difficult: In such cases, knowing the second neighborhoods of the chosen actions is indispensable for constructing reliable loss estimates (Kocák et al., 2014).

Finally, let us comment on the apparent conceptual complexity of our algorithm. In particular, one might wonder if it might be possible to obtain a reliable estimate $\hat{r}$ of r by a procedure similar to Algorithm 2 and then use $\hat{r}$ in place of r to compute the loss estimates (1). For instance, one can construct an estimate $\hat{r}$ based on K samples that guarantees $|r-\hat{r}|=O_P(\sqrt{r/K})$. Plugging this estimate into Equation (1) to compute $\hat{\ell}_{t,i}$, one can only show $|\hat{\ell}_{t,i}-\hat{\ell}^*_{t,i}|=O(1/\sqrt{Kr^3})$, which necessarily results in an additive $O(T\cdot\sqrt{1/(Kr^3)})$ term in the expected regret on top of the regret of K suffered for the exploration period. Optimizing the resulting bound gives a guarantee of $O(T^{2/3}/r)$. A similar obstacle was faced by Kanade et al. (2009), who studied the so-called sleeping bandit problem and gave an algorithm with regret of $O(T^{4/5})$. The main terms of their bound come from an estimation procedure similar to the one outlined above. The bound of Kanade et al. was recently improved by Neu and Valko (2014) to $O(T^{2/3})$ by using a similar technique as presented here.

References

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. (2013). From Bandits to Experts: A Tale of Domination and Independence. In *NIPS-26*.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002a). The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 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.

Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Now Publishers Inc.

Buccapatnam, S., Eryilmaz, A., and Shroff, Ness, B. (2014). Stochastic bandits with side observations on networks. In *ACM SIGMETRICS'14*.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. (2012). Leveraging Side Observations in Stochastic Bandits. In *Uncertainty in Artificial Intelligence*, pages 142–151.

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

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

Kanade, V., McMahan, H. B., and Bryan, B. (2009). Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *AISTATS 2009*, pages 272–279.

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

Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations. In *NIPS-24*, pages 684–692.

Neu, G. and Bartók, G. (2013). An efficient algorithm for learning with semi-bandit feedback. In *ALT 2013*, pages 234–248.

Neu, G. and Valko, M. (2014). Online combinatorial optimization with stochastic decision sets and adversarial losses. In *NIPS-27*.

Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori, Y. (2014). Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, page 280287.

A Full proof of Theorem 2

As in Theorem 1 we split the algorithm into two sub-algorithms, and analyze them separately. We define

$$J_k = \{j \in [J] : j \equiv k \,(\mathrm{mod}\ 2)\} \quad \text{for } k \in \{0,\, 1\}.$$

Following the same line of proof as for Theorem 1, we obtain

$$\sum_{j\in J_k}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}_{j,i}$$

$$\le \sum_{j\in J_k}\left(\frac{\log W_j}{\eta_j} - \frac{\log W'_{j+2}}{\eta_j}\right) + \sum_{j\in J_k}\frac{\eta_j}{2}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}^2_{j,i}$$

$$= -\frac{\log W_{J+2}}{\eta_{J+2}} + \sum_{j\in J_k}\frac{\eta_j}{2}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}^2_{j,i}.$$

Taking expectation, we have

$$\mathbb{E}\left[\sum_{j\in J_k}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}_{j,i}\right]$$

$$\le \mathbb{E}\left[-\frac{\log W_{J+2}}{\eta_{J+2}}\right] + \mathbb{E}\left[\sum_{j\in J_k}\frac{\eta_j}{2}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}^2_{j,i}\right]$$

$$\le \mathbb{E}\left[\frac{\log N}{\eta_{J+2}}\right] + \mathbb{E}\left[\widehat{L}_{J,k}\right] + \mathbb{E}\left[\sum_{j\in J_k}\frac{\eta_j}{2}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}^2_{j,i}\right]$$

First, we lower-bound the first expectation. By generalizing Lemma 1 to episodes of length A, we get

$$\mathbb{E}\left[\sum_{j\in J_k}\sum_{i=1}^{N} p_{j,i}\Delta\widehat{L}_{j+1,i}\right] \le \sum_{j\in J_k}\sum_{t=(j-1)A+1}^{jA}\mathbb{E}\left[\ell_{t,I_t}\right] - \frac{\sqrt{T}}{2},$$

We set

$$\eta_j = \sqrt{\frac{\log N}{A^2N^2 + \sum_{s\in\tau_j}\sum_{i=1}^{N} p_{s,i}(\Delta\widehat{L}_{s,i})^2}},$$

where $\tau_j = \{s \in [T] : s \equiv j \,(\mathrm{mod}\ 2);\ s < j\}$. Now for any $g \in [N]$ we get

$$R_{J_k} = \sum_{j\in J_k}\sum_{t=(j-1)A+1}^{jA}\mathbb{E}\left[\ell_{t,I_t}\right] + \max_{j\in[N]}\mathbb{E}\left[-\widehat{L}_{J,g}\right]$$

$$\le 2\mathbb{E}\left[\sqrt{\log N\left(A^2N^2 + \textstyle\sum_{j\in J_k}\sum_{i=1}^{N} p_{t,i}\Delta\widehat{L}^2_{j,i}\right)}\right]$$

$$\le 2\sqrt{\log N\left(A^2N^2 + \textstyle\sum_{j\in J_k}\sum_{i=1}^{N} p_{t,i}\mathbb{E}\left[\Delta\widehat{L}^2_{j,i}\right]\right)}$$

The only challenge left is controlling $\mathbb{E}_j\left[\Delta\widehat{L}^2_{j,i}\right]$. This is done as

$$\mathbb{E}_j\left[\Delta\widehat{L}^2_{j,i}\right] = \mathbb{E}_j\left[\ell^2_{j,i}G^2_{j,i}\left(\sum_{t=(j-1)A+1}^{jA} O_{t,i}\right)^2\right]$$

$$\le \frac{2-o_{j,i}}{o^2_{j,i}}\mathbb{E}\left[\sum_{t=(j-1)A+1}^{jA} O^2_{t,i} + \sum_{t,s=(j-1)A+1,s\neq t}^{jA} O_{t,i}O_{r,i}\right]$$

$$\le \frac{2-o_{j,i}}{o^2_{j,i}}\left(Ao_{j,i} + A(A-1)o^2_{j,i}\right)$$

$$\le \frac{2}{o_{j,i}}\left(A + A(A-1)o_{j,i}\right) \le 2A\left(\frac{1}{r} + (A-1)\right).$$

where in the last line we used that the expectation of $O^2_{t,i}$ is $o_{j,i}$ and $O_{t,i}$ and $O_{r,i}$ are conditionally independent for $t \neq r$.

Finally, putting everything together we get a regret bound for each sub-algorithm

$$R_{J_g} = \sum_{j\in J_k}\sum_{t=(j-1)A+1}^{jA}\mathbb{E}\left[\ell_{t,I_t}\right] + \max_{j\in[N]}\mathbb{E}\left[-\widehat{L}_{J,g}\right]$$

$$\le 2\sqrt{T\log(N)\left(\frac{1}{r} + (A-1)\right) + A^2N^2\log(N)} + \frac{\sqrt{T}}{2}$$

and a regret bound for DuplExp3($\underline{r}$) is a sum of regret bounds of the two sub-algorithms:

$$R_T = R_{J_0} + R_{J_1}.$$