HAL
archives-ouvertes.fr

Efficient learning by implicit exploration in bandit problems with side observations

Tomáš Kocák, Gergely Neu, Michal Valko, Rémi Munos

▶ **To cite this version:**

Tomáš Kocák, Gergely Neu, Michal Valko, Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. Neural Information Processing Systems, Dec 2014, Montréal, Canada. <hal-01079351v2>

HAL Id: hal-01079351
https://hal.inria.fr/hal-01079351v2

Submitted on 3 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficient learning by implicit exploration in bandit problems with side observations

Tomáš Kocák **Gergely Neu** **Michal Valko** **Rémi Munos***
SequeL team, INRIA Lille – Nord Europe, France
{tomas.kocak,gergely.neu,michal.valko,remi.munos}@inria.fr

Abstract

We consider online learning problems under a a partial observability model capturing situations where the information conveyed to the learner is between full information and bandit feedback. In the simplest variant, we assume that in addition to its own loss, the learner also gets to observe losses of some other actions. The revealed losses depend on the learner's action and a directed *observation system* chosen by the environment. For this setting, we propose the first algorithm that enjoys near-optimal regret guarantees without having to know the observation system before selecting its actions. Along similar lines, we also define a new partial information setting that models online combinatorial optimization problems where the feedback received by the learner is between semi-bandit and full feedback. As the predictions of our first algorithm cannot be always computed efficiently in this setting, we propose another algorithm with similar properties and with the benefit of always being computationally efficient, at the price of a slightly more complicated tuning mechanism. Both algorithms rely on a novel exploration strategy called *implicit exploration*, which is shown to be more efficient both computationally and information-theoretically than previously studied exploration strategies for the problem.

1 Introduction

Consider the problem of sequentially recommending content for a set of users. In each period of this online decision problem, we have to assign content from a news feed to each of our subscribers so as to maximize clickthrough. We assume that this assignment needs to be done well in advance, so that we only observe the actual content after the assignment was made and the user had the opportunity to click. While we can easily formalize the above problem in the classical multi-armed bandit framework [3], notice that we will be throwing out important information if we do so! The additional information in this problem comes from the fact that several news feeds can refer to the same content, giving us the opportunity to infer clickthroughs for a number of assignments that we *did not actually make*. For example, consider the situation shown on Figure 1a. In this simple example, we want to suggest one out of three news feeds to each user, that is, we want to choose a matching on the graph shown on Figure 1a which covers the users. Assume that news feeds 2 and 3 refer to the same content, so *whenever we assign news feed 2 or 3 to any of the users, we learn the value of both of these assignments*. The relations between these assignments can be described by a graph structure (shown on Figure 1b), where nodes represent user-news feed assignments, and edges mean that the corresponding assignments reveal the clickthroughs of each other. For a more compact representation, we can group the nodes by the users, and rephrase our task as having to choose one node from each group. Besides its own reward, each selected node reveals the rewards assigned to all their neighbors.

*Current affiliation: Google DeepMind

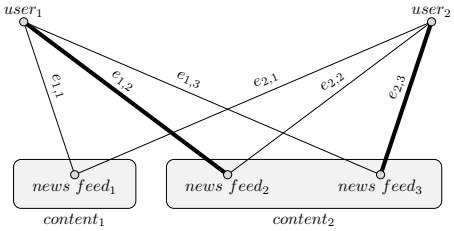

Figure 1a: Users and news feeds. The thick edges represent one potential matching of users to feeds, grouped news feeds show the same content.

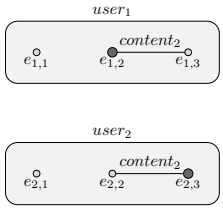

Figure 1b: Users and news feeds. Connected feeds mutually reveal each others clickthroughs.

The problem described above fits into the framework of *online combinatorial optimization* where in each round, a learner selects one of a very large number of available actions so as to minimize the losses associated with its sequence of decisions. Various instances of this problem have been widely studied in recent years under different feedback assumptions [7, 2, 8], notably including the so-called *full-information* [13] and *semi-bandit* [2, 16] settings. Using the example in Figure 1a, assuming full information means that clickthroughs are observable for *all* assignments, whereas assuming semi-bandit feedback, clickthroughs are only observable on the actually realized assignments. While it is unrealistic to assume full feedback in this setting, assuming semi-bandit feedback is far too restrictive in our example. Similar situations arise in other practical problems such as packet routing in computer networks where we may have additional information on the delays in the network besides the delays of our own packets.

In this paper, we generalize the partial observability model first proposed by Mannor and Shamir [15] and later revisited by Alon et al. [1] to accommodate the feedback settings situated between the full-information and the semi-bandit schemes. Formally, we consider a sequential decision making problem where in each time step t the (potentially *adversarial*) *environment* assigns a loss value to each out of d components, and generates an *observation system* whose role will be clarified soon. Obliviously of the environment's choices, the *learner* chooses an action $\boldsymbol{V}_t$ from a fixed action set $S \subset \{0,1\}^d$ represented by a binary vector with at most m nonzero components, and incurs the sum of losses associated with the nonzero components of $\boldsymbol{V}_t$. At the end of the round, the learner observes the individual losses along the chosen components *and some additional feedback based on its action and the observation system*. We represent this observation system by a directed *observability graph* with d nodes, with an edge connecting $i \to j$ if and only if the loss associated with j is revealed to the learner whenever $V_{t,i} = 1$. The goal of the learner is to minimize its total loss obtained over T repetitions of the above procedure. The two most well-studied variants of this general framework are the multi-armed bandit problem [3] where each action consists of a single component and the observability graph is a graph without edges, and the problem of prediction with expert advice [17, 14, 5] where each action consists of exactly one component and the observability graph is complete. In the true combinatorial setting where $m > 1$, the empty and complete graphs correspond to the semi-bandit and full-information settings respectively.

Our model directly extends the model of Alon et al. [1], whose setup coincides with $m = 1$ in our framework. Alon et al. themselves were motivated by the work of Mannor and Shamir [15], who considered *undirected* observability systems where actions mutually uncover each other's losses. Mannor and Shamir proposed an algorithm based on linear programming that achieves a regret of $\tilde{\mathcal{O}}(\sqrt{cT})$, where c is the number of cliques into which the graph can be split. Later, Alon et al. [1] proposed an algorithm called EXP3-SET that guarantees a regret of $\mathcal{O}(\sqrt{\alpha T \log d})$, where α is an upper bound on the *independence numbers* of the observability graphs assigned by the environment. In particular, this bound is tighter than the bound of Mannor and Shamir since $\alpha \leq c$ for any graph. Furthermore, EXP3-SET is much more efficient than the algorithm of Mannor and Shamir as it only requires running the EXP3 algorithm of Auer et al. [3] on the decision set, which runs in time linear in d. Alon et al. [1] also extend the model of Mannor and Shamir in allowing the observability graph to be directed. For this setting, they offer another algorithm called EXP3-DOM with similar guarantees, although with the serious drawback that it *requires access to the observation system before choosing its actions*. This assumption poses severe limitations to the practical applicability of EXP3-DOM, which also needs to solve a sequence of set cover problems as a subroutine.

In the present paper, we offer two computationally and information-theoretically efficient algorithms for bandit problems with directed observation systems. Both of our algorithms circumvent the costly exploration phase required by EXP3-DOM by a trick that we will refer to IX as in Implicit eXploration. Accordingly, we name our algorithms EXP3-IX and FPL-IX, which are variants of the well-known EXP3 [3] and FPL [12] algorithms enhanced with implicit exploration. Our first algorithm EXP3-IX is specifically designed[1] to work in the setting of Alon et al. [1] with $m = 1$ and does not need to solve any set cover problems or have any sort of prior knowledge concerning the observation systems chosen by the adversary.[2] FPL-IX, on the other hand, does need either to solve set cover problems or have a prior upper bound on the independence numbers of the observability graphs, but can be computed efficiently for a wide range of true combinatorial problems with $m > 1$. We note that our algorithms do not even need to know the number of rounds T and our regret bounds scale with the *average* independence number $\bar{\alpha}$ of the graphs played by the adversary rather than the largest of these numbers. They both employ adaptive learning rates and unlike EXP3-DOM, they do not need to use a doubling trick to be anytime or to aggregate outputs of multiple algorithms to optimally set their learning rates. Both algorithms achieve regret guarantees of $\tilde{\mathcal{O}}(m^{3/2}\sqrt{\bar{\alpha}T})$ in the combinatorial setting, which becomes $\tilde{\mathcal{O}}(\sqrt{\bar{\alpha}T})$ in the simple setting.

Before diving into the main content, we give an important graph-theoretic statement that we will rely on when analyzing both of our algorithms. The lemma is a generalized version of Lemma 13 of Alon et al. [1] and its proof is given in Appendix A.

Lemma 1. *Let G be a directed graph with vertex set $V = \{1, \dots, d\}$. Let N_i^- be the in-neighborhood of node i, i.e., the set of nodes j such that $(j \to i) \in G$. Let α be the independence number of G and $p_1, \dots, p_d$ are numbers from $[0, 1]$ such that $\sum_{i=1}^d p_i \leq m$. Then*

$$\sum_{i=1}^{d} \frac{p_i}{\frac{1}{m}p_i + \frac{1}{m}P_i + c} \leq 2m\alpha \log\left(1 + \frac{m\lceil d^2/c \rceil + d}{\alpha}\right) + 2m,$$

where $P_i = \sum_{j \in N_i^-} p_j$ and c is a positive constant.

2 Multi-armed bandit problems with side information

In this section, we start by the simplest setting fitting into our framework, namely the multi-armed bandit problem with side observations. We provide intuition about the implicit exploration procedure behind our algorithms and describe EXP3-IX, the most natural algorithm based on the IX trick.

The problem we consider is defined as follows. In each round $t = 1, 2, \dots, T$, the environment assigns a loss vector $\boldsymbol{\ell}_t \in [0, 1]^d$ for d actions and also selects an observation system described by the directed graph G_t. Then, based on its previous observations (and likely some external source of randomness) the learner selects action I_t and subsequently incurs and observes loss ℓ_{t,I_t}. Furthermore, the learner also observes the losses $\ell_{t,j}$ for all j such that $(I_t \to j) \in G_t$, denoted by the indicator $O_{t,i}$. Let $\mathcal{F}_{t-1} = \sigma(I_{t-1}, \dots, I_1)$ capture the interaction history up to time t. As usual in online settings [6], the performance is measured in terms of (total expected) regret, which is the difference between a total loss received and the total loss of the best single action chosen in hindsight,

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

where the expectation integrates over the random choices made by the learning algorithm. Alon et al. [1] adapted the well-known EXP3 algorithm of Auer et al. [3] for this precise problem. Their algorithm, EXP3-DOM, works by maintaining a weight $w_{t,i}$ for each individual arm $i \in [d]$ in each round, and selecting I_t according to the distribution

$$\mathbb{P}\left[I_t = i \,|\mathcal{F}_{t-1}\right] = (1-\gamma)p_{t,i} + \gamma\mu_{t,i} = (1-\gamma)\frac{w_{t,i}}{\sum_{j=1}^{d} w_{t,j}} + \gamma\mu_{t,i},$$

[1] EXP3-IX can also be efficiently implemented for some specific combinatorial decision sets even with $m > 1$, see, e.g., Cesa-Bianchi and Lugosi [7] for some examples.

[2] However, it is still necessary to have access to the observability graph to construct low bias estimates of losses, but only after the action is selected.

where $\gamma \in (0,1)$ is parameter of the algorithm and $\boldsymbol{\mu}_t$ is an *exploration distribution* whose role we will shortly clarify. After each round, EXP3-DOM defines the loss estimates

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i}} \mathbb{1}_{\{(I_t \to i) \in G_t\}} \quad \text{where} \quad o_{t,i} = \mathbb{E}\left[O_{t,i} \,|\mathcal{F}_{t-1}\right] = \mathbb{P}\left[(I_t \to i) \in G_t \,|\mathcal{F}_{t-1}\right]$$

for each $i \in [d]$. These loss estimates are then used to update the weights for all i as

$$w_{t+1,i} = w_{t,i} e^{-\gamma \hat{\ell}_{t,i}}.$$

It is easy to see that the these loss estimates $\hat{\ell}_{t,i}$ are unbiased estimates of the true losses whenever $p_{t,i} > 0$ holds for all i. This requirement along with another important technical issue justify the presence of the exploration distribution $\boldsymbol{\mu}_t$. The key idea behind EXP3-DOM is to compute a *dominating set* $D_t \subseteq [d]$ of the observability graph G_t in each round, and define $\boldsymbol{\mu}_t$ as the uniform distribution over D_t. This choice ensures that $o_{t,i} \geq p_{t,i} + \gamma/|D_t|$, a crucial requirement for the analysis of [1]. In what follows, we propose an exploration scheme that does not need any fancy computations but, more importantly, works *without any prior knowledge of the observability graphs*.

2.1 Efficient learning by implicit exploration

In this section, we propose the simplest exploration scheme imaginable, which consists of *merely pretending to explore*. Precisely, we simply sample our action I_t from the distribution defined as

$$\mathbb{P}\left[I_t = i \,|\mathcal{F}_{t-1}\right] = p_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{d} w_{t,j}}, \tag{1}$$

without explicitly mixing with any exploration distribution. Our key trick is to define the loss estimates for all arms i as

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i} + \gamma_t} \mathbb{1}_{\{(I_t \to i) \in G_t\}},$$

where $\gamma_t > 0$ is a parameter of our algorithm. It is easy to check that $\hat{\ell}_{t,i}$ is a *biased* estimate of $\ell_{t,i}$. The nature of this bias, however, is very special. First, observe that $\hat{\ell}_{t,i}$ is an *optimistic* estimate of $\ell_{t,i}$ in the sense that $\mathbb{E}\left[\hat{\ell}_{t,i} \,|\mathcal{F}_{t-1}\right] \leq \ell_{t,i}$. That is, our bias always ensures that, on expectation, we underestimate the loss of any fixed arm i. Even more importantly, our loss estimates also satisfy

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{d} p_{t,i} \hat{\ell}_{t,i} \,\middle|\, \mathcal{F}_{t-1}\right] &= \sum_{i=1}^{d} p_{t,i} \ell_{t,i} + \sum_{i=1}^{d} p_{t,i} \ell_{t,i} \left(\frac{o_{t,i}}{o_{t,i} + \gamma_t} - 1\right) \\ &= \sum_{i=1}^{d} p_{t,i} \ell_{t,i} - \gamma_t \sum_{i=1}^{d} \frac{p_{t,i} \ell_{t,i}}{o_{t,i} + \gamma_t}, \end{aligned} \tag{2}$$

that is, the bias of the estimated losses *suffered by our algorithm* is directly controlled by γ_t. As we will see in the analysis, it is sufficient to control the bias of our own estimated performance as long as we can guarantee that the loss estimates associated with any fixed arm are optimistic—which is precisely what we have. Note that this slight modification ensures that the denominator of $\hat{\ell}_{t,i}$ is lower bounded by $p_{t,i} + \gamma_t$, which is a very similar property as the one achieved by the exploration scheme used by EXP3-DOM. We call the above loss estimation method *implicit exploration* or IX, as it gives rise to the same effect as explicit exploration without actually having to implement any exploration policy. In fact, explicit and implicit explorations can both be regarded as two different approaches for bias-variance tradeoff: while explicit exploration biases the *sampling distribution* of I_t to reduce the variance of the loss estimates, implicit exploration achieves the same result by biasing *the loss estimates themselves*.

From this point on, we take a somewhat more predictable course and define our algorithm EXP3-IX as a variant of EXP3 using the IX loss estimates. One of the twists is that EXP3-IX is actually based on the adaptive learning-rate variant of EXP3 proposed by Auer et al. [4], which avoids the necessity of prior knowledge of the observability graphs in order to set a proper learning rate. This algorithm is defined by setting $\widehat{L}_{t-1,i} = \sum_{s=1}^{t-1} \hat{\ell}_{s,i}$ and for all $i \in [d]$ computing the weights as

$$w_{t,i} = (1/d) e^{-\eta_t \widehat{L}_{t-1,i}}.$$

These weights are then used to construct the sampling distribution of I_t as defined in (1). The resulting EXP3-IX algorithm is shown as Algorithm 1.

2.2 Performance guarantees for EXP3-IX

Our analysis follows the footsteps of Auer et al. [3] and Györfi and Ottucsák [9], who provide an improved analysis of the adaptive learning-rate rule proposed by Auer et al. [4]. However, a technical subtlety will force us to proceed a little differently than these standard proofs: for achieving the tightest possible bounds and the most efficient algorithm, we need to tune our learning rates according to some random quantities that depend on the performance of EXP3-IX. In fact, the key quantities in our analysis are the terms

$$Q_t = \sum_{i=1}^{d} \frac{p_{t,i}}{o_{t,i} + \gamma_t},$$

which depend on the interaction history $\mathcal{F}_{t-1}$ for all t. Our theorem below gives the performance guarantee for EXP3-IX using a parameter setting adaptive to the values of Q_t. A full proof of the theorem is given in the supplementary material.

Algorithm 1 EXP3-IX

Input: Set of actions $\mathcal{S} = [d]$,
  parameters $\gamma_t \in (0,1)$, $\eta_t > 0$ for $t \in [T]$.
for $t = 1$ **to** T **do**
  $w_{t,i} \leftarrow (1/d)\exp\left(-\eta_t \widehat{L}_{t-1,i}\right)$ for $i \in [d]$
  An adversary privately chooses losses $\ell_{t,i}$ for $i \in [d]$ and generates a graph G_t
  $W_t \leftarrow \sum_{i=1}^{d} w_{t,i}$
  $p_{t,i} \leftarrow w_{t,i}/W_t$
  Choose $I_t \sim \boldsymbol{p}_t = (p_{t,1}, \ldots, p_{t,d})$
  Observe graph G_t
  Observe pairs $\{i, \ell_{t,i}\}$ for $(I_t \to i) \in G_t$
  $o_{t,i} \leftarrow \sum_{(j \to i) \in G_t} p_{t,j}$ for $i \in [d]$
  $\hat{\ell}_{t,i} \leftarrow \frac{\ell_{t,i}}{o_{t,i}+\gamma_t} \mathbb{1}_{\{(I_t \to i) \in G_t\}}$ for $i \in [d]$
end for

Theorem 1. *Setting* $\eta_t = \gamma_t = \sqrt{(\log d)/(d + \sum_{s=1}^{t-1} Q_s)}$ *, the regret of* EXP3-IX *satisfies*

$$R_T \le 4\mathbb{E}\left[\sqrt{\left(d + \textstyle\sum_{t=1}^{T} Q_t\right)\log d}\right]. \tag{3}$$

Proof sketch. Following the proof of Lemma 1 in Györfi and Ottucsák [9], we can prove that

$$\sum_{i=1}^{d} p_{t,i}\hat{\ell}_{t,i} \le \frac{\eta_t}{2}\sum_{i=1}^{d} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2 + \left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right). \tag{4}$$

Taking *conditional* expectations, using Equation (2) and summing up both sides, we get

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

Using Lemma 3.5 of Auer et al. [4] and plugging in η_t and γ_t, this becomes

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_{t,i}\ell_{t,i} \le 3\sqrt{\left(d + \textstyle\sum_{t=1}^{T} Q_t\right)\log d} + \sum_{t=1}^{T}\mathbb{E}\left[\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right)\middle|\, \mathcal{F}_{t-1}\right].$$

Taking expectations on both sides, the second term on the right hand side telescopes into

$$\mathbb{E}\left[\frac{\log W_1}{\eta_1} - \frac{\log W_{T+1}}{\eta_{T+1}}\right] \le \mathbb{E}\left[-\frac{\log w_{T+1,j}}{\eta_{T+1}}\right] = \mathbb{E}\left[\frac{\log d}{\eta_{T+1}}\right] + \mathbb{E}\left[\hat{L}_{T,j}\right]$$

for any $j \in [d]$, giving the desired result as

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_{t,i}\ell_{t,i} \le \sum_{t=1}^{T}\ell_{t,j} + 4\mathbb{E}\left[\sqrt{\left(d + \textstyle\sum_{t=1}^{T} Q_t\right)\log d}\right],$$

where we used the definition of η_T and the optimistic property of the loss estimates. □

Setting $m = 1$ and $c = \gamma_t$ in Lemma 1, gives the following *deterministic* upper bound on each Q_t.

Lemma 2. *For all* $t \in [T]$,

$$Q_t = \sum_{i=1}^{d} \frac{p_{t,i}}{o_{t,i} + \gamma_t} \le 2\alpha_t \log\left(1 + \frac{\lceil d^2/\gamma_t \rceil + d}{\alpha_t}\right) + 2.$$

Combining Lemma 2 with Theorem 1 we prove our main result concerning the regret of EXP3-IX.

Corollary 1. *The regret of* EXP3-IX *satisfies*

$$R_T \leq 4\sqrt{\left(d + 2\sum_{t=1}^{T}\left(H_t\alpha_t + 1\right)\right)\log d},$$

where

$$H_t = \log\left(1 + \frac{\lceil d^2\sqrt{td/\log d}\rceil + d}{\alpha_t}\right) = \mathcal{O}(\log(dT)).$$

3 Combinatorial semi-bandit problems with side observations

We now turn our attention to the setting of online combinatorial optimization (see [13, 7, 2]). In this variant of the online learning problem, the learner has access to a possibly huge action set $\mathcal{S} \subseteq \{0,1\}^d$ where each action is represented by a binary vector $\boldsymbol{v}$ of dimensionality d. In what follows, we assume that $\|\boldsymbol{v}\|_1 \leq m$ holds for all $\boldsymbol{v} \in \mathcal{S}$ and some $1 \leq m \ll d$, with the case $m = 1$ corresponding to the multi-armed bandit setting considered in the previous section. In each round $t = 1, 2, \ldots, T$ of the decision process, the learner picks an action $\boldsymbol{V}_t \in \mathcal{S}$ and incurs a loss of $\boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t$. At the end of the round, the learner receives some feedback based on its decision $\boldsymbol{V}_t$ and the loss vector $\boldsymbol{\ell}_t$. The regret of the learner is defined as

$$R_T = \max_{\boldsymbol{v}\in\mathcal{S}} \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{V}_t - \boldsymbol{v}\right)^\mathsf{T}\boldsymbol{\ell}_t\right].$$

Previous work has considered the following feedback schemes in the combinatorial setting:

- The full information scheme where the learner gets to observe $\boldsymbol{\ell}_t$ regardless of the chosen action. The minimax optimal regret of order $m\sqrt{T\log d}$ here is achieved by COMPONENTHEDGE algorithm of [13], while the Follow-the-Perturbed-Leader (FPL) [12, 10] was shown to enjoy a regret of order $m^{3/2}\sqrt{T\log d}$ by [16].
- The semi-bandit scheme where the learner gets to observe the components $\ell_{t,i}$ of the loss vector where $V_{t,i} = 1$, that is, the losses along the components chosen by the learner at time t. As shown by [2], COMPONENTHEDGE achieves a near-optimal $\mathcal{O}(\sqrt{mdT\log d})$ regret guarantee, while [16] show that FPL enjoys a bound of $\mathcal{O}(m\sqrt{dT\log d})$.
- The bandit scheme where the learner only observes its own loss $\boldsymbol{V}_t^\mathsf{T}\boldsymbol{\ell}_t$. There are currently no known efficient algorithms that get close to the minimax regret in this setting—the reader is referred to Audibert et al. [2] for an overview of recent results.

In this section, we define a new feedback scheme situated between the semi-bandit and the full-information schemes. In particular, we assume that the learner gets to observe the losses of some other components not included in its own decision vector $\boldsymbol{V}_t$. Similarly to the model of Alon et al. [1], the relation between the chosen action and the side observations are given by a directed observability G_t (see example in Figure 1). We refer to this feedback scheme as *semi-bandit with side observations*. While our theoretical results stated in the previous section continue to hold in this setting, combinatorial EXP3-IX could rarely be implemented efficiently—we refer to [7, 13] for some positive examples. As one of the main concerns in this paper is computational efficiency, we take a different approach: we propose a variant of FPL that efficiently implements the idea of implicit exploration in combinatorial semi-bandit problems with side observations.

3.1 Implicit exploration by geometric resampling

In each round t, FPL bases its decision on some estimate $\widehat{\boldsymbol{L}}_{t-1} = \sum_{s=1}^{t-1}\hat{\boldsymbol{\ell}}_s$ of the total losses $\boldsymbol{L}_{t-1} = \sum_{s=1}^{t-1}\boldsymbol{\ell}_s$ as follows:

$$\boldsymbol{V}_t = \operatorname*{arg\,min}_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T}\left(\eta_t\widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t\right). \tag{5}$$

Here, $\eta_t > 0$ is a parameter of the algorithm and $\boldsymbol{Z}_t$ is a perturbation vector with components drawn independently from an exponential distribution with unit expectation. The power of FPL lies in that it only requires an oracle that solves the (offline) optimization problem $\min_{\boldsymbol{v}\in\mathcal{S}} \boldsymbol{v}^\mathsf{T}\boldsymbol{\ell}$ and thus

can be used to turn any efficient offline solver into an online optimization algorithm with strong guarantees. To define our algorithm precisely, we need to some further notation. We redefine $\mathcal{F}_{t-1}$ to be $\sigma(\boldsymbol{V}_{t-1}, \dots, \boldsymbol{V}_1)$, $O_{t,i}$ to be the indicator of the observed *component* and let

$$q_{t,i} = \mathbb{E}\left[V_{t,i} \,|\mathcal{F}_{t-1}\right] \qquad \text{and} \qquad o_{t,i} = \mathbb{E}\left[O_{t,i} \,|\mathcal{F}_{t-1}\right].$$

The most crucial point of our algorithm is the construction of our loss estimates. To implement the idea of implicit exploration by optimistic biasing, we apply a modified version of the geometric resampling method of Neu and Bartók [16] constructed as follows: Let $\boldsymbol{O}'_t(1), \boldsymbol{O}'_t(2), \dots$ be independent copies[3] of $\boldsymbol{O}_t$ and let $U_{t,i}$ be geometrically distributed random variables for all $i = [d]$ with parameter γ_t. We let

$$K_{t,i} = \min\left(\left\{k : O'_{t,i}(k) = 1\right\} \cup \{U_{t,i}\}\right) \tag{6}$$

and define our loss-estimate vector $\hat{\boldsymbol{\ell}}_t \in \mathbb{R}^d$ with its i-th element as

$$\hat{\ell}_{t,i} = K_{t,i} O_{t,i} \ell_{t,i}. \tag{7}$$

By definition, we have $\mathbb{E}\left[K_{t,i} \,|\mathcal{F}_{t-1}\right] = 1/(o_{t,i} + (1 - o_{t,i})\gamma_t)$, implying that our loss estimates are *optimistic* in the sense that they lower bound the losses in expectation:

$$\mathbb{E}\left[\hat{\ell}_{t,i} \middle| \mathcal{F}_{t-1}\right] = \frac{o_{t,i}}{o_{t,i} + (1 - o_{t,i})\gamma_t} \ell_{t,i} \le \ell_{t,i}.$$

Here we used the fact that $O_{t,i}$ is independent of $K_{t,i}$ and has expectation $o_{t,i}$ given $\mathcal{F}_{t-1}$. We call this algorithm Follow-the-Perturbed-Leader with Implicit eXploration (FPL-IX, Algorithm 2).

Note that the geometric resampling procedure can be terminated as soon as $K_{t,i}$ becomes well-defined for all i with $O_{t,i} = 1$. As noted by Neu and Bartók [16], this requires generating at most d copies of $\boldsymbol{O}_t$ on expectation. As each of these copies requires one access to the linear optimization oracle over $\mathcal{S}$, we conclude that the expected running time of FPL-IX is at most d times that of the expected running time of the oracle. A high-probability guarantee of the running time can be obtained by observing that $U_{t,i} \le \log\left(\frac{1}{\delta}\right)/\gamma_t$ holds with probability at least $1 - \delta$ and thus we can stop sampling after at most $d \log\left(\frac{d}{\delta}\right)/\gamma_t$ steps with probability at least $1 - \delta$.

3.2 Performance guarantees for FPL-IX

The analysis presented in this section combines some techniques used by Kalai and Vempala [12], Hutter and Poland [11], and Neu and Bartók [16] for analyzing FPL-style learners. Our proofs also heavily rely on some specific properties of the IX loss estimate defined in Equation 7. The most important difference from the analysis presented in Section 2.2 is that now we are not able to use random learning rates as we cannot compute the values corresponding to Q_t efficiently. In fact, these values are observable in the information-theoretic sense, so we could prove bounds similar to Theorem 1 had we had access to infinite computational resources. As our focus in this paper is on computationally efficient algorithms, we choose to pursue a different path. In particular, our learning rates will be tuned according to efficiently computable approximations $\widetilde{\alpha}_t$ of the respective independence numbers α_t that satisfy $\alpha_t/C \le \widetilde{\alpha}_t \le \alpha_t \le d$ for some $C \ge 1$. For the sake of simplicity, we analyze the algorithm in the oblivious adversary model. The following theorem states the performance guarantee for FPL-IX in terms of the learning rates and random variables of the form

$$\widetilde{Q}_t(c) = \sum_{i=1}^{d} \frac{q_{t,i}}{o_{t,i} + c}.$$

Algorithm 2 FPL-IX

Input: Set of actions $\mathcal{S}$,
  parameters $\gamma_t \in (0,1)$, $\eta_t > 0$ for $t \in [T]$.
for $t = 1$ **to** T **do**
  An adversary privately chooses losses $\ell_{t,i}$ for all $i \in [d]$ and generates a graph G_t
  Draw $Z_{t,i} \sim \mathrm{Exp}(1)$ for all $i \in [d]$
  $\boldsymbol{V}_t \leftarrow \arg\min_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\intercal \left(\eta_t \widehat{\boldsymbol{L}}_{t-1} - \boldsymbol{Z}_t\right)$
  Receive loss $\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t$
  Observe graph G_t
  Observe pairs $\{i, \ell_{t,i}\}$ for all i, such that $(j \to i) \in G_t$ and $\boldsymbol{v}(I_t)_j = 1$
  Compute $K_{t,i}$ for all $i \in [d]$ using Eq. (6)
  $\hat{\ell}_{t,i} \leftarrow K_{t,i} O_{t,i} \ell_{t,i}$
end for

[3] Such independent copies can be simply generated by sampling independent copies of $\boldsymbol{V}_t$ using the FPL rule (5) and then computing $\boldsymbol{O}'_t(k)$ using the observability G_t. Notice that this procedure requires no interaction between the learner and the environment, although each sample requires an oracle access.

Theorem 2. *Assume $\gamma_t \le 1/2$ for all t and $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_T$. The regret of* FPL-IX *satisfies*

$$R_T \le \frac{m\left(\log d + 1\right)}{\eta_T} + 4m \sum_{t=1}^{T} \eta_t \mathbb{E}\left[\widetilde{Q}_t\left(\frac{\gamma_t}{1-\gamma_t}\right)\right] + \sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\widetilde{Q}_t(\gamma_t)\right].$$

Proof sketch. As usual for analyzing FPL methods [12, 11, 16], we first define a hypothetical learner that uses a time-independent perturbation vector $\widetilde{\boldsymbol{Z}} \sim \boldsymbol{Z}_1$ and has access to $\hat{\boldsymbol{\ell}}_t$ on top of $\widehat{\boldsymbol{L}}_{t-1}$

$$\widetilde{\boldsymbol{V}}_t = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{S}} \boldsymbol{v}^\intercal \left(\eta_t \widehat{\boldsymbol{L}}_t - \widetilde{\boldsymbol{Z}}\right).$$

Clearly, this learner is infeasible as it uses observations from the future. Also, observe that this learner does not actually interact with the environment and depends on the predictions made by the actual learner only through the loss estimates. By standard arguments, we can prove

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(\widetilde{\boldsymbol{V}}_t - \boldsymbol{v}\right)^\intercal \hat{\boldsymbol{\ell}}_t\right] \le \frac{m\left(\log d + 1\right)}{\eta_T}.$$

Using the techniques of Neu and Bartók [16], we can relate the performance of $\boldsymbol{V}_t$ to that of $\widetilde{\boldsymbol{V}}_t$, which we can further upper bounded after a long and tedious calculation as

$$\mathbb{E}\left[(\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t)^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] \le \eta_t \, \mathbb{E}\left[\left(\widetilde{\boldsymbol{V}}_{t-1}^\intercal \hat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1}\right] \le 4m\eta_t \mathbb{E}\left[\widetilde{Q}_t\left(\frac{\gamma}{1-\gamma}\right) \middle| \mathcal{F}_{t-1}\right].$$

The result follows by observing that $\mathbb{E}\left[\boldsymbol{v}^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] \le \boldsymbol{v}^\intercal \boldsymbol{\ell}_t$ for any fixed $\boldsymbol{v} \in \mathcal{S}$ by the optimistic property of the IX estimate and also from the fact that by the definition of the estimates we infer that

$$\mathbb{E}\left[\widetilde{\boldsymbol{V}}_{t-1}^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] \ge \mathbb{E}\left[\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t \middle| \mathcal{F}_{t-1}\right] - \gamma_t \mathbb{E}\left[\widetilde{Q}_t(\gamma_t)\right]. \qquad \square$$

The next lemma shows a suitable upper bound for the last two terms in the bound of Theorem 2. It follows from observing that $o_{t,i} \ge (1/m) \sum_{j \in \{N_{t,i}^- \cup \{i\}\}} q_{t,j}$ and applying Lemma 1.

Lemma 3. *For all $t \in [T]$ and any $c \in (0,1)$,*

$$\widetilde{Q}_t(c) = \sum_{i=1}^{d} \frac{q_{t,i}}{o_{t,i} + c} \le 2m\alpha_t \log\left(1 + \frac{m\lceil d^2/c \rceil + d}{\alpha_t}\right) + 2m.$$

We are now ready to state the main result of this section, which is obtained by combining Theorem 2, Lemma 3, and Lemma 3.5 of Auer et al. [4] applied to the following upper bound

$$\sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{d + \sum_{s=1}^{t-1} \widetilde{\alpha}_s}} \le \sum_{t=1}^{T} \frac{\alpha_t}{\sqrt{\sum_{s=1}^{t} \alpha_s / C}} \le 2\sqrt{C \sum_{t=1}^{T} \alpha_t} \le 2\sqrt{d + C \sum_{t=1}^{T} \alpha_t}.$$

Corollary 2. *Assume that for all $t \in [T]$, $\alpha_t / C \le \widetilde{\alpha}_t \le \alpha_t \le d$ for some $C > 1$, and assume $md > 4$. Setting $\eta_t = \gamma_t = \sqrt{\left(\log d + 1\right) / \left(m\left(d + \sum_{s=1}^{t-1} \widetilde{\alpha}_s\right)\right)}$, the regret of* FPL-IX *satisfies*

$$R_T \le H m^{3/2} \sqrt{\left(d + C \sum_{t=1}^{T} \alpha_t\right)\left(\log d + 1\right)}, \quad \textit{where } H = \mathcal{O}(\log(mdT)).$$

Conclusion We presented an efficient algorithm for learning with side observations based on implicit exploration. This technique gave rise to multitude of improvements. Remarkably, our algorithms no longer need to know the observation system before choosing the action unlike the method of [1]. Moreover, we extended the partial observability model of [15, 1] to accommodate problems with large and structured action sets and also gave an efficient algorithm for this setting.

Acknowledgements The research presented in this paper was supported by French Ministry of Higher Education and Research, by European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270327 (CompLACS), and by FUI project Hermès.

References

[1] 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*.

[2] Audibert, J. Y., Bubeck, S., and Lugosi, G. (2014). Regret in Online Combinatorial Optimization. *Mathematics of Operations Research*, 39:31–45.

[3] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002a). The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77.

[4] 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.

[5] Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D., Schapire, R., and Warmuth, M. (1997). How to use expert advice. *Journal of the ACM*, 44:427–485.

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

[7] Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78:1404–1422.

[8] Chen, W., Wang, Y., and Yuan, Y. (2013). Combinatorial Multi-Armed Bandit: General Framework and Applications. In *International Conference on Machine Learning*, pages 151–159.

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

[10] Hannan, J. (1957). Approximation to Bayes Risk in Repeated Play. *Contributions to the theory of games*, 3:97–139.

[11] Hutter, M. and Poland, J. (2004). Prediction with Expert Advice by Following the Perturbed Leader for General Weights. In *Algorithmic Learning Theory*, pages 279–293.

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

[13] Koolen, W. M., Warmuth, M. K., and Kivinen, J. (2010). Hedging structured concepts. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 93–105.

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

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

[16] Neu, G. and Bartók, G. (2013). An Efficient Algorithm for Learning with Semi-bandit Feedback. In Jain, S., Munos, R., Stephan, F., and Zeugmann, T., editors, *Algorithmic Learning Theory*, volume 8139 of *Lecture Notes in Computer Science*, pages 234–248. Springer Berlin Heidelberg.

[17] Vovk, V. (1990). Aggregating strategies. In *Proceedings of the third annual workshop on Computational learning theory (COLT)*, pages 371–386.

A Proof of Lemma 1

The proof relies on the following two statements borrowed from Alon et al. [1].

Lemma 4. (cf. Lemma 10 of [1]) *Let G be a directed graph, with $V = \{1, \dots, d\}$. Let d_i^- be the indegree of the node i and $\alpha = \alpha(G)$ be the independence number of G. Then*

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

Lemma 5. (cf. Lemma 12 of [1]) *If $a,\ b \geq 0$ and $a + b \geq B > A > 0$, then*

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

Proof.

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

□

We are now ready to prove Lemma 1. Our proof is obtained as a generalization of the proof of Lemma 13 by Alon et al. [1].

Let $M = \lceil d^2/c \rceil$ and d_i^- be the indegree of node i. We begin by constructing a discretization of the values p_i for all i such that the discretized version of p_i satisfies $\hat{p}_i = k/M$ for some integer k and $\hat{p}_i - 1/M < p_i \leq \hat{p}_i$. By straightforward algebraic manipulations and the fact that $x/(x+a)$ is increasing in x for nonnegative x and a, we obtain the bound

$$\begin{aligned}
\sum_{i=1}^{d} \frac{p_i}{\frac{1}{m}p_i + \frac{1}{m}P_i + c} &= m \sum_{i=1}^{d} \frac{p_i}{p_i + \sum_{j \in N_i^-} p_j + mc} \\
&\leq m \sum_{i=1}^{d} \frac{\hat{p}_i}{\hat{p}_i + \sum_{j \in N_i^-} \hat{p}_j + mc - d_i^-/M} \\
&\leq m \sum_{i=1}^{d} \frac{\hat{p}_i}{\hat{p}_i + \sum_{j \in N_i^-} \hat{p}_j + mc} + m \sum_{i=1}^{d} \frac{d_i^-/M}{mc - d_i^-/M} \\
&\leq m \sum_{i=1}^{d} \frac{M\hat{p}_i}{M\hat{p}_i + \sum_{j \in N_i^-} M\hat{p}_j} + 2m,
\end{aligned}$$

where the second to last inequality holds by Lemma 5 with $a = \hat{p}_i$, $b = \sum_{j \in N_i^-} \hat{p}_j$, $A = d_i^-/M$, and $B = mc$. It remains to find a suitable upper bound for the first sum on the right hand side. To this end, we construct a graph G' from our original graph G, where that we replace each node i of G by a clique C_i with Mp_i nodes. In this expanded graph, we connect all vertices in clique C_i with all vertices in C_j if and only if there is an edge from i to j in original graph G. Note that our new graph G' has the same independence number α as the original graph G. Also observe that the indegree $\hat{d}_k^-$ of a node k in clique C_i is equal to $Mp_i - 1 + \sum_{j \in N_i^-} Mp_j$. Therefore, the remaining term can be rewritten as

$$\sum_{i=1}^{d} \frac{M\hat{p}_i}{M\hat{p}_i + \sum_{j \in N_i^-} M\hat{p}_j} = \sum_{i=1}^{d} \sum_{k \in C_i} \frac{1}{1+\hat{d}_k^-}$$

which in turn can be bounded using Lemma 4 by

$$2\alpha \ln\left(1 + \frac{\sum_{i=1}^{d} M\hat{p}_i}{\alpha}\right) \leq 2\alpha \ln\left(1 + \frac{mM+d}{\alpha}\right).$$

Using this bound we get

$$\sum_{i=1}^{d} \frac{p_i}{\frac{1}{m}p_i + \frac{1}{m}P_i + c} \leq 2m\alpha \ln\left(1 + \frac{mM+d}{\alpha}\right) + 2m$$

as advertised.

B Full proof of Theorem 1

Proof (Theorem 1). We start by introducing some notation. Let

$$\widehat{L}_{t-1,i} = \sum_{s=1}^{t-1} \hat{\ell}_{s,i} \qquad \text{and} \qquad W'_t = \frac{1}{d}\sum_{i=1}^{d} e^{-\eta_{t-1}\widehat{L}_{t-1,i}}.$$

Following the proof of Lemma 1 of Györfi and Ottucsák [9], we track the evolution of $\log W'_{t+1}/W_t$ to control the regret. We have

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

where we used the inequality $\exp(-x) \le 1 - x + x^2/2$ that holds for $x \ge 0$. Using the inequality $\log(1-x) \le -x$ that holds for all x, we get

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

The second term in brackets on the right hand side can be bounded as

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

where we applied 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 prove Equation (4) as

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

Taking *conditional* expectations and summing up both sides over the time, we get

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

The first term in the above inequality is controlled as

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{d}p_{t,i}\hat{\ell}_{t,i}\middle|\mathcal{F}_{t-1}\right] &= \sum_{i=1}^{d}p_{t,i}\ell_{t,i} + \sum_{i=1}^{d}p_{t,i}\ell_{t,i}\left(\frac{o_{t,i}}{o_{t,i}+\gamma_t} - 1\right) \\
&= \sum_{i=1}^{d}p_{t,i}\ell_{t,i} - \sum_{i=1}^{d}p_{t,i}\ell_{t,i}\left(\frac{\gamma_t}{o_{t,i}+\gamma_t}\right) \\
&\ge \sum_{i=1}^{d}p_{t,i}\ell_{t,i} - \gamma_t Q_t,
\end{aligned}$$

while the first one on the right hand side as

$$\mathbb{E}\left[\sum_{i=1}^{d} p_{t,i}(\hat{\ell}_{t,i})^2 \middle| \mathcal{F}_{t-1}\right] = \sum_{i=1}^{d} p_{t,i}\frac{\ell_{t,i}^2}{(o_{t,i}+\gamma_t)^2}o_{t,i} \le \sum_{i=1}^{d} p_{t,i}\frac{\ell_{t,i}^2}{(o_{t,i}+\gamma_t)o_{t,i}}o_{t,i}$$
$$\le \sum_{i=1}^{d} p_{t,i}\frac{1}{(o_{t,i}+\gamma_t)o_{t,i}}o_{t,i} = \sum_{i=1}^{d}\frac{p_{t,i}}{o_{t,i}+\gamma_t} = Q_t.$$

Combining these bounds yields

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

To proceed, we substitute the parameter choice $\eta_t = \gamma_t = \sqrt{(\log d)/(d+\sum_{s=1}^{t-1}Q_s)}$ and use a standard algebraic lemma [4, Lemma 3.5] to get

$$\sum_{t=1}^{T}\sum_{i=1}^{d} p_{t,i}\ell_{t,i} \le 3\sqrt{\left(d+\sum_{t=1}^{T}Q_t\right)\log d} + \sum_{t=1}^{T}\mathbb{E}\left[\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right)\middle|\mathcal{F}_{t-1}\right].$$

Taking expectation on both sides, the second term on the right hand side telescopes into

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\frac{\log W_t}{\eta_t} - \frac{\log W_{t+1}}{\eta_{t+1}}\right)\right] = \mathbb{E}\left[\frac{\log W_1}{\eta_1} - \frac{\log W_{T+1}}{\eta_{T+1}}\right] \le \mathbb{E}\left[-\frac{\log w_{T+1,j}}{\eta_{T+1}}\right]$$
$$= \mathbb{E}\left[\frac{-1}{\eta_{T+1}}\log\left(\frac{1}{d}e^{-\eta_{T+1}\hat{L}_{T,j}}\right)\right] = \mathbb{E}\left[\frac{\log d}{\eta_{T+1}}\right] + \mathbb{E}\left[\hat{L}_{T,j}\right],$$

for any $j \in [d]$, where we used that $W_{T+1} \ge w_{T+1,j}$ and $W_1 = 1$ since $w_{1,i} = 1/d$ by definition for all $i \in [d]$. Substituting η_{T+1}, we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{d} p_{t,i}\ell_{t,i}\right] \le 3\mathbb{E}\left[\sqrt{\log d\left(d+\sum_{t=1}^{T}Q_t\right)}\right] + \mathbb{E}\left[\sqrt{\log d\left(d+\sum_{t=1}^{T}Q_t\right)}\right] + \mathbb{E}\left[\hat{L}_{T,j}\right],$$

which together with the fact that our estimates $\hat{L}_{T,j}$ are optimistic yields the theorem.

□

C Full proof of Theorem 2

We begin with a statement that concerns the performance of the imaginary learner that predicts $\widetilde{\boldsymbol{V}}_t$ in round t.

Lemma 6. *Assume $\eta_1 \ge \eta_2 \ge \cdots \ge \eta_T$. For any sequence of loss estimates, the expected regret of the hypothetical learner against any fixed action $\boldsymbol{v} \in \mathcal{S}$ satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\widetilde{\boldsymbol{V}}_t - \boldsymbol{v}\right)^{\mathsf{T}}\hat{\boldsymbol{\ell}}_t\right] \le \frac{m\left(\log d + 1\right)}{\eta_T}.$$

Proof. For simplicity, define $\beta_t = 1/\eta_t$ for $t \ge 1$ and $\beta_0 = 0$. We start by applying the classical follow-the-leader/be-the-leader lemma (see, e.g., [6, Lemma 3.1]) to the loss sequence defined as $(\hat{\boldsymbol{\ell}}_1 - \widetilde{\boldsymbol{Z}}\beta_1, \hat{\boldsymbol{\ell}}_2 - \widetilde{\boldsymbol{Z}}(\beta_2-\beta_1), \ldots, \hat{\boldsymbol{\ell}}_T - \widetilde{\boldsymbol{Z}}(\beta_T - \beta_{T-1}))$ to obtain

$$\sum_{t=1}^{T}\widetilde{\boldsymbol{V}}_t^{\mathsf{T}}\left(\hat{\boldsymbol{\ell}}_t - \widetilde{\boldsymbol{Z}}\left(\beta_t - \beta_{t-1}\right)\right) \le \widetilde{\boldsymbol{V}}_T^{\mathsf{T}}\left(\widehat{\boldsymbol{L}}_T - \widetilde{\boldsymbol{Z}}\beta_T\right) \le \boldsymbol{v}^{\mathsf{T}}\left(\widehat{\boldsymbol{L}}_T - \widetilde{\boldsymbol{Z}}\beta_T\right).$$

After reordering and observing that $-\boldsymbol{v}^\intercal \widetilde{\boldsymbol{Z}} \le 0$, we get

$$\begin{aligned}\sum_{t=1}^{T} \left(\widetilde{\boldsymbol{V}}_t - \boldsymbol{v}\right)^\intercal \hat{\boldsymbol{\ell}}_t &\le \sum_{t=1}^{T} (\beta_t - \beta_{t-1}) \widetilde{\boldsymbol{V}}_t^\intercal \widetilde{\boldsymbol{Z}} \\ &\le \left\|\widetilde{\boldsymbol{V}}_t\right\|_1 \left\|\widetilde{\boldsymbol{Z}}\right\|_\infty \sum_{t=1}^{T} (\beta_t - \beta_{t-1}) = \left\|\widetilde{\boldsymbol{V}}_t\right\|_1 \left\|\widetilde{\boldsymbol{Z}}\right\|_\infty \beta_T.\end{aligned}$$

The result follows from using our uniform upper bound on $\|\boldsymbol{v}\|_1$ for all $\boldsymbol{v}$ and the well-known bound $\mathbb{E}\left[\left\|\widetilde{\boldsymbol{Z}}\right\|_\infty\right] \le \log d + 1$. □

The following result can be extracted from the proof of Theorem 1 of Neu and Bartók [16].

Lemma 7. *For any sequence of nonnegative loss estimates,*

$$\mathbb{E}\left[(\widetilde{\boldsymbol{V}}_{t-1} - \widetilde{\boldsymbol{V}}_t)^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_t\right] \le \eta_t \, \mathbb{E}\left[\left(\widetilde{\boldsymbol{V}}_{t-1}^\intercal \hat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_t\right].$$

Using these two lemmas, we can prove the following lemma that upper bounds the total expected regret of FPL-IX in terms of the sum of the variables

$$\widetilde{Q}_t(c) = \sum_{i=1}^{d} \frac{q_{t,i}}{o_{t,i} + c}.$$

Lemma 8. *Assume that $\gamma_t \le 1/2$ for all t. Then,*

$$\sum_{t=1}^{T} \mathbb{E}\left[\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t \middle| \mathcal{F}_{t-1}\right] \le \sum_{t=1}^{T} \mathbb{E}\left[\widetilde{\boldsymbol{V}}_t^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] + 4m \sum_{t=1}^{T} \eta_t \mathbb{E}\left[\widetilde{Q}_t\left(\frac{\gamma_t}{1-\gamma_t}\right)\right] + \sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\widetilde{Q}_t(\gamma_t)\right].$$

Proof. First, note that Lemma 7 implies

$$\mathbb{E}\left[(\widetilde{\boldsymbol{V}}_{t-1} - \widetilde{\boldsymbol{V}}_t)^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] \le \eta_t \, \mathbb{E}\left[\left(\widetilde{\boldsymbol{V}}_{t-1}^\intercal \hat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1}\right]$$

by the tower rule of expectation. We start by observing that

$$\begin{aligned}\mathbb{E}\left[\widetilde{\boldsymbol{V}}_{t-1}^\intercal \hat{\boldsymbol{\ell}}_t \middle| \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\sum_{i=1}^{d} q_{t,i} \hat{\ell}_{t,i} \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\sum_{i=1}^{d} q_{t,i} \frac{\ell_{t,i}}{o_{t,i} + (1-o_{t,i})\gamma_t} O_{t,i} \middle| \mathcal{F}_{t-1}\right] \\ &\ge \mathbb{E}\left[\sum_{i=1}^{d} q_{t,i} \frac{\ell_{t,i}}{o_{t,i} + (1-o_{t,i})\gamma_t} \left(O_{t,i} + (1-o_{t,i})\gamma_t\right) - \gamma_t \sum_{i=1}^{d} q_{t,i} \frac{1-o_{t,i}}{o_{t,i} + (1-o_{t,i})\gamma_t} \middle| \mathcal{F}_{t-1}\right] \\ &\ge \sum_{i=1}^{d} q_{t,i} \ell_{t,i} - \gamma_t \mathbb{E}\left[\sum_{i=1}^{d} \frac{q_{t,i}(1-o_{t,i})}{o_{t,i} + (1-o_{t,i})\gamma_t} \middle| \mathcal{F}_{t-1}\right] \\ &\ge \sum_{i=1}^{d} q_{t,i} \ell_{t,i} - \gamma_t \mathbb{E}\left[\sum_{i=1}^{d} \frac{q_{t,i}}{o_{t,i} + \gamma_t} \middle| \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\boldsymbol{V}_t^\intercal \boldsymbol{\ell}_t \middle| \mathcal{F}_{t-1}\right] - \gamma_t \widetilde{Q}_t(\gamma_t).\end{aligned}$$

To simplify some notation, let us fix a time t and define $\boldsymbol{V} = \widetilde{\boldsymbol{V}}_{t-1}$. We deduce that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\widetilde{\boldsymbol{V}}_{t-1}^{\intercal}\hat{\boldsymbol{\ell}}_t\right)^2 \middle| \mathcal{F}_{t-1}\right] \\
&\quad= \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\left(V_j\hat{\ell}_{t,j}\right)\left(V_k\hat{\ell}_{t,k}\right)\middle|\mathcal{F}_{t-1}\right] \\
&\quad= \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\left(V_jK_{t,j}O_{t,j}\ell_{t,j}\right)\left(V_kK_{t,k}O_{t,k}\ell_{t,k}\right)\middle|\mathcal{F}_{t-1}\right] && \text{(def. of } \hat{\boldsymbol{\ell}}_t\text{)} \\
&\quad\le \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{K_{t,j}^2+K_{t,k}^2}{2}\left(V_jO_{t,j}\ell_{t,j}\right)\left(V_kO_{t,k}\ell_{t,k}\right)\middle|\mathcal{F}_{t-1}\right] && (2K_{t,j}K_{t,k}\le K_{t,j}^2+K_{t,k}^2) \\
&\quad\le \mathbb{E}\left[\sum_{j=1}^{d}\sum_{k=1}^{d}K_{t,j}^2\left(V_jO_{t,j}\ell_{t,j}\right)\left(V_kO_{t,k}\ell_{t,k}\right)\middle|\mathcal{F}_{t-1}\right] && \text{(symmetry of } j \text{ and } k\text{)} \\
&\quad\le 2\mathbb{E}\left[\sum_{j=1}^{d}\frac{1}{(o_{t,j}+(1-o_{t,j})\gamma_t)^2}\left(V_jO_{t,j}\ell_{t,j}\right)\sum_{k=1}^{d}V_k\ell_{t,k}\middle|\mathcal{F}_{t-1}\right] && \text{(def. of } K_{t,j} \text{ and } O_{t,k}\le 1\text{)} \\
&\quad\le 2m\mathbb{E}\left[\sum_{j=1}^{d}\frac{V_j\ell_{t,j}}{o_{t,j}+(1-o_{t,j})\gamma_t}\middle|\mathcal{F}_{t-1}\right] \\
&\quad\le 2m\sum_{j=1}^{d}\frac{q_{t,j}}{o_{t,j}+(1-o_{t,j})\gamma_t} = \frac{2m}{1-\gamma_t}\sum_{j=1}^{d}\frac{q_{t,j}}{o_{t,j}+\gamma_t/(1-\gamma_t)} \\
&\quad= \frac{2m}{1-\gamma_t}\widetilde{Q}_t\left(\frac{\gamma_t}{1-\gamma_t}\right) \le 4m\widetilde{Q}_t\left(\frac{\gamma_t}{1-\gamma_t}\right),
\end{aligned}
$$

where we used our assumption on γ_t in the last line. The first statement follows from combining the above terms with Lemma 7 and using $\mathbb{E}\left[\boldsymbol{v}^{\intercal}\hat{\boldsymbol{\ell}}_t\middle|\mathcal{F}_{t-1}\right] \le \boldsymbol{v}^{\intercal}\boldsymbol{\ell}_t$ by the optimistic property of the loss estimates $\hat{\boldsymbol{\ell}}_t$. □

D Proof of Lemma 3

We start with proving the lower bound on

$$o_{t,i} \ge \frac{1}{m}\sum_{j\in\{N_{t,i}^{-}\cup\{i\}\}} q_{t,j}.$$

We prove this by first proving $O_{t,i} \ge (1/m)\sum_{j\in\{N_{t,i}^{-}\cup\{i\}\}} V_{t,j}$ as follows: First, assume that $O_{t,j} = 0$, in which case the bound trivially holds, since both sides evaluate to zero by definition of $O_{t,i}$. Otherwise, we have

$$\frac{1}{m}\sum_{j\in\{N_{t,i}^{-}\cup\{i\}\}} V_{t,j} \le \frac{1}{m}\sum_{j=1}^{d} V_{t,j} \le 1 = O_{t,i},$$

where we used $\sum_{j\in V} V_{t,j} \le m$ in the last inequality. Taking expectations gives the desired lower bound on $o_{t,i}$. Then we get

$$\sum_{i=1}^{d}\frac{q_{t,i}}{o_{t,i}+c} \le \sum_{i=1}^{d}\frac{q_{t,i}}{q_{t,i}+\sum_{j\in N_{t,i}^{-}} q_{t,j}+c}.$$

The proof is completed using Lemma 1.