A Near-Optimal Best-of-Both-Worlds Algorithm for Online Learning with Feedback Graphs

Chloé Rouyer
Dept. of Computer Science
University of Copenhagen, Denmark
chloe@di.ku.dk

Dirk van der Hoeven
Dept. of Computer Science
Università degli Studi di Milano, Italy
dirk@dirkvanderhoeven.com

Nicolò Cesa-Bianchi
Dept. of Computer Science
Università degli Studi di Milano, Italy
nicolo.cesa-bianchi@unimi.it

Yevgeny Seldin
Dept. of Computer Science
University of Copenhagen, Denmark
seldin@di.ku.dk

Abstract

We consider online learning with feedback graphs, a sequential decision-making framework where the learner's feedback is determined by a directed graph over the action set. We present a computationally efficient algorithm for learning in this framework that simultaneously achieves near-optimal regret bounds in both stochastic and adversarial environments. The bound against oblivious adversaries is $\tilde{O}(\sqrt{\alpha T})$, where T is the time horizon and α is the independence number of the feedback graph. The bound against stochastic environments is $O\big((\ln T)^2 \max_{S\in\mathcal{I}(G)} \sum_{i\in S} \Delta_i^{-1}\big)$ where $\mathcal{I}(G)$ is the family of all independent sets in a suitably defined undirected version of the graph and Δ_i are the suboptimality gaps. The algorithm combines ideas from the EXP3++ algorithm for stochastic and adversarial bandits and the EXP3.G algorithm for feedback graphs with a novel exploration scheme. The scheme, which exploits the structure of the graph to reduce exploration, is key to obtain best-of-both-worlds guarantees with feedback graphs. We also extend our algorithm and results to a setting where the feedback graphs are allowed to change over time.

1 Introduction

Online learning is a general framework for studying sequential decision-making in unknown environments (see, for example, (Cesa-Bianchi and Lugosi, 2006; Bubeck and Cesa-Bianchi, 2012; Orabona, 2019)). We consider a setting where, at each round, the player chooses an action (a.k.a. arm) from a fixed set of K actions and incurs the loss associated with the chosen action. The performance of the learner is quantified in terms of regret, which is the difference between the total loss incurred by the learner over the duration of the game, and the smallest cumulative loss obtained by a player that would only ever play the same action throughout the game.

The smallest achievable regret is determined by a number of parameters. One of these parameters is the amount of feedback that the learner receives at each round. There is a whole spectrum of problems, characterized by the amount of feedback received by the learner. At the one extreme of this spectrum is the bandit setting, where the learner only observes the loss of the action taken. At the other extreme is the full information setting, where the learner observes the full loss vector at the end of each round, irrespective of the action played.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

There are two common ways to interpolate between full information and bandit feedback. One is to allow the learner to make a limited number of additional observations without restricting how the additional observations are selected. Then no additional observations correspond to the bandit setting and $K-1$ additional observations correspond to the full information setting. This way of interpolation was proposed by Seldin et al. (2014) in two variants, "prediction with limited advice" and "multiarmed bandits with paid observations". It was also studied by Thune and Seldin (2018).

The second way of interpolation, which we focus on in this paper, is via feedback graphs (Alon et al., 2017). In this setting observations of the learner are governed by a feedback graph on the actions. When an action is played, the learner observes the losses of all of its neighbors in the feedback graph. A complete graph corresponds to the full information setting, whereas a graph containing only self-loops corresponds to the bandit setting. This setting has multiple variants, depending on whether the graph is directed or undirected, observed or unobserved, static or dynamic.

Another important parameter characterizing online learning problems is the type of environment. The two primary types that we focus on are stochastic and adversarial environments. In stochastic environments each action is associated with a fixed, but unknown distribution, and in each round the loss of each action is sampled independently from the corresponding distribution. In adversarial environments the loss sequence is chosen arbitrarily. We consider oblivious adversarial environments, where the loss sequences are chosen independently of the actions taken by the learner.

For a long time stochastic and adversarial environments where studied separately, but in practice the exact nature of environment is rarely known. In recent years this has led to a growing interest in "best-of-both-worlds" algorithms that are robust against adversarial loss sequences and, at the same time, provide tighter regret guarantees in the stochastic regime. Most work has focused on the bandit setting (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018), where the Tsallis-INF algorithm proposed by Zimmert and Seldin (2019, 2021) was shown to achieve the optimal regret rates in both stochastic and adversarial regimes, as well as a number of intermediate regimes. The analysis was further improved by Masoudian and Seldin (2021) and Ito (2021). In the full information setting Mourtada and Gaïffas (2019) have shown that the well-known Hedge algorithm originally designed for the adversarial setting (Littlestone and Warmuth, 1994) also achieves the optimal stochastic regret. Best-of-both-world results also spilled over to other domains, including additional approaches to full information games and online convex optimization (Koolen et al., 2016; Van Erven et al., 2021; Negrea et al., 2021), decoupled exploration and exploitation (Rouyer and Seldin, 2020), combinatorial bandits (Zimmert et al., 2019), bandits with switching costs (Rouyer et al., 2021), MDPs (Jin and Luo, 2020; Jin et al., 2021), and linear bandits (Lee et al., 2021).

In the context of online learning with feedback graphs the only best-of-both-worlds result known to us is by Erez and Koren (2021) for undirected graphs. They present an intricate algorithm based on the Follow The Regularized Leader (FTRL) framework with a regularization function that is a product of the Tsallis and Shannon entropies. The algorithm simultaneously enjoys an $O\big(\sqrt{\chi T}\big(\ln(KT)\big)^2\big)$ pseudo-regret bound in the adversarial regime and an $O\big(\big(\ln(KT)\big)^4 \sum_k \frac{\ln T}{\Delta_k}\big)$ pseudo-regret bound in the stochastic regime, where T is the number of prediction rounds, χ is the clique covering number of the undirected feedback graph, and the summation in the second bound is on the smallest non-zero gap within each clique.

It is tempting to apply an FTRL-based algorithm with Tsallis entropy regularization to online learning with feedback graphs, since Tsallis entropy with power $a = 1/2$ leads to the optimal Tsallis-INF algorithm for the bandit setting (Zimmert and Seldin, 2021) and Tsallis entropy with power $a = 1$ leads to the Hedge algorithm, which is optimal in the full information setting. However, as also noted by Erez and Koren (2021), extension of the analysis to online learning with feedback graphs when the power $a \in (1/2, 1)$ is not straightforward and, so far, there was no success in this direction. Furthermore, at the moment it is unclear whether it is possible to derive bounds that take further advantage of the graph structure and depend on the independence number of the graph when $a < 1$.

Our contribution. We significantly extend and improve on the bounds of Erez and Koren (2021). Our results hold for directed graphs (with self-loops), depend on the independence number of the graph, have a better dependence on T in the stochastic regime, and extend to time-varying feedback graphs. Our approach takes advantage of the common structure shared by two exponential weights algorithms: EXP3.G (Alon et al., 2015) and EXP3++ (Seldin and Slivkins, 2014; Seldin and Lugosi,

2017), to obtain near-optimal best-of-both worlds guarantees. By using similar ideas as in the proof of the regret bound of EXP3.G, the proposed algorithm adapts to the independence number of the graph. We derive an $\min\left\{O(\sqrt{\tilde{\alpha}T\ln K}), O\left(\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T}\right)\right\}$ pseudo-regret bound against adversarial sequences of losses, where α is the independence number of the graph and $\tilde{\alpha}$ is its strong independence number, which is a graph dependent quantity smaller than the clique covering number. For undirected graphs, independence number and strong independence number are equal and the result matches the best known lower bound $\Omega(\sqrt{\alpha T})$ within logarithmic factors (Alon et al., 2017). In the stochastic setting we use the idea of injected exploration from EXP3++ to estimate the suboptimality gaps of each arm. By introducing a novel dynamic exploration set and an appropriate exploration rate, we derive an almost optimal regret bound in the stochastic setting. Along the way, we also improve the regret bound of EXP3++ in the stochastic bandit setting. Our exploration set is constructed by sorting the arms by ascending gap estimates, and then adding a new arm to the exploration set if the arm cannot be observed by playing another arm previously added to the set. If we play each arm i in the exploration set at a rate $1/\hat{\Delta}_i^2$, where $\hat{\Delta}_i$ is the gap estimate, then all arms j in the graph are observed with probability at least $1/\hat{\Delta}_j^2$.

To present our main result we introduce some notations. Let $G = (V, E)$ be a directed feedback graph with independence number α (where the independence number is computed on G ignoring edge directions). We define a strongly independent set on G as an independent set on the subgraph $G' = (V, E')$, where $(i, j) \in E'$ if and only if $(i, j) \in E$ and $(j, i) \in E$. We use $\tilde{\alpha}$ to denote the strong independence number of G, and $\mathcal{I}(G)$ to denote the collection of all the strongly independent sets in G. We note that $\alpha = \tilde{\alpha}$ for undirected graphs and $\alpha \leq \tilde{\alpha}$ for directed graphs. Now we can present an informal statement of our main result.

Theorem 1 (Informal)**.** *Given a directed feedback graph $G = (V, E)$ with independence number α and strong independence number $\tilde{\alpha}$, there exists an algorithm (Algorithm 1) whose pseudo-regret can simultaneously be bounded by $\min\left\{O(\sqrt{\tilde{\alpha}T\ln K}), O\left(\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T}\right)\right\}$ against adversarial loss sequences and by $O\big((\ln T)^2 \max_{S\in\mathcal{I}(G)}\sum_{i\in S}\Delta_i^{-1}\big)$ against stochastic loss sequences.*

We emphasize that Algorithm 1 requires neither prior knowledge of the type of the environment (adversarial or stochastic), nor the time horizon.

1.1 Additional Related Work

The study of bandits with feedback graphs was initiated by Mannor and Shamir (2011) in the adversarial regime and by Caron et al. (2012) in the stochastic regime. In the adversarial regime, the optimal regret rates for arbitrary directed graphs were characterized (up to log factors) by Alon et al. (2015). They showed an $\Omega(T)$ lower bound for graphs that have non-observable nodes (i.e., with an empty in-neighborhood). For graphs with observable nodes, they derived pseudo-regret bounds of order $O\big(\sqrt{\alpha T}\log(KT)\big)$ when all nodes are strongly observable (i.e., they have a self-loop or their in-neighborhood contains all of the other nodes) and of order $O\big((\delta\ln K)^{1/3}T^{2/3}\big)$ for weakly observable graphs (where each non-strongly observable node is in the out-neighborhood of some observable node). Here α is the independence number of the graph and δ is the dominating number of the weakly observable portion of the graph. Van der Hoeven et al. (2021) derived results for the multiclass classification with feedback graphs setting. The setting where the graph can adversarially change over time has been studied by Alon et al. (2017) in the case of directed graphs with self-loops. For learners that are allowed to observe the feedback graph at the beginning of each round, they achieved a bound of $O\big(\ln K\sqrt{\ln(KT)\sum_{t=1}^T\alpha_t}\big)$, where α_t is the independence number of the graph at time t. For the case of undirected graphs, they proved a refined bound $O\big(\sqrt{\ln K\sum_{t=1}^T\alpha_t}\big)$ that holds even when the learner can only observe the graph at the end of each round. Note that, as shown by Cohen et al. (2016), in order to take advantage of the graph structure in the adversarial regime, it not sufficient to observe the neighborhood of the played action at the end of each round.

In the stochastic regime, Buccapatnam et al. (2014, 2017) considered a fixed, possibly directed, feedback graph. They derived an asymptotic lower bound showing that the regret scales as $\Omega(c^*\ln T)$, where c^*—which is related to the domination number of the graph—is the solution to a linear program expressing the trade-off between the loss incurred from playing an action and the observations that

can be gathered from playing that action. They proposed an algorithm that can achieve a matching $O\big(c^* \ln T + Kd\big)$ pseudo-regret bound, where d is the maximum degree in the feedback graph. In the case of graphs that change over time, Cohen et al. (2016) derived an $O(\sum_{i \in S}(\ln T)/\Delta_i)$ bound, where S is a set containing an order of α arms (up to log factors), and α is an upper bound on the independence number of the graphs in the sequence. They achieved this result without requiring to observe the graphs fully, and having only access to the neighbourhood of the arm played at the end of the round. Both of these approaches are based on arm elimination algorithms, which—by construction—are not suitable for best-of-both worlds guarantees. The proof strategy of Cohen et al. (2016) was adapted by Lykouris et al. (2020) to provide refined bounds for both UCB-N and Thompson Sampling-N, which are variants of UCB1 (Auer et al., 2002) and Thompson Sampling (Thompson, 1933). In both cases, Lykouris et al. (2020) considered undirected feedback graphs and obtained pseudo-regret bounds that scale as $O\big(\max_{\mathrm{Ind} \in \mathcal{I}(G)} \sum_{i \in \mathrm{Ind}} \ln(KT)(\ln T)/\Delta_i\big)$, where $\mathcal{I}(G)$ is the collection of all the independence sets of the graph.

Concurrently to our work, several other papers in online learning with feedback graphs have appeared. Ito et al. (2022) derive an algorithm with nearly optimal regret bounds in both the stochastic and adversarial setting. While their results are more general than ours (they do not require self-loops in the feedback graph), their regret bounds in the stochastic regime are worse that ours, of order $\frac{\ln(T)^3}{\Delta_{\min}}$, where $\Delta_{\min}$ is the minimum suboptimality gap. Similarly to Erez and Koren (2021), the algorithm of Ito et al. (2022) is based on the FTRL framework. They use the entropic regularization, which makes their algorithm equivalent to EXP3 (Auer et al., 2002). Moreover, Ito et al. (2022) rely on the self-bounding technique of Zimmert et al. (2019) together with an intricate tuning to simultaneously obtain regret bounds in the stochastic regime and the adversarial regime, as well as in intermediate ones. In the stochastic regime, Marinov et al. (2022) provide an improved characterization of the difficulty of online learning with feedback graphs in both finite-time and asymptotic cases. Finally, Esposito et al. (2022) study the more general model of stochastic feedback graphs.

2 Problem Setting and Definitions

Problem Setting We consider a sequential decision-making game, where in each round $t = 1, 2, \ldots$, the learner repeatedly plays an action $I_t \in V$, where $|V| = K$, receives a feedback based on a feedback graph $G = (V, E)$, and suffers a loss ℓ_{t,I_t}. We consider directed feedback graphs with self-loops, meaning that $(i, i) \in E$ for each vertex $i \in V$. The feedback received by the learner at the end of round t is $\big\{(i, \ell_{t,i}) \,:\, i \in N^{\mathrm{out}}(I_t)\big\}$, where $N^{\mathrm{out}}(i) = \{j \in V : (i, j) \in E\}$ is the out-neighbourhood of i. Similarly, we define $N^{\mathrm{in}}(i) = \{j \in V : (j, i) \in E\}$ to be the in-neighborhood of i. For each arm $i \in V$, $\ell_{t,i} \in [0, 1]$ for $t \geq 1$. In the adversarial regime the losses are generated arbitrarily by an oblivious adversary. In the stochastic regime they are independently drawn from a fixed but unknown distribution with expectation $\mathbb{E}[\ell_{1,i}]$. The performance of the learner is measured in terms of the pseudo-regret:

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_{i \in V} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right].$$

In the stochastic regime, we define the best arm i^* as the arm with the smallest expected loss, i.e. $i^* = \mathrm{argmin}_{i \in V} \mathbb{E}[\ell_{1,i}]$. The pseudo-regret can then be expressed in terms of the suboptimality gaps $\Delta_i = \mathbb{E}[\ell_{1,i} - \ell_{1,i^*}]$,

$$\mathcal{R}_T = \sum_{t=1}^{T} \sum_{i \in V} \mathbb{E}\left[p_{t,i}\right] \Delta_i, \tag{1}$$

where $p_{t,i}$ is the probability that the learner plays action i at round t. We define the smallest suboptimality gap $\Delta_{\min} = \min_{i:\Delta_i > 0} \{\Delta_i\}$, and for all i, we define $\bar{\Delta}_i = \max\{\Delta_{\min}, \Delta_i\}$, so that $\bar{\Delta}_{i^*} = \Delta_{\min}$. We use $\mathbb{E}_t$ to express expectation conditioned on all randomness up to round t.

Properties of Graphs Recall that a dominating set in G is a subset $D \subseteq V$, such that for all $i \in V$ there exists $j \in D$, such that $(j, i) \in E$. An independent set in G is a subset $S \subseteq V$, such that for all $i, j \in S$, $(i, j) \notin E$ and $(j, i) \notin E$. We define the independence number $\alpha(G)$ as the size of the largest independent set in the graph G. For clarity, we restate below here the definition of the strong independence number which was already mentioned in the introduction.

Algorithm 1: EXP3.G++

Input: Feedback graph $G = (V, E)$,
Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$; exploration rates $\varepsilon_{t,i}$ for $i \in V$, see Equation (2)
Initialize: $\tilde{L}_0 = \mathbf{0}_K$, $\hat{L}_0 = \mathbf{0}_K$ and $O_0 = \mathbf{0}_K$. Play each arm once to initialize $\hat{L}$ and O
for $t = K+1, K+2, \ldots$ **do**

$$\forall i \in V : \mathrm{UCB}_{t,i} = \min\left\{1, \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} + \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}}\right\}$$

$$\forall i \in V : \mathrm{LCB}_{t,i} = \max\left\{0, \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} - \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}}\right\}$$

$\forall i \in V : \hat{\Delta}_{t,i} = \max\{0, \mathrm{LCB}_{t,i} - \min_j \mathrm{UCB}_{t,j}\}$
$\forall i \in V$: update $\varepsilon_{t,i}$ based on the gap estimates $\hat{\Delta}_t$

$$\forall i \in V : q_{t,i} = \frac{\exp(-\eta_t \tilde{L}_{t,i})}{\sum_{i \in V} \exp(-\eta_t \tilde{L}_{t,j})},\ p_{t,i} = \left(1 - \sum_{j \in V} \varepsilon_{t,j}\right) q_{t,i} + \varepsilon_{t,i}$$

Sample $I_t \sim p_t$ and play it
Observe $\{(j, \ell_{t,j}) : j \in N^{\mathrm{out}}(I_t)\}$ and suffer ℓ_{t,I_t}.

$$\forall i \in V : \tilde{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\left[i \in N^{\mathrm{out}}(I_t)\right]}{P_{t,i}}, \text{ where } P_{t,i} = \textstyle\sum_{j \in N^{\mathrm{in}}(i)} p_{t,j}$$

$\forall\, i \in V : \quad \tilde{L}_{t,i} = \tilde{L}_{t-1,i} + \tilde{\ell}_{t,i}$

$$\forall i \in V : \hat{L}_{t,i} = \hat{L}_{t-1,i} + \ell_{t,i} \mathbb{1}\left[i \in N^{\mathrm{out}}(I_t)\right] \text{ and } O_{t,i} = O_{t-1,i} + \mathbb{1}\left[i \in N^{\mathrm{out}}(I_t)\right]$$

end for

Definition 1. *Let $G = (V, E)$ be a directed graph. We define a strongly independent set on G as an independent set on the subgraph $G' = (V, E')$, where $(i, j) \in E'$ if and only if $(i, j) \in E$ and $(j, i) \in E$. Furthermore, we define $\tilde{\alpha}(G)$ as the independence number of the subgraph G'.*

We use $\mathcal{I}(G)$ to denote a collection of all the strongly independent sets in G. We note that $\alpha = \tilde{\alpha}$ for undirected graphs and $\alpha \leq \tilde{\alpha}$ for directed graphs.

3 Algorithm

We present the EXP3.G++ algorithm (Algorithm 1), which is a combination of the EXP3.G algorithm of Alon et al. (2015) and the EXP3++ algorithm of Seldin and Lugosi (2017) with a novel exploration scheme described in Algorithm 2. This scheme ensures that the additional feedback the learner obtains (relative to the bandit setting) is used nearly optimally.

To understand the motivation behind the novel exploration scheme, note that in the stochastic setting EXP3.G++ needs to ensure that the loss of each arm is observed sufficiently often. However, if we would play each arm too often, the regret would scale with the number of arms, rather than with the independence number or some other graph-theoretic quantity. To avoid that, we exploit the central property of feedback graphs: since we can gather information on certain arms by playing adjacent arms in the graph, we can restrict exploration to a subset of nodes and yet obtain sufficient information on *all* the arms. We exploit this observation, to design a strategy for selecting an exploration set S_t at each round t. S_t is defined in terms of estimated suboptimality gaps $\hat{\Delta}_{t,i}$, which are maintained by EXP3.G++. Crucially, the exploration set ensures that, with high probability, the empirical gaps are reliable estimates of the true suboptimality gaps Δ_i. In turn, this ensures that we observe the loss of each arm sufficiently often.

The construction of the exploration set S_t is detailed in Algorithm 2, which is used by EXP3.G++ to update the exploration rates $\varepsilon_{t,i}$ according to Equation (2). Algorithm 2 starts by sorting the arms according to their gap estimates in ascending order. The exploration set is then greedily constructed by sequentially selecting the next arm with the smallest $\hat{\Delta}_{t,i}$, and discarding all the arms in the out-neighborhood of that arm. The exploration set can be constructed in $O(K^3)$ time, but note that

Algorithm 2: Exploration Set Construction

Input: K arms with associated gaps: $\Delta_1, \Delta_2, \ldots$
Initialize: Exploration set $S = \emptyset$.
Let Λ be the list of arms sorted in ascending order of their associated gaps.
for $i \in \Lambda$ **do**
  Add i to S
  for $j \in N^{\text{out}}(i)$ **do**
    remove j from Λ
  end for
end for
Output: S

we only need to recompute it only when the order of the estimated suboptimality gaps changes. The exploration set S_t has several useful properties, as shown in Proposition 1 below.

Proposition 1. *Let $G = (V, E)$ be a directed feedback graph on K arms with self-loops, and let $\hat{\Delta}_1, \ldots, \hat{\Delta}_K$ be a sequence of suboptimality gaps estimates. Let S be the exploration set constructed by Algorithm 2 based on the sequence of suboptimality gaps. Then S is a dominating set of G with the following property: for all $i \in V$ there exists $j \in S$, such that $i \in N^{\text{out}}(j)$ and $\hat{\Delta}_j \leq \hat{\Delta}_i$. Furthermore, S is also a strongly independent set of G.*

Proof. Let S be the output of Algorithm 2. Since G contains self-loops, if $i \in S$, then $i \in N^{\text{out}}(i)$ and $\hat{\Delta}_i \leq \hat{\Delta}_i$. If $i \notin S$, then i was removed from Λ because $i \in N^{\text{out}}(j)$ for some j that, in a previous iteration, was added to S. Since j was considered before i, we must have $\hat{\Delta}_j \leq \hat{\Delta}_i$. Now, for all $i, j \in S$, we know by construction that $j \notin N^{\text{out}}(i)$. Thus (i, j) is not a directed edge in G, and so S is a strongly independent set in G. □

We define the exploration rates at round t in terms of the exploration set S_t, which is constructed using the aforementioned procedure. For all arms i in V,

$$\varepsilon_{t,i} = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}, \xi_{t,i}\right\}, \tag{2}$$

for some constant $\lambda \in [1, K]$ and where $\xi_{t,i}$ depends on whether $i \in S_t$ or not:

$$\xi_{t,i} = \begin{cases} (\beta \ln t)/(t\hat{\Delta}_{t,i}^2), & \text{if } i \in S_t, \\ 4/t^2, & \text{otherwise,} \end{cases} \tag{3}$$

where $\beta > 0$ is a constant. The role of $\xi_{t,i}$ changes depending on whether we are in an adversarial or stochastic environment. In an adversarial environment, we use $4/t^2 \leq \xi_{t,i}$ to ensure that we sample each arm with a small positive probability, which is essential to bound the second-order term in the regret bound in terms of the independence number. Note that $\varepsilon_{t,i} \leq \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}$, so choosing $\lambda = \alpha$ ensures that the cost of exploration is bounded by $\tilde{O}(\sqrt{\alpha T})$. In the stochastic environment, the construction of the exploration set and the choice of $\xi_{t,i}$ ensure that, at each round t, each $i \in V$ is observed with probability at least $(\beta \ln t)/(t\hat{\Delta}_{t,i}^2)$, independently of whether i is in the exploration set at round t.

Formally, our procedure ensures that we can lower bound the probability with which any arm is observed. In the algorithm we use $P_{t,i} = \mathbb{P}[i \in N^{\text{out}}(I_t)]$ to denote the probability that arm i is observed at round t. We can lower bound this quantity by only considering the minimum rate at which each arm is observed according to the exploration rate $\varepsilon_{t,i}$ and our construction of the exploration sets. We use $o_{t,i}$ to denote that quantity, and we have for all t and i,

$$P_{t,i} \geq o_{t,i} = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}. \tag{4}$$

The definition of $o_{t,i}$ uses that S_t is a dominating set. The difference between $\varepsilon_{t,i}$ and $o_{t,i}$ is key to take advantage of the graph structure. First, we need to lower bound $o_{t,i}$ to ensure that enough observations (counted by $O_{t,i}$ in Algorithm 1) are made for each arm, such that our gap estimates are reliable. Simultaneously, we upper bound $\varepsilon_{t,i}$ to ensure that the extra exploration is not too costly. Here we benefit from the fact that S_t is a strongly independent set on G.

We ensure that all arms get sufficiently many observations and derive the following concentration bounds on the gap estimates $\hat{\Delta}_{t,i}$ computed by Algorithm 1. Concentration of the gap estimates around the true gaps is crucial for bounding the regret in the stochastic setting.

Lemma 1. *If Algorithm 1 is run with parameters $\gamma \geq 3$, $\beta \geq 64(\gamma+1) \geq 256$, and exploration rates $\varepsilon_{t,i}$, such that for all $t \geq 1$ and $i \in V$, $P_{t,i}$ satisfies equation* (4) *for some $\lambda \in [1, K]$, then for all $i \in V$ and $t \geq 1$,*

$$\mathbb{P}\left[\hat{\Delta}_{t,i} \geq \overline{\Delta}_i\right] \leq \frac{1}{Kt^{\gamma-1}}.$$

Furthermore, for any arm i with $\Delta_i > 0$ let $t_{\min}(i) := \max\left\{t \geq 0 : \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}} \leq \frac{\beta \ln t}{t\Delta_i^2}\right\}$. Then for any arm i with $\Delta_i > 0$ and $t \geq t_{\min}(i)$,

$$\mathbb{P}\left[\hat{\Delta}_{t,i} \leq \frac{1}{2}\Delta_i\right] \leq \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} + \frac{2}{Kt^{\gamma-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{10}}. \tag{5}$$

A proof of the lemma is provided in Appendix C.

We run the algorithm with $\gamma = 4$ and $\beta = 64(\gamma+1) = 320$ which is a different parameterization from the EXP3++ algorithm (Seldin and Lugosi, 2017), which uses $\gamma = 3$ and $\beta = 256$. Picking a larger value of γ means that the confidence intervals are slightly larger, which allows us to obtain a better dependency on the suboptimality gaps.

Indeed, under the same assumptions as in Lemma 1, if $\gamma = 4$ and $t \geq t_{\min}(i)$, we have that

$$\frac{(\ln t)^2}{t} \leq \frac{\lambda \Delta_i^4 \ln K}{4K^2\beta^2}, \text{ implying } \left(\frac{\ln t}{t\Delta_i^2}\right)^2 = \frac{(\ln t)^2}{t^2\Delta_i^4} = \frac{(\ln t)^2}{t}\frac{1}{t\Delta_i^4} \leq \frac{\lambda\Delta_i^4 \ln K}{4K^2\beta^2}\frac{1}{t\Delta_i^4} = \frac{1}{t}\frac{\lambda \ln K}{4K^2\beta^2}.$$

4 Adversarial Analysis

Our result for the adversarial regime generalizes the analysis of both Alon et al. (2017) and Alon et al. (2015) as we derive a bound that depends on the both the independence number and the strong independence number simultaneously. In order to do so, we define the quantity:

$$\theta_t := \sum_{i \in V} \frac{p_{t,i}}{P_{t,i}}, \tag{6}$$

which is the sum of the ratios of the probability of playing an arm to the probability of observing its loss. Bounding this sum of ratios is key to obtain a dependency on graph quantities, and Alon et al. (2017) and Alon et al. (2015) respectively bound equation (6) in terms of the strong independence number (Lemma 8) and the independence number (Lemma 7) at the cost of a logarithmic factor. By defining the learning rate in terms of θ, it is possible to obtain both bounds simultaneously.

Theorem 2. *Assume that Algorithm 1 is run with a directed feedback graph $G = (V, E)$, with learning rate $\eta_t = \sqrt{\frac{\ln K}{2\sum_{s=K}^{t-1}\theta_s}}$ and the exploration rate defined in* (2)–(3) *with $\gamma = 4$, and $\beta = 320$. For any $\lambda \in [1, \min(\tilde{\alpha}, \alpha \ln T)]$, the pseudo-regret against any oblivious loss sequence satisfies*

$$\mathcal{R}_T \leq \min\left\{4\sqrt{\tilde{\alpha}T\ln K}, 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T}\right\} + 2K,$$

where $K = |V|$, α is the independence number of G and $\tilde{\alpha}$ is its strong independence number.

On undirected graphs, the first part of the bound is always smaller, and it matches the bound of Alon et al. (2017). This implies that in the adversarial regime we are not paying a price for the extra guarantees that we derive in the stochastic regime. On directed graphs, if the difference between α and $\tilde{\alpha}$ is large, the second half of the bound may be advantageous. Furthermore, we note that the extra

logarithmic factor is only of order $\sqrt{\ln(T)}$, which is a slight improvement on the $\ln T$ dependency of Alon et al. (2015).

We give a sketch of the proof here and defer the detailed proof to Appendix B.

Proof sketch. We separate the first K rounds, in which the algorithm plays deterministically, from the remaining rounds, where we bound separately the contributions to the regret from the exponential weights and from the extra exploration. To bound the contribution of the extra exploration, we use that $\varepsilon_{t,i} \leq \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}$ for all t and i, meaning that the extra exploration contributes at most $O(\sqrt{\lambda T \ln K})$ to the regret. For bounding the contribution of the exponential weights to the regret, we follow the standard analysis of EXP3 with time varying learning rate (Bubeck and Cesa-Bianchi, 2012). We bound the second order term by exploiting the fact that $p_{t,i}$ and $q_{t,i}$ are close to each other because $\varepsilon_{t,i} \leq \frac{1}{2K}$ for all t and i. This allows us to bound the second order term in terms of θ_t, which simplifies with the learning rate. This quantity can be bounded by the strong independence number of the graph (Lemma 10 Alon et al. (2017)) or the independence number of the graph at the cost of a logarithmic factor (Lemma 5 Alon et al. (2015)), which gives the two parts of the bound.

5 Stochastic Analysis

In the stochastic regime, tuning λ affects the tightness of the bound. If the learner has knowledge of the independence and strong independence numbers but does not know the time horizon, picking $\lambda = \alpha$ is a safe choice to ensure that Theorem 2 holds.

Our result for the stochastic regime is given in the following theorem.

Theorem 3. *Let $G = (V, E)$ be a directed feedback graph with $K = |V|$ and independence number α and strong independence number $\tilde{\alpha}$. Under the same conditions as in Theorem 2 and choosing $\lambda = \alpha$, the pseudo-regret of Algorithm 1 against any stochastic stochastic loss sequence, satisfies:*

$$\mathcal{R}_T \leq \max_{\mathrm{Ind} \in \mathcal{I}(G)} \left\{ \sum_{i \in \mathrm{Ind}\,:\,\Delta_i > 0} \frac{4\beta\,(\ln T)^2}{\Delta_i} \right\} + 2\alpha \ln T$$
$$+ \sum_{i:\Delta_i>0} \frac{16K}{\Delta_i} + \frac{1020\beta K}{\Delta_{min}^2} \left(\ln \left(\frac{\beta K}{\Delta_{min}} \right) \right)^{3/2},$$

where $\mathcal{I}(G)$ is the collection of all strongly independent subsets of G.

We remark that the last two terms do not depend on T. Moreover, the leading coefficient of the term scaling with $(\ln T)^2$ sums over an independence set (as opposed to summing over the entire action set). The lower bound for this problem scales as $\Omega(c^* \ln T)$, where c^* is a graph dependent quantity which takes the size of the suboptimality gaps into account (Buccapatnam et al., 2014). Compared to that, our result is suboptimal by a logarithmic factor and our dependency on the strong independence number of G is weaker. The algorithms of Buccapatnam et al. (2014, 2017) and Cohen et al. (2016) almost match the lower bound, but their elimination based structure prevents them from being applicable in best-of-both-worlds settings. In the undirected case, we obtain the same dependence on T and on the set of arms as the UCB-N algorithm analysed by Lykouris et al. (2020).

We provide a sketch of the proof here. The detailed version can be found in Appendix D.

Proof sketch. Let $t_{\min} = \max_{i:\Delta_i>0}\{t_{\min}(i)\} = \max\left\{t \geq 0 : \frac{1}{2}\sqrt{\frac{\alpha \ln K}{tK^2}} \leq \frac{\beta \ln t}{t\Delta_{\min}^2}\right\}$. The pseudo-regret can be decomposed by treating the first $t_{\min}$ rounds like in the adversarial case, and by using a refined bound for the stochastic regime in the remaining rounds.

$$R_T = R_{t_{\min}} + \sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \mathbb{E}\left[p_{t,i}\right] \leq R_{t_{\min}} + \sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \big(\mathbb{E}\left[q_{t,i}\right] + \mathbb{E}\left[\varepsilon_{t,i}\right]\big). \tag{7}$$

Note that $t_{\min}$ is time independent: $t_{\min} = \frac{c}{\Delta_{\min}^4}\left(\ln\left(\frac{c}{\Delta_{\min}^4}\right)\right)^2$ for a positive constant c, therefore,

$$R_{t_{\min}} = C_0\sqrt{\alpha\, t_{\min}} \log\left(t_{\min}\right) = C_1 \frac{K}{\Delta_{\min}^2}\left(\ln\left(\frac{K}{\Delta_{\min}}\right)\right), \tag{8}$$

where the first equality follows from the second part of the bound presented in Theorem 2 and C_0, C_1 are universal constants. After the initial $t_{\min}$ rounds, enough observations on all arms have been gathered to ensure with high probability that the gap estimates of all arms are close to their true gaps, as stated in Lemma 1. These concentration inequalities allow us to show that the two following propositions hold.

Proposition 2 (informal)**.** *The contribution of the exponential weights to the pseudo-regret can be bounded as:*

$$\sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \mathbb{E}\left[q_{t,i}\right] \leq C_2 \sum_{i\,:\,\Delta_i>0} \frac{K}{\Delta_i} + O\left(\alpha \ln T\right)$$

for a universal constant C_2.

Proposition 3 (informal)**.** *The contribution of the extra exploration to the pseudo-regret can be bounded as:*

$$\sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \mathbb{E}\left[\varepsilon_{t,i}\right] = O\left(\max_{\mathrm{Ind}\in\mathcal{I}(G)} \left\{ \sum_{i\in\mathrm{Ind}\,:\,\Delta_i>0} \frac{\ln^2 T}{\Delta_i} \right\} + \alpha \ln T \right).$$

Formal statements and proofs of the above propositions are in Appendix D. These propositions ensure that after $t_{\min}$ steps the exponential weights of all suboptimal arms i are small, the extra exploration $\varepsilon_{t,i}$ achieves the correct rate, and that the sum of the probabilities that the suboptimality gap estimates fail in any of the rounds is of order $O(\alpha \ln T)$. Applying these propositions to Equation (7) finishes the proof.

Our approach to bound the pseudo-regret in the initial rounds differs from the one of Seldin and Lugosi (2017) as we take advantage of the adversarial bound in these rounds. (Mourtada and Gaïffas (2019) used a similar approach to derive best-of-both worlds guarantees for the Hedge algorithm.) This refinement improves upon the result of Seldin and Lugosi (2017) by replacing $\sum_{i:\Delta_i>0} \frac{1}{\Delta_i^3}$ with $\frac{1}{\Delta_{\min}^2}$ (numerical constants ignored) in the time-independent part of the bound.

For instances where the independence number and the strong independence number are close to each other, in particular in the case of undirected graphs, the analysis of the initial rounds can be improved by using the first part of Theorem 2, which depends on $\tilde{\alpha}$ rather than the second part, which depends on α when bounding the regret on the initial t_{min} rounds.

Corollary 1. *Let $G = (V, E)$ be a directed feedback graph with $K = |V|$ and a strong independence number $\tilde{\alpha}$. Under the same conditions as in Theorem 2, the pseudo-regret of Algorithm 1 against any stochastic stochastic loss sequence, satisfies:*

$$\mathcal{R}_T \leq \max_{\mathrm{Ind}\in\mathcal{I}(G)} \left\{ \sum_{i\in\mathrm{Ind}\,:\,\Delta_i>0} \frac{4\beta\left(\ln T\right)^2}{\Delta_i} \right\} + 2\tilde{\alpha}\ln T + \sum_{i\,:\,\Delta_i>0} \frac{16K}{\Delta_i} + \frac{161\beta K}{\Delta_{min}^2} \ln\left(\frac{\sqrt{\beta}K}{\Delta_{min}}\right).$$

6 Extension to Time Varying Feedback Graphs

The results presented in Theorem 3 and Corollary 1 assume that the learner has knowledge of the independence and strong independence numbers of the graph. Computing those numbers are NP-hard problems, which could lead to prohibitively large computation times. This is particularly true if one considers a natural extension of our results to the setting where the feedback graphs are allowed to change over time.

We consider a setting where an oblivious adversary chooses a feedback graph at each round and the algorithm observes the graph at the beginning of the round. In the stochastic regime, the knowledge of the full feedback graph is required at the beginning of the round in order to construct the exploration set.

As we do not know the independence numbers ahead of time, we tune the exploration rates defined in equation (2) with $\lambda = 1$ to ensure that the exploration is never too large. This exploration rate allows us to apply Lemma 1 with $\lambda = 1$, and derive the following result.

Theorem 4. *Assume that Algorithm 1 is run on a sequence of arbitrarily generated feedback graphs $G_1, G_2, \dots$ with learning rate $\eta_t = \sqrt{\frac{\ln K}{2\sum_{s=K}^{t-1}\theta_s}}$ and exploration rates defined in* (2) *and* (3) *with $\lambda = 1$, $\gamma = 4$ and $\beta = 320$. Then the pseudo-regret against any oblivious loss sequence satisfies*

$$\mathcal{R}_T \leq \min\left\{4\sqrt{\sum_{t=1}^{T}\tilde{\alpha}_t \ln K},\ 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\sum_{t=1}^{T}\alpha_t}\right\} + 2K,$$

where for all $t \geq 1$, α_t and $\tilde{\alpha}_t$ are the independence and strong independence numbers of G_t. Simultaneously, the pseudo-regret against stochastic losses satisfies:

$$R_T \leq \inf_{0\leq n\leq T}\left\{\max_{S\subset V:|S|=\tilde{A}_n}\left\{\sum_{i\in S:\Delta_i>0}\frac{4\beta\ln^2 T}{\Delta_i}\right\} + n\right\}$$
$$+ 2\ln T + \sum_{i:\Delta_i>0}\frac{16K}{\Delta_i} + \frac{161\beta K^{3/2}}{\Delta_{min}^2}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{min}}\right).,$$

where $\tilde{A}_n$ is the nth largest element in the set containing the strong independence number of all the G_t, for $t \leq T$.

A proof of this theorem is provided in Appendix E.

In the adversarial regime, adapting to graphs that change over time is seamless and does not come at any cost, as using a sequence of fixed graphs exactly recovers the bound of Theorem 2. In the stochastic regime, using $\lambda = 1$ allows us to obtain the same tight constants as in Corollary 1, and only comes at the cost of a multiplicative $\sqrt{K}$ factor in the last term of the bound. Furthermore, the first term of the bound is a sum over the $\tilde{A}_n$ arms that have the smallest non-zero suboptimality gaps. In the case of undirected graphs, if we upper bound the infimum by taking $n = 0$, we have $\tilde{\alpha}_0 = \max_{t>1}\{\alpha(G_t)\}$, which matches the dependency on gaps achieved by Cohen et al. (2016), who got an $O\left(\max_{S\subset V\setminus\{i^*\}:|S|=O(\alpha)}\sum_{i\in S}\frac{\ln T}{\Delta_i}\right)$ bound. This trick is particularly useful if most of the graphs have a small strong independence number and very few have a large independence number, as we can consider the graphs that have a large independence number separately at the cost of an additive constant and in return the dominating term will scale with the strong independence number of the remaining graphs, which may be much smaller.

7 Conclusion

Erez and Koren (2021) left open the following questions: is it possible to achieve best-of-both-worlds regret bounds in terms of the independence number, and can the dependence on T in their regret bounds be improved? We partially answered these questions with the EXP3.G++ algorithm and derived near-optimal best-of-both-worlds guarantees for directed feedback graphs. Our regret bounds depend on the independence number of the feedback graphs and improve upon the results of Erez and Koren (2021) by poly-logarithmic factors in both the adversarial and stochastic regimes. Furthermore, we extended our results to time-varying feedback graphs with a computationally efficient algorithm.

Acknowledgments and Disclosure of Funding

CR and YS acknowledge partial support by the Independent Research Fund Denmark, grant number 9040-00361B. DvdH and NCB gratefully acknowledge partial support from the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR) and the EU Horizon 2020 ICT-48 research and innovation action under grant agreement 951847, project ELISE (European Learning and Intelligent Systems Excellence).

References

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2015.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 2017.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2016.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 2002.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2012.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff. Stochastic bandits with side observations on networks. *Association for Computing Machinery*, 2014.

Swapna Buccapatnam, Fang Liu, Atilla Eryilmaz, and Ness B. Shroff. Reward maximization under uncertainty: Leveraging side-observations on networks. *Journal of Machine Learning Research*, 2017.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 2013.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tim van Erven, Wouter M. Koolen, and Dirk van der Hoeven. Metagrad: Adaptation using multiple learning rates in online learning. *Journal of Machine Learning Research*, 2021.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge, online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2014.

Dirk van der Hoeven, Federico Fusco, and Nicolo Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2021.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic MDPs with known transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tiancheng Jin, Longbo Huang, and Haipeng Luo. The best of both worlds: Stochastic and adversarial episodic MDPs with unknown transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wouter M Koolen, Peter Grünwald, and Tim Van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108, 1994.

Thodoris Lykouris, Éva Tardos, and Drishti Wali. Feedback graph regret bounds for Thompson Sampling and UCB. *Proceedings of Machine Learning Research*, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Teodor V. Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Saeed Masoudian and Yevgeny Seldin. Improved analysis of the Tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2021.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20, 2019.

Jeffrey Negrea, Blair Bilodeau, Nicolò Campolongo, Francesco Orabona, and Dan Roy. Minimax optimal quantile and semi-adversarial regret via root-logarithmic regularizers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Francesco Orabona. A modern introduction to online learning. *CoRR*, 2019.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2020.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25, 1933.

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

A Tools to Bound Series

We use the following lemmas to bound series.

Lemma 2 (Lemma 11 (Seldin and Lugosi, 2017))**.** *For $\gamma \geq 2$ and $m \geq 1$:*

$$\sum_{k=m}^{n} \frac{1}{k^\gamma} \leq \frac{1}{2m^{\gamma-1}}.$$

Lemma 3 (Lemma 8 (Seldin et al., 2014))**.** *For any sequence of non-negative numbers $a_1, a_2, \ldots,$ such that $a_1 > 0$ and any power $\gamma \in (0,1)$ we have:*

$$\sum_{t=1}^{T} \frac{a_t}{\left(\sum_{s=1}^{t} a_s\right)^\gamma} \leq \frac{1}{1-\gamma}\left(\sum_{t=1}^{T} a_t\right)^{1-\gamma}.$$

We also require a variation of this bound to handle the case where the denominator of the sum only sums up to index $t-1$. The proof of this Lemma follows from Gaillard et al. (2014, Lemma 14) that we generalized to adapt to sequences of a_t that are not restricted to the $[0,1]$ interval.

Lemma 4. *For any sequence $a_1, a_2, \ldots,$ such that $a_s \in [1, K]$ for all s, we have:*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{K + \sum_{s<t} a_s}} \leq 2\sqrt{\sum_{t=1}^{T} a_t} + \sqrt{K}.$$

Proof. Let $s_t = \sum_{n=1}^{t} a_t$, and define $s_0 := 0$. We want to bound $\sum_{t=1}^{T} \frac{a_t}{\sqrt{K+\sum_{s<t} a_s}} = \sum_{t=1}^{T} \frac{a_t}{\sqrt{K+s_{t-1}}}$, where $\frac{1}{\sqrt{K+s}}$ is a decreasing function of s. Thus we have:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{a_t}{\sqrt{K+s_{t-1}}} &= \sum_{t=1}^{T} \frac{a_t}{\sqrt{K+s_t}} + \sum_{t=1}^{T} a_t \left(\frac{1}{\sqrt{K+s_{t-1}}} - \frac{1}{\sqrt{K+s_t}}\right) \\
&\leq \sum_{t=1}^{T} \frac{a_t}{\sqrt{s_t}} + K \sum_{t=1}^{T} \left(\frac{1}{\sqrt{K+s_{t-1}}} - \frac{1}{\sqrt{K+s_t}}\right) \\
&\leq \sum_{t=1}^{T} \frac{a_t}{\sqrt{s_t}} + K \frac{1}{\sqrt{K+s_0}} \\
&\leq 2\sqrt{s_t} + \sqrt{K},
\end{aligned}$$

where we use Lemma 3 in the last step. □

Lemma 5 (Lemma 3 (Thune and Seldin, 2018))**.** *For $c > 0$ we have*

$$\sum_{t=1}^{\infty} e^{-c\sqrt{t}} \leq \frac{2}{c^2} \qquad \textit{and} \qquad \sum_{t=1}^{\infty} e^{-ct} \leq \frac{1}{c}.$$

B Analysis of the Adversarial Regime

We follow the proof structure of Theorem 2 from Alon et al. (2015), and use Lemma 7 from Seldin and Slivkins (2014) where $X_{t,i} = \tilde{\ell}_{t,i}$ for all t, i as a base for the analysis of EXP3.

Lemma 6 (Lemma 7 (Seldin and Slivkins, 2014))**.** *For any K sequences of non-negative numbers $X_{1,i}, X_{2,i}, \ldots$ indexed by $i \in [K]$, and any non-increasing positive sequence $\eta_1, \eta_2, \ldots$, for $q_{t,i} = \frac{\exp(-\eta_t \sum_{s=1}^{t-1} X_{s,i})}{\sum_{j\in[K]} \exp(-\eta_t \sum_{s=1}^{t-1} X_{s,j})}$ (assuming for $t = 1$ the sum in the exponent is 0) we have:*

$$\sum_{t=1}^{T}\sum_{i=1}^{K} q_{t,i} X_{t,i} - \min_{k\in[K]} \sum_{t=1}^{T} X_{t,k} \leq \frac{\ln K}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t}{2} \left(\sum_{i\in[K]} q_{t,i} X_{t,i}^2 \right).$$

We then consider two ways to take advantage of the graph structure. In the first case, we rely on Lemma 5 from Alon et al. (2015) in order to derive a bound that scales with the independence number.

Lemma 7 (Lemma 5 (Alon et al., 2015))**.** *Let $G = (V, E)$ be a directed graph with $|V| = K$, in which each node $i \in V$ is assigned a positive weight w_i. Assume that $\sum_{i\in V} w_i \leq 1$, and that $w_i \geq \epsilon$ for all $i \in V$ for some constant $0 < \epsilon < \frac{1}{2}$. Then*

$$\sum_{i\in V} \frac{w_i}{w_i + \sum_{j\in N^{\text{in}}(i)} w_j} \leq 4\alpha \ln\left(\frac{4K}{\alpha\epsilon}\right),$$

where $\alpha = \alpha(G)$ is the independence number of G.

In the second case, we want to derive a bound that scales with the strong independence number of the graph. To do so, we rely on Lemma 10 from Alon et al. (2017). That Lemma depends on a different graph dependent quantity: the maximum acyclic subgraph of a feedback graph G, which is defined by Alon et al. as follows. We show that we can upper bound the maximum acyclic subgraph of any graph G in terms of its strong independence number.

Definition 2. *Given a directed graph $G = (V, E)$, an acyclic subgraph of G is any $G' = (V', E')$ such that $V' \subseteq V$ and $E' = E \cap (V' \times V')$, with no (directed) cycles. We denote by $\text{mas}(G) = |V'|$ the maximum size of such a V'.*

A key property of the maximum acyclic subgraph is that for any graph G, $\alpha(G) \leq \text{mas}(G)$ and for undirected graphs, $\alpha(G) = \text{mas}(G)$ (Alon et al., 2017). We now show that for any directed graph G, the maximum acyclic subset of G can be upper bounded by its strong independence number.

Proposition 4. *Let $G = (V, E)$ be a directed graph.* $\text{mas}(G) \leq \tilde{\alpha}(G)$.

Proof. Let $G' = (V', E')$ be an acyclic subgraph of G, where $V' \subseteq V$ and $E' = E \cap (V' \times V')$. For any $i, j \in V'$, we know that $(i, j) \notin E'$ or $(j, i) \notin E'$, otherwise i and j would be part of a cycle which contradicts the definition of G'. Thus i and j are strongly independent and V' is a strongly independent set. As this holds for all acyclic subgraphs of G, we deduce that $\text{mas}(G) \leq \tilde{\alpha}(G)$ which finishes the proof. □

This characterization allows us to use the following lemma and derive bounds that scale with the strong independence number.

Lemma 8 (Lemma 10 Alon et al. (2017))**.** *let $G = (V, E)$ be a directed graph with vertex set $V = \{1, \ldots, K\}$, and arc set V. Then, for any distribution p over V we have:*

$$\sum_{i=1}^{K} \frac{p_i}{p_i + \sum_{j\in N^{\text{in}}(i)} p_j} \leq \text{mas}(G).$$

With those results, we can move on to the proof of Theorem 2.

Proof of Theorem 2. Without loss of generality, we assume that $K \geq 2$.

Recall that the algorithm initializes by playing each arm once, which adds at most K to the regret. The EXP3 part of the analysis starts from round $K + 1$. We can upper trivially upper bound the

first K rounds by 1 and then analyse the algorithm from round $t = K+1$. Precisely, we bound the pseudo-regret as:

$$\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] \\
&\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=K+1}^{T} \ell_{t,i}\right] \\
&= K + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} p_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] \\
&\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} q_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} \varepsilon_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right], \quad (9)
\end{aligned}$$

where $i^* = \text{argmin}\left\{\sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right\}$, and $\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] = \ell_{t,i}$. Equation (9) follows from $p_{t,i} \leq q_{t,i} + \varepsilon_{t,i}$. We can consider the contribution of $q_{t,i}$ and $\varepsilon_{t,i}$ separately.

We recall that the learning rate is defined from index $t \geq K+1$ by:

$$\eta_t = \sqrt{\frac{\ln K}{2\sum_{s=K}^{t-1}\theta_s}}, \qquad \text{where } \theta_t = \sum_{i\in V}\frac{p_{t,i}}{P_{t,i}}.$$

As the quantities $p_{t,i}$ are not defined for $t \geq K$, we set $\theta_K := K$ to ensure that the learning rate is well defined and non-increasing at all the rounds where we use exponential weights. As the learning rate is a random variable, we have:

$$\begin{aligned}
\mathcal{R}_T &\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} q_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} \varepsilon_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right] \\
&\leq K + \mathbb{E}\left[\mathbb{E}_t\left[\frac{\ln K}{\eta_T}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}_t\left[\sum_{i\in V}\frac{\eta_t}{2}\frac{q_{t,i}}{P_{t,i}}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} \mathbb{E}_t\left[\varepsilon_{t,i}\tilde{\ell}_{t,i}\right]\right]. \quad (10)
\end{aligned}$$

We now want to bound each term as a function of the θ_t.

The first term becomes:

$$\mathbb{E}\left[\mathbb{E}_t\left[\frac{\ln K}{\eta_T}\right]\right] \leq \sqrt{2\ln K}\,\mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{s=K}^{T-1}\theta_s}\right]\right].$$

To bound the second term, we first note that using $\frac{1}{2K}$ as an upper bound on ε_t, we ensure that for all t and i, $p_{t,i} \geq (1 - \sum_{j\in V}\varepsilon_{t,j})q_{t,i} \geq \frac{1}{2}q_{t,i}$ which gives:

$$\sum_{i\in V}\frac{q_{t,i}}{P_{t,i}} \leq 2\sum_{i\in V}\frac{p_{t,i}}{P_{t,i}} = 2\theta_t.$$

Then, the second term can be bounded as:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}_t\left[\sum_{i\in V}\frac{\eta_t}{2}\frac{q_{t,i}}{P_{t,i}}\right]\right] &\leq \mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}_t\left[\eta_t\sum_{i\in V}\frac{p_{t,i}}{P_{t,i}}\right]\right] \\
&\leq \sqrt{\ln K}\,\mathbb{E}\left[\sum_{t=K+1}^{T} \mathbb{E}_t\left[\frac{\theta_t}{\sqrt{2\sum_{s=K}^{t-1}\theta_s}}\right]\right] \\
&\leq \sqrt{2\ln K}\,\mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{K}, \quad (11)
\end{aligned}$$

where equation (11) follows from Lemma 4.

For the last term, we recall that we bounded $\varepsilon_{t,i} \leq \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}$, and we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K}\varepsilon_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right] &\leq \sum_{t=K+1}^{T}\mathbb{E}\left[\sum_{i=1}^{K}\frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right] \\
&\leq \sum_{t=1}^{T}\frac{1}{2}\sqrt{\frac{\lambda \ln K}{t}} \\
&\leq \sqrt{\lambda T \ln K}.
\end{aligned}$$

Using those three bounds in equation 10 gives:

$$\mathcal{R}_T \leq 2K + 2\sqrt{2\ln K}\,\mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{\lambda T \ln K}. \tag{12}$$

All that remains is to bound θ and λ. To obtain the first part of the bound, we use Proposition 4 and Lemma 8, and deduce that for all $t \leq K+1$:

$$\theta_t \leq \tilde{\alpha}\text{ , which gives }\sqrt{\sum_{t=K+1}^{T}\theta_t} \leq \sqrt{\tilde{\alpha}T}. \tag{13}$$

Using $\lambda \leq \tilde{\alpha}$, we deduce that:

$$\begin{aligned}
\mathcal{R}_T &\leq 2K + 2\sqrt{2\ln K}\,\sqrt{\tilde{\alpha}T} + \sqrt{\tilde{\alpha}T\ln K} \\
&\leq 4\sqrt{\tilde{\alpha}T\ln K} + 2K.
\end{aligned} \tag{14}$$

For the second part of the bound, we use Lemma 7 and recall that for all t and i, $\varepsilon_{t,i} \geq \frac{4}{t^2}$. We deduce the following upper bound:

$$\theta_t \leq 4\alpha\ln\left(\frac{t^2K}{\alpha}\right) \leq 8\alpha\ln(KT)\text{ , which gives }\sqrt{\sum_{t=K+1}^{T}\theta_t} \leq \sqrt{8\alpha T\ln(KT)}.$$

Using $\lambda \leq \alpha \ln T$, we deduce that:

$$\begin{aligned}
\mathcal{R}_T &\leq 2K + 2\sqrt{2\ln K}\,\sqrt{8\alpha T\ln(KT)} + \sqrt{\ln K}\sqrt{\ln T}\sqrt{\alpha T} \\
&\leq 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T} + 2K,
\end{aligned} \tag{15}$$

Taking the minimum between equations (14) and (15) finishes the proof.

□

C Properties of the Gaps Estimates

In this section, we provide upper and lower high probability bounds for the estimates of the suboptimality gaps. We decompose the proof of Lemma 1 in two parts.

C.1 Upper bound

We start by deriving a high probability upper bound. For this bound, we have to be careful with the fact that the gap estimates are clipped in the $[0,1]$ interval. We first upper derive bounds on UCB and LCB.

Lemma 9. *The confidence intervals satisfy:*

$$\mathbb{P}\left[UCB_{t,i} \leq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}$$
$$\textit{and} \qquad \mathbb{P}\left[LCB_{t,i} \geq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}.$$

Proof. Let $\overline{\text{UCB}}_t$ and $\overline{\text{LCB}}_t$ be the non clipped versions of the UCB_t and LCB_t. In other words, for all i and t:

$$\overline{\text{UCB}}_{t,i} = \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} + \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}}$$
$$\text{and} \qquad \overline{\text{LCB}}_{t,i} = \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} - \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}}.$$

Then, through standard UCB analysis using Hoeffding's inequality (see for example Seldin and Lugosi (2017)), we have:

$$\mathbb{P}\left[\overline{\text{UCB}}_{t,i} \leq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}$$
$$\text{and} \qquad \mathbb{P}\left[\overline{\text{LCB}}_{t,i} \geq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}.$$

By definition, we have $\text{UCB}_{t,i} = \min\left\{1, \overline{\text{UCB}}_{t,i}\right\} \leq \overline{\text{UCB}}_{t,i}$ and $\text{LCB}_{t,i} = \max\left\{0, \overline{\text{LCB}}_{t,i}\right\} \geq \overline{\text{LCB}}_{t,i}$, so:

$$\mathbb{P}\left[\text{UCB}_{t,i} \leq \mu_i\right] \leq \mathbb{P}\left[\overline{\text{UCB}}_{t,i} \leq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}$$
$$\text{and} \qquad \mathbb{P}\left[\text{LCB}_{t,i} \geq \mu_i\right] \leq \mathbb{P}\left[\overline{\text{LCB}}_{t,i} \geq \mu_i\right] \leq \frac{1}{KT^{\gamma-1}}.$$

□

Using this result, we can move on to bound the gap estimates.

Proof of the first part of Lemma 1. We recall that $\hat{\Delta}_{t,i} = \max\left\{0, \text{LCB}_{t,i} - \min_{j\neq i} \text{UCB}_{t,j}\right\}$. Then using Lemma 9, we have:

$$\begin{aligned}
\mathbb{P}\left[\hat{\Delta}_{t,i} \geq \overline{\Delta_i}\right] &= \mathbb{P}\left[\text{LCB}_{t,i} - \min_{j\neq i} \text{UCB}_{t,j} \geq \overline{\Delta_i}\right] \\
&\leq \mathbb{P}\left[\text{LCB}_{t,i} - \min_{j\neq i} \text{UCB}_{t,j} \geq \Delta_i\right] \\
&\leq \mathbb{P}\left[\text{LCB}_{t,i} \geq \mu_i\right] + \sum_{j\neq i} \mathbb{P}\left[\text{UCB}_{t,j} \leq \mu_j\right] \\
&\leq K\frac{1}{Kt^{\gamma-1}} = \frac{1}{t^{\gamma-1}},
\end{aligned}$$

where the first step takes advantage of the fact that $\overline{\Delta}_i > 0$ for all i, allowing to remove the maximum. The second step relies on $\Delta_i \leq \overline{\Delta}_i$, and we finish the proof with a union bound and applying Lemma 9. □

C.2 Lower bound

To derive a lower bound on the gap estimates and prove the second part of Lemma 1, we start by proving some intermediate results. recall that we use $o_{t,i}$ to lower bound the probability of observing the loss of arm i at round t, and that by construction we have for all t, i:

$$o_{t,i} = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}.$$

We also recall that for all i such that $\Delta_i > 0$, we defined $t_{\min}(i)$ as:

$$t_{\min}(i) = \max\left\{t \geq 0 : \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}} \leq \frac{\beta \ln t}{t\Delta_i^2}\right\}.$$

A lower bound for $o_{t,i}$. As $\hat{\Delta}_{t,i}$ is a random variable, we derive a high probability lower bounds on $o_{t,i}$.

Definition 3. *We define the following events:*

$$\mathcal{E}(i,t) = \left\{\forall s \in [K+1,t] : o_{s,i} \geq \frac{\beta \ln t}{t\Delta_i^2}\right\},$$
$$\mathcal{E}(i^*,i,t) = \left\{\forall s \in [K+1,t] : o_{s,i^*} \geq \frac{\beta \ln t}{t\Delta_i^2}\right\},$$

where i^* is an optimal arm and i a suboptimal arm.

Note that the second event lower bounds the rate at which observations on optimal arm i^* are gathered in terms of the gap with the suboptimal arm i.

Lemma 10. *For any i suboptimal arm and i^* optimal arm, and $t \geq t_{\min}(i)$ and $\gamma \geq 3$, we have:*

$$\mathbb{P}\left[\overline{\mathcal{E}(i,t)}\right] \leq \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2},$$
$$\mathbb{P}\left[\overline{\mathcal{E}(i^*,i,t)}\right] \leq \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2}.$$

Proof of Lemma 10. The proof is very similar for the two inequalities. By definition for all s and i, we have $\hat{\Delta}_{s,i} \leq 1$. Thus, $\frac{\beta \ln s}{s\hat{\Delta}_{s,i}^2} \geq \frac{\beta \ln s}{s}$. Then for $s \in \left[K+1, \frac{t\Delta_i^2}{\ln t}\right]$, we have $\frac{\beta \ln s}{s} \geq \frac{\beta \ln s \ln t}{t\Delta_i^2} \geq \frac{\beta \ln t}{t\Delta_i^2}$, as $s > K \geq 2$, so $\ln s \geq 1$. Furthermore, as $t \geq t_{\min}(i)$ then for all $s \in [K+1,t]$, we have $\frac{1}{2}\sqrt{\frac{\lambda \ln K}{sK^2}} \geq \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}} \geq \frac{\beta \ln t}{t\Delta_i^2}$ and $\frac{1}{2K} \geq \frac{\beta \ln t}{t\Delta_i^2}$. We deduce:

$$\begin{aligned}
\mathbb{P}\left[\overline{\mathcal{E}(i,t)}\right] &= \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : o_{s,i} \leq \frac{\beta \ln t}{t\Delta_i^2}\right] \\
&\leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \hat{\Delta}_{s,i} \geq \Delta_i\right] \qquad (16) \\
&\leq \sum_{s=\frac{t\Delta_i^2}{\ln t}} \frac{1}{s^{\gamma-1}} \\
&\leq \frac{1}{2}\left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2},
\end{aligned}$$

where the last summation follows from Lemma 2. The proof of the second inequality is similar, Equation (16) only requiring the extra step:

$$\mathbb{P}\left[\exists s \in \left[\frac{t\Delta_{i^*}^2}{\ln t}, t\right] : \hat{\Delta}_{s,i} \geq \Delta_i\right] \leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_{i^*}^2}{\ln t}, t\right] : \hat{\Delta}_{s,i} \geq \overline{\Delta_{i^*}}\right],$$

which follows from $\Delta_i \geq \Delta_{\min} = \overline{\Delta_{i^*}}$ as i is a suboptimal arm and i^* is an optimal arm. □

A lower bound for $O_{t,i}$ We now want to lower bound the number of observations of an arm up to round t. We rely on the following concentration inequality.

Theorem 5 (Theorem 8 (Seldin and Lugosi, 2017))**.** *Let $X_1, \ldots, X_n$ be Bernoulli random variables adapted to filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$ (in particular, X_s may depend on $X_1, \ldots, X_{s-1}$). Let $\mathcal{E}_\lambda$ be the event $\mathcal{E}_\lambda = \{\forall s : \mathbb{E}[X_s|\mathcal{F}_{s-1}] \geq \lambda\}$. Then,*

$$\mathbb{P}\left[\left(\sum_{s=1}^{n} X_s \leq \frac{1}{2}n\lambda\right) \wedge \mathcal{E}_\lambda\right] \leq e^{-n\lambda/8}.$$

We recall that the first K rounds of the algorithm are deterministic, and that each arm is observed at least once. We use $O_{[K+1:t],i}$ to refer to the number of observations from rounds $K+1$ to t, and we note that $O_{t,i} \geq O_{[K+1:t],i} + 1$. We have:

$$\begin{aligned}\mathbb{P}\left[O_{t,i} \leq \frac{\beta \ln t}{2\Delta_i^2}\right] &\leq \mathbb{P}\left[O_{[K+1:t],i} \leq \frac{\beta \ln t}{2\Delta_i^2} - 1\right] \\ &\leq \mathbb{P}\left[O_{[K+1:t],i} \leq \frac{\beta \ln t}{2\Delta_i^2}\frac{t-K}{t}\right],\end{aligned}$$

where the second step follows from, $\frac{\beta \ln t}{2\Delta_i^2} - 1 \leq \frac{\beta \ln t}{2\Delta_i^2}\frac{t-K}{t} \Leftrightarrow \frac{K\beta \ln t}{2t\Delta_i^2} \leq 1$, which is true for $t \geq t_{\min}(i)$ as

$$\frac{K\beta \ln t}{2t\Delta_i^2} \leq \frac{K\beta \ln t}{2\Delta_i^2}\frac{\Delta_i^4 \lambda \ln K}{4K^2\beta^2 \ln^2 t} \leq \frac{\Delta_i^2 \ln K}{8 \ln t} \leq \frac{\Delta_i^2}{8} \leq 1.$$

We can apply Theorem 5 on the $t-K$ random variables $\mathbb{1}\left[i \in N^{\text{out}}(I_s)\right]$ for $s \in [K+1, t]$ and we get:

$$\begin{aligned}\mathbb{P}\left[O_{t,i} \leq \frac{\beta \ln t}{2\Delta_i^2}\right] &\leq \mathbb{P}\left[\left(O_{[K+1:t],i} \leq \frac{\beta \ln t}{2\Delta_i^2}\frac{t-K}{t}\right) \wedge \mathcal{E}_{t,i}\right] + \mathbb{P}\left[\overline{\mathcal{E}_{t,i}}\right] \\ &\leq e^{-\frac{t-K}{t}\frac{\beta \ln t}{8\Delta_i^2}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} \\ &\leq e^{-\frac{3}{4}\frac{\beta \ln t}{8\Delta_i^2}} + \frac{1}{2}\left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} \\ &\leq \left(\frac{1}{t}\right)^{\beta/10} + \frac{1}{2}\left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2},\end{aligned}$$

where we use that $t \geq t_{\min}(i) \geq 4K$, so $\frac{t-K}{t} \geq \frac{3}{4}$.

A lower bound for $\hat{\Delta}_{t,i}$ Using Lemma 9, we know that the upper and lower confidence bounds satisfy: $\mathbb{P}\left[\text{UCB}_{t,i^*} \leq \mu_{i^*} \vee \text{LCB}_{t,i} \geq \mu_i\right] \leq \frac{2}{Kt^{\gamma-1}}$. Then assuming that $\text{UCB}_{t,i^*} \geq \mu_{i^*}$ and $\text{LCB}_{t,i} \leq \mu_i$, we have:

$$\begin{aligned}\hat{\Delta}_{t,i} \geq& \text{LCB}_{t,i} - \min_{j \neq i} \text{UCB}_{t,i} \\ \geq& \text{LCB}_{t,i} - \text{UCB}_{t,i^*} \\ \geq& \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} - \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - \frac{\hat{L}_{t-1,i^*}}{O_{t-1,i^*}} - \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}} \\ =& \frac{\hat{L}_{t-1,i}}{O_{t-1,i}} + \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} \\ &- \left(\frac{\hat{L}_{t-1,i^*}}{O_{t-1,i^*}} - \sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}}\right) - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}} \\ =& \text{UCB}_{t,i} - \text{LCB}_{t,i^*} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}} \\ \geq& \Delta_i - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}}.\end{aligned}$$

Using the previously derived high probability bounds, assuming that $O_{t,i} \geq \frac{\beta \ln t}{2\Delta_i^2}$ and $O_{t,i^*} \geq \frac{\beta \ln t}{2\Delta_i^2}$, and using that $t \geq t_{\min}(i) \geq K$, we have:

$$\begin{aligned}
\hat{\Delta}_{t,i} \geq & \Delta_i - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}} \\
\geq & \Delta_i - 4\sqrt{\frac{2\Delta_i^2 \gamma \ln\left(tK^{1/\gamma}\right)}{2\beta \ln t}} \\
\geq & \Delta_i - 4\sqrt{\frac{\Delta_i^2 (\gamma+1) \ln\left(tK^{1/\gamma}\right)}{\beta \ln t}} \qquad (17) \\
= & \Delta_i \left(1 - 4\sqrt{\frac{\gamma+1}{\beta}}\right),
\end{aligned}$$

where equation (17) follows from $t \geq K$, so $\gamma \ln\left(tK^{1/\gamma}\right) = \gamma \ln\left(t^\gamma K\right) \leq \ln\left(t^\gamma t\right) = (\gamma+1)\ln t$. Using that $\beta \geq 64\,(\gamma+1)$, we have:

$$\mathbb{P}\left[\hat{\Delta}_{t,i} \leq \frac{1}{2}\Delta_i\right] \leq \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} + \frac{2}{Kt^{\gamma-1}} + 2\left(\frac{1}{t}\right)^{\beta/10}.$$

D Analysis of the Stochastic Regime

In the stochastic regime, we decompose the regret bound into three terms that we bound separately. First, during the initial $t_{\min} = \max_{i:\Delta_i>0}\{t_{\min}(i)\}$ rounds we use the adversarial bound. Then, in the remaining rounds we bound the contribution of the exponential weights and of the exploration separately.

D.1 Control over the Initial Rounds

We start by deriving a time independent upper bound on $t_{\min}(i)$ for all vertices i such that $\Delta_i > 0$.

Proposition 5. *For any constant $c > e^2$, we have:*

$$\max_t \left\{t \leq c(\ln t)^2\right\} \leq 25c\,(\ln c)^2.$$

Proof. First, we note that for $t = 3$,

$$c(\ln t)^2 \geq e^2(\ln 3)^2 \geq 3 = t,$$

so the inequality is fulfilled at $t = 3$.

Furthermore, $(c(\ln t)^2)' = 2c\frac{\ln t}{t}$ is a decreasing function of t for $t \geq e$ and such that $\lim_{t\to\infty} 2c\frac{\ln t}{t} = 0$, whereas $(t)' = 1$ is constant. Thus, $\max_t\left\{t \leq c(\ln t)^2\right\}$ exists and is solution of

$$t = c\,(\ln t)^2\,.$$

Let's upper bound this t. We denote by W_{-1} the product log function. Then we have:

$$\begin{aligned}
t =& c\,(\ln t)^2 \\
\sqrt{t} =& \sqrt{c}\ \ln t \\
\sqrt{t} =& 2\sqrt{c}\ln(\sqrt{t}) \\
x =& b\ln(x) && b = 2\sqrt{c},\ x = \sqrt{t} \\
x =& -bW_{-1}\left(-\frac{1}{b}\right) && \text{for } b \geq e.
\end{aligned}$$

By Chatzigeorgiou (2013, Theorem 1), we have for $b \geq e$:

$$-bW_{-1}\left(-\frac{1}{b}\right) = -bW_{-1}\left(-\exp\left(-\ln\left(\frac{b}{e}\right)-1\right)\right) \leq b\left(1+\sqrt{2\ln\left(\frac{b}{e}\right)}+\ln\left(\frac{b}{e}\right)\right).$$

Thus, we have that for $c \geq e^2$,

$$\begin{aligned} t \leq & 4c\left(1+\sqrt{2\ln\left(\frac{2\sqrt{c}}{e}\right)}+\ln\left(\frac{2\sqrt{c}}{e}\right)\right)^2 \\ \leq & 4c\left(1+2\ln c\right)^2 \\ \leq & 25c\left(\ln c\right)^2, \end{aligned}$$

where the last step follows from $(1+2\ln c)^2 \leq (\frac{1}{2}\ln(e^2)+2(\ln c))^2 \leq (2.5\ln c)^2 = 6.25\left(\ln c\right)^2$. □

Proposition 6. *Under the conditions of Lemma 1 with $\gamma = 4$, $\beta = 320$ and $\lambda \in [1, K]$, the contribution of the initial $t_{\min}$ rounds to the regret can be bounded as:*

$$R_{t_{\min}} \leq \min\left\{\frac{160\beta K}{\Delta_{min}^2}\sqrt{\frac{\tilde{\alpha}}{\lambda}}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{min}}\right), \frac{1019\beta K}{\Delta_{min}^2}\sqrt{\frac{\alpha}{\lambda}}\left(\ln\left(\frac{\beta K}{\Delta_{min}}\right)\right)^{3/2}\right\} + 2K.$$

Proof. By definition, we have $t_{\min} = \max\left\{t \geq 0 : \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}} \leq \frac{\beta \ln t}{t\Delta_{\min}^2}\right\}$. By proposition 5, we have that $t_{\min} \leq 25d\left(\ln d\right)^2$, where $d = \frac{4\beta^2K^2}{\lambda \ln K \Delta_{\min}^4}$.

Then, we can use the first half Theorem 2 and deduce that:

$$\begin{aligned} R_{t_{\min}} &\leq 4\sqrt{\tilde{\alpha}\ln K\, t_{\min}} + 2K \\ &\leq 4\sqrt{\tilde{\alpha}\ln K\, 25d}\ln(d) + 2K \\ &= 4\sqrt{\tilde{\alpha}\ln K\, 25\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}}\ln\left(\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}\right) + 2K \\ &\leq \frac{160\beta K}{\Delta_{\min}^2}\sqrt{\frac{\tilde{\alpha}}{\lambda}}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K. \end{aligned}$$

For the second part of the bound, we use the second part of Theorem 2, and we deduce:

$$\begin{aligned} R_{t_{\min}} &\leq 9\sqrt{\alpha t_{\min}}\sqrt{\ln(Kt_{\min})}\sqrt{\ln K} + 2K \\ &\leq 9\sqrt{\alpha\ln K\, 25d}\ln(d)\sqrt{\ln\left(25Kd^2\right)} + 2K \qquad (18) \\ &\leq 9\sqrt{\alpha\ln K\, 25\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}}\ln\left(\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}\right)\sqrt{\ln\left(25K\left(\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}\right)^2\right)} + 2K \\ &\leq 9*5*2\frac{\beta K}{\Delta_{\min}^2}\sqrt{\frac{\alpha}{\lambda}}\ln\left(\frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}\right)\sqrt{\ln\left(\frac{400\beta^4K^5}{\Delta_{\min}^8}\right)} + 2K \qquad (19) \\ &\leq \frac{1019\beta K}{\Delta_{\min}^2}\sqrt{\frac{\alpha}{\lambda}}\left(\ln\left(\frac{\beta K}{\Delta_{\min}}\right)\right)^{3/2} + 2K. \end{aligned}$$

where the equation (18) uses that for $d > 1$ we have $d\left(\ln d\right)^2 \leq d^2$, and equation (19) uses that $400 \leq 320^4 \leq \beta^4$.

□

D.2 Control over the Exponential Weights

Proposition 2 introduced in the proof sketch of Theorem 3 is based on the following result.

Proposition 7. *Under the conditions of Lemma 1 with $\gamma = 4$, $\beta = 320$ and $\lambda \in [1, K]$, the sum of exponential weights with sequence of learning rates $\eta_1, \eta_2, \dots$ of each suboptimal arm i can be bounded as:*

$$\sum_{t=t_{\min}(i)}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \sum_{t=t_{\min}(i)}^{T} \left(e^{-\frac{1}{2}t\eta_t\Delta_i} + \frac{1}{t}\left(\frac{\lambda \ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right)$$

To prove this result, we can follow the same derivation as in Seldin and Lugosi (2017). We want to bound the $q_{t,i}$ for all i such that $\Delta_i > 0$ and $t \geq t_{\min}(i)$. First, we note that

$$\begin{aligned}
q_{t,i} &= \frac{\exp(-\eta_t \tilde{L}_{t,i})}{\sum_{j\in V}\exp(-\eta_t \tilde{L}_{t,j})} \\
&= \frac{\exp(-\eta_t(\tilde{L}_{t,i} - \tilde{L}_{t,i^*}))}{\sum_{j\in V}\exp(-\eta_t(\tilde{L}_{t,j} - \tilde{L}_{t,i^*})))} \\
&\leq \exp(-\eta_t(\tilde{L}_{t,i} - \tilde{L}_{t,i^*})) \\
&:= \exp(-\eta_t \tilde{\Delta}_{t,i}),
\end{aligned}$$

where i^* is the best arm, and where the inequality holds because one term of the sum is $\exp(-\eta_t(\tilde{L}_{t,i^*} - \tilde{L}_{t,i^*})) = 1$ and the other terms are positive, so the denominator is greater than 1. We now want to want to ensure that $\tilde{\Delta}_{t,i} := \tilde{L}_{t,i} - \tilde{L}_{t,i^*}$ is close to $t\Delta_i$. To do so, we want to apply a variant of Bernstein's inequality on the martingale sequence difference $t\Delta_i - \tilde{\Delta}_{t,i} = \sum_{s=1}^{t} X_s$, where each single term of the sequence is defined as $X_s = \Delta_i - (\tilde{\ell}_{s,i} - \tilde{\ell}_{s,i^*})$.

Theorem 6 (Bernstein's inequality for martingales)**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence with respect to filtration $\mathcal{F}_1, \dots, \mathcal{F}_n$, where each X_j is bounded from above, and let $S_i = \sum_{j=1}^{i} X_j$ be the associated martingale. Let $\nu_n = \sum_{j=1}^{n} \mathbb{E}\left[(X_j)^2 | \mathcal{F}_{j-1}\right]$ and $\kappa_n = \max_{1\leq j\leq n}\{X_j\}$. Then, for any $\delta > 0$:*

$$\mathbb{P}\left[\left(S_n \geq \sqrt{2\nu \ln\left(\frac{1}{\delta}\right)} + \frac{\kappa \ln\left(\frac{1}{\delta}\right)}{3}\right) \wedge (\nu_n \leq \nu) \wedge (\kappa_n \leq \kappa)\right] \leq \delta.$$

In order to apply the this theorem, we need to bound $\max_{1\leq s\leq n}\{X_s\}$ and $\sum_{s=1}^{n} \mathbb{E}\left[(X_s)^2 | \mathcal{F}_{s-1}\right]$.

Control of $\max_{1\leq s\leq t}\{X_s\}$ For each s we have:

$$\begin{aligned}
X_s &= \Delta_i - (\tilde{\ell}_{s,i} - \tilde{\ell}_{s,i^*}) \\
&\leq 1 + \tilde{\ell}_{s,i^*} \\
&\leq 1 + \frac{1}{P_{t,i^*}} \\
&\leq 1 + \max\left\{2K, 2\sqrt{\frac{sK^2}{\lambda \ln K}}, \frac{s\hat{\Delta}_{s,i^*}^2}{\beta \ln s}\right\} \qquad (20) \\
&\leq 1.25 \max\left\{2K, 2\sqrt{\frac{sK^2}{\lambda \ln K}}, \frac{s\hat{\Delta}_{s,i^*}^2}{\beta \ln s}\right\}
\end{aligned}$$

where equation (20) holds by definition of o_{t,i^*}. Using the same argument as in the proof of Lemma 1, we know that $t \geq t_{\min}(i)$, and if $s \leq \frac{t\Delta_i^2}{\ln t}$ then $\frac{s\hat{\Delta}_i^2}{\ln s} \leq \frac{s}{\beta} \leq \frac{t\Delta_i^2}{\beta \ln t}$ then:

$$\mathbb{P}\left[\exists s \leq t : \max\left\{2K, 2\sqrt{\frac{sK^2}{\lambda \ln K}}, \frac{s\hat{\Delta}_{s,i^*}^2}{\beta \ln s}\right\} \geq \frac{t\Delta_i^2}{\beta \ln t}\right] = \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \Delta_i\right]$$
$$\leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \overline{\Delta_{i^*}}\right],$$

because $\Delta_i \geq \Delta_{\min} = \bar{\Delta}_{i^*}$. Let $\kappa_t = \max_{1 \leq s \leq t}\{X_s\}$, and we deduce:

$$\mathbb{P}\left[\kappa_t \geq \frac{1.25t\Delta_i^2}{\beta \ln t}\right] \leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \overline{\Delta_{i^*}}\right].$$

Control of $\nu_t = \sum_{s=1}^t \mathbb{E}\left[(X_s)^2 | \mathcal{F}_{s-1}\right]$ We start by looking at each individual element of the sum.

$$\mathbb{E}\left[(X_s)^2 | \mathcal{F}_{s-1}\right] = \mathbb{E}\left[(\Delta_i - (\tilde{\ell}_{s,i} - \tilde{\ell}_{s,i^*}))^2 | \mathcal{F}_{s-1}\right]$$
$$\leq \mathbb{E}\left[(\tilde{\ell}_{s,i} - \tilde{\ell}_{s,i^*})^2 | \mathcal{F}_{s-1}\right]$$
$$\leq \mathbb{E}\left[\tilde{\ell}_{s,i}^2 | \mathcal{F}_{s-1}\right] + \mathbb{E}\left[\tilde{\ell}_{s,i^*}^2 | \mathcal{F}_{s-1}\right],$$

where the last equation holds, because for all non-negative a and b, we have: $(a-b)^2 \leq a^2 + b^2$. Then, note that:

$$E[\tilde{\ell}_{s,i}^2 | \mathcal{F}_{s-1}] \leq \frac{1}{P_{t,i}},$$

so

$$\mathbb{E}\left[(X_s)^2 | \mathcal{F}_{s-1}\right] \leq \frac{1}{P_{s,i}} + \frac{1}{P_{s,i^*}}.$$

Using the same argument as before to bound $\frac{1}{P_{s,i^*}}$ and $\frac{1}{P_{s,i}}$, we have:

$$\mathbb{P}\left[\sum_{s=1}^t \mathbb{E}\left[(X_s)^2 | \mathcal{F}_{s-1}\right] \geq \frac{2t^2\Delta_i^2}{\beta \ln t}\right] \leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \overline{\Delta_{i^*}}\right]$$
$$+ \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \overline{\Delta_{i^*}}\right].$$

Noting that the bounds on κ_t and ν_t depend on the same events, we deduce that:

$$\mathbb{P}\left[\left(\kappa_t \geq \frac{1.25t\Delta_i^2}{\beta \ln t}\right) \vee \left(\nu_t \geq \frac{2t^2\Delta_i^2}{\beta \ln t}\right)\right] \leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i} \geq \overline{\Delta_i}\right]$$
$$+ \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \Delta_{s,i^*} \geq \overline{\Delta_{i^*}}\right]$$
$$\leq \frac{1}{t}\frac{\lambda \ln K}{4K^2\beta^2},$$

where for the last step we use that for all $j, k \in V$ and $\gamma = 4$,

$$\mathbb{P}\left[\exists s \in \left[\frac{t\Delta_k^2}{\ln t}, t\right] : \Delta_{s,j} \geq \overline{\Delta_j}\right] \leq \sum_{s=\frac{t\Delta_k^2}{\ln t}}^t \mathbb{P}\left[\Delta_{s,j} \geq \overline{\Delta_j}\right]$$
$$\leq \sum_{s=\frac{t\Delta_k^2}{\ln t}}^t \frac{1}{s^3}$$
$$\leq \frac{1}{2}\left(\frac{\ln t}{t\Delta_k^2}\right)^2$$
$$\leq \frac{1}{t}\frac{\lambda \ln K}{8K^2\beta^2},$$

and the last step follows by definition of $t_{\min}$.

Control of $\tilde{\Delta}_{t,i}$ We have that:

$$\begin{aligned}\mathbb{P}\left[\tilde{\Delta}_{t,i} \leq \frac{1}{2}t\Delta_i\right] &= \mathbb{P}\left[t\Delta_i - \tilde{\Delta}_{t,i} \geq \frac{1}{2}t\Delta_i\right] \\ &\leq \mathbb{P}\left[\left(t\Delta_i - \tilde{\Delta}_{t,i} \geq \frac{1}{2}t\Delta_i\right) \wedge \left(\kappa_t \leq \frac{1.25t\Delta_i^2}{\beta \ln t}\right) \wedge \left(\nu_t \leq \frac{4t^2\Delta_i^2}{\beta \ln t}\right)\right] \\ &\quad + \mathbb{P}\left[\left(\kappa_t \geq \frac{1.25t\Delta_i^2}{\beta \ln t}\right) \vee \left(\nu_t \geq \frac{2t^2\Delta_i^2}{\beta \ln t}\right)\right]\end{aligned}$$

We set $\nu = \frac{2t^2\Delta_i^2}{\beta \ln t}$, $\kappa = \frac{1.25t\Delta_i^2}{\beta \ln t}$, $\delta = \frac{1}{Kt}$, and we recall that $\beta = 320$. Then:

$$\begin{aligned}\sqrt{2\nu \ln\left(\frac{1}{\delta}\right)} + \frac{\kappa \ln\left(\frac{1}{\delta}\right)}{3} &= \sqrt{2\frac{2t^2\Delta_i^2}{\beta \ln t}\ln(Kt)} + \frac{\frac{1.25t\Delta_i^2}{\beta \ln t}\ln(Kt)}{3} \\ &\leq \sqrt{8\frac{t^2\Delta_i^2}{\beta \ln t}\ln(t)} + \frac{\frac{1.25t\Delta_i^2}{\beta \ln t}\ln t}{3} \qquad (21) \\ &\leq t\Delta_i\left(\frac{2\sqrt{2}}{\sqrt{\beta}} + \frac{2.5}{3\beta}\right) \\ &\leq \frac{1}{2}t\Delta_i,\end{aligned}$$

where equation (21) is due to $t \geq t_{\min}(i) \geq K$, so $\ln(Kt) \leq 2\ln t$. We can then use Theorem 6 and get:

$$\mathbb{P}\left[\tilde{\Delta}_{t,i} \leq \frac{1}{2}t\Delta_i\right] \leq \frac{1}{t}\frac{\lambda \ln K}{4K^2\beta^2} + \frac{1}{Kt} = \frac{1}{t}\left(\frac{\lambda \ln K}{4K^2\beta^2} + \frac{1}{K}\right).$$

Using this bound, summing on t gives

$$\begin{aligned}\sum_{t=t_{\min}(i)}^{T}\mathbb{E}\left[q_{t,i}\right] &\leq \sum_{t=t_{\min}(i)}^{T}\mathbb{E}\left[e^{-\frac{1}{2}t\eta_t\Delta_i}\mathbb{1}\left[\tilde{\Delta}_{t,i} \geq \frac{1}{2}t\Delta_i\right] + \mathbb{1}\left[\tilde{\Delta}_{t,i} \leq \frac{1}{2}t\Delta_i\right]\right] \\ &\leq \sum_{t=t_{\min}(i)}^{T}\left(e^{-\frac{1}{2}t\eta_t\Delta_i} + \frac{1}{t}\left(\frac{\lambda \ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right),\end{aligned}$$

which finishes the proof.

D.3 Control over the Exploration

We now provide a more general version of Proposition 3.

Proposition 8. *Let $S_1, S_2, \ldots$ be a sequence of exploration sets generated by playing algorithm 1 the conditions of Lemma 1 with $\gamma = 4$, $\beta = 320$ and $\lambda \in [1, K]$. Then, the contribution of the extra exploration can be bounded as:*

$$\sum_{t=t_{\min}}^{T}\sum_{i\,:\,\Delta_i>0}\Delta_i\mathbb{E}\left[\varepsilon_{t,i}\right] \leq \sum_{t=t_{\min}}^{T}\mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0}\frac{4\beta \ln t}{t\Delta_i}\right] + \frac{\lambda \ln T \ln K}{4K\beta^2} + 12K + 3.$$

Proof. By definition of $\xi_{t,i}$, we can decompose the contribution of the extra exploration as follows.

$$\begin{aligned}\sum_{t=t_{\min}}^{T} \mathbb{E}\left[\varepsilon_{t,i}\right] \sum_{i:\Delta_i>0} \Delta_i &\leq \sum_{t=t_{\min}}^{T} \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}\left[\min\left\{1, \xi_{t,i}\right\}\right] \\ &\leq \sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \Delta_i \mathbb{E}\left[\min\left\{1, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}\right]\right] \\ &\quad + \sum_{t=t_{\min}}^{T} \sum_{i:\Delta_i>0} \Delta_i \frac{4}{t^2},\end{aligned}$$

where in the first step we use the $\min$ term to ensure that we have an upper bound on this quantity in the cases where the bounds on $\hat{\Delta}_{t,i}$ do not hold. The last step consists in upper bounding the exploration of arms that are not in S_t by adding $\frac{4}{t^2}$ to all arms, and in counting $\mathbb{E}\left[\frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right]$ only for arms i that are in the exploration set S_t.

Thus the second term is bounded as:

$$\sum_{t=t_{\min}}^{T} \sum_{i:\Delta_i>0} \Delta_i \frac{4}{t^2} \leq K \sum_{t=1}^{T} \frac{4}{t^2} \leq 8K. \tag{22}$$

In order to bound the first term, we recall that for $t \geq t_{\min}$, for any $i \in V$:

$$\begin{aligned}\mathbb{E}\left[\min\left\{1, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}\right] &\leq \mathbb{E}\left[\frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2} \mathbb{1}\left[\hat{\Delta}_{t,i} \geq \frac{1}{2}\overline{\Delta}_i\right] + \mathbb{1}\left[\hat{\Delta}_{t,i} \leq \frac{1}{2}\overline{\Delta}_i\right]\right] \\ &\leq \frac{4\beta \ln t}{t\Delta_i^2} + \mathbb{P}\left[\hat{\Delta}_{t,i} \leq \frac{1}{2}\overline{\Delta}_i\right] \\ &\leq \frac{4\beta \ln t}{t\Delta_i^2} + \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} + \frac{2}{Kt^{\gamma-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{10}} \\ &\leq \frac{4\beta \ln t}{t\Delta_i^2} + \frac{1}{t}\frac{\lambda \ln K}{4K^2\beta^2} + \frac{2}{Kt^3} + 2\left(\frac{1}{t}\right)^2,\end{aligned}$$

which gives:

$$\begin{aligned}&\sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0} \Delta_i \mathbb{E}\left[\min\left\{1, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}\right]\right] \\ &\leq \sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0} \Delta_i \left(\frac{\beta \ln t}{t\Delta_i^2} + \frac{1}{t}\frac{\lambda \ln K}{4K^2\beta^2} + \frac{2}{Kt^3} + 2\left(\frac{1}{t}\right)^2\right)\right] \\ &\leq \sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0} \frac{4\beta \ln t}{t\Delta_i}\right] + \frac{\lambda \ln T \ln K}{4K\beta^2} + 3 + 4K.\end{aligned}$$

□

D.4 Proof of Theorem 3 and Corollary 1

The proof of Theorem 3 follows from the propositions in this section.

Proof of Theorem 3. We want to bound the pseudo-regret of algorithm 1 run with parameters defined in Lemma 1 with $\gamma = 4$, $\beta = 320$ and $\lambda = \alpha$. The pseudo-regret can be decomposed by treating the

first $t_{\min}$ rounds like in the adversarial case, and by using a refined bound in the stochastic regime.

$$R_T = R_{t_{\min}} + \sum_{i\,:\,\Delta_i>0} \sum_{t=t_{\min}}^{T} \Delta_i \mathbb{E}\left[p_{t,i}\right]$$
$$\leq R_{t_{\min}} + \sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \left(\mathbb{E}\left[q_{t,i}\right] + \mathbb{E}\left[\varepsilon_{t,i}\right]\right), \tag{23}$$

First, we apply the second part of Proposition 6 with $\lambda = \alpha$, and deduce that:

$$R_{t_{\min}} \leq \frac{1019\beta K}{\Delta_{\min}^2} \left(\ln\left(\frac{\beta K}{\Delta_{\min}}\right)\right)^{3/2} + 2K. \tag{24}$$

Then we bound the contribution of exponential weights by applying Proposition 7 with $\lambda = \alpha$ and $\eta_t = \sqrt{\frac{\ln K}{2\sum_{s=K}^{t-1}\theta_s}}$. By definition of θ_s, $\sum_{s=K}^{t-1}\theta_s \leq tK$, so $\eta_t \geq \sqrt{\frac{\ln K}{2tK}}$. This gives:

$$\sum_{t=t_{\min}(i)}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \sum_{t=t_{\min}(i)}^{T} \left(e^{-\frac{1}{2}t\eta_t\Delta_i} + \frac{1}{t}\left(\frac{\alpha\ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right)$$
$$\leq \sum_{t=t_{\min}(i)}^{T} \left(e^{-\Delta_i\sqrt{\frac{\ln K}{8K}}\sqrt{t}} + \frac{1}{t}\left(\frac{\alpha\ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right)$$
$$\leq \frac{16K}{\Delta_i^2} + \ln T\left(\frac{\alpha\ln K}{4K^2\beta^2} + \frac{1}{K}\right),$$

and then:

$$\sum_{i:\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \sum_{i:\Delta_i>0} \Delta_i \left(\frac{6K}{\Delta_i^2} + \ln T\left(\frac{\alpha\ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right)$$
$$\leq \ln T\left(\frac{\alpha\ln K}{4K\beta^2} + 1\right) + \sum_{i:\Delta_i\geq 0} \frac{16K}{\Delta_i}, \tag{25}$$

where the last step follows from Lemma 5. Furthermore, we bound the contribution of the extra exploration by applying Proposition 8 with $\lambda = \alpha$, which gives:

$$\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}(i)}^{T} \mathbb{E}\left[\varepsilon_{t,i}\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0} \frac{4\beta\ln t}{t\Delta_i}\right] + \frac{\alpha\ln K}{4K\beta^2}\ln T + 12K + 3$$
$$\leq \max_{Ind\in\mathcal{I}(G)} \left\{\sum_{i\in Ind:\Delta_i>0} \frac{4\beta\ln^2 T}{\Delta_i}\right\} + \frac{\alpha\ln K}{4K\beta^2}\ln T + 12K + 3, \tag{26}$$

where the last step follows from Proposition 1: by definition, for all t, S_t is a strongly independent set on G, and we can upper bound by taking the maximum over all the strongly independent sets of G.

Finally, summing over equations (24), (25) and (26) and noting that $14K+3 \leq \frac{\beta K}{\Delta_{\min}^2}$ and $\frac{2\alpha\ln K}{4K\beta^2}+1 \leq 2\alpha$ finishes the proof. □

The Corollary 1 follows the same structure.

Proof of Corollary 1. We decompose the regret following equation (23), where t_{min} is defined using $\lambda = \tilde{\alpha}$.

$R_{t_{min}}$ is bounded the first part of Proposition 5, which gives:

$$R_{t_{min}} \leq \frac{160\beta K}{\Delta_{\min}^2}\sqrt{\frac{\tilde{\alpha}}{\lambda}}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K. \tag{27}$$

Then $\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[q_{t,i}\right]$ is bounded by Proposition 7 with $\lambda = \tilde{\alpha}$, and using the derivation leading to (25), which gives:

$$\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \ln T \left(\frac{\tilde{\alpha}\ln K}{4K\beta^2} + 1\right) + \sum_{i:\Delta_i\geq 0} \frac{16K}{\Delta_i}, \tag{28}$$

and $\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[\varepsilon_{t,i}\right]$ follows from the derivation leading to (26), which gives:

$$\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}(i)}^{T} \mathbb{E}\left[\varepsilon_{t,i}\right] \leq \max_{Ind\in\mathcal{I}(G)} \left\{ \sum_{i\in Ind:\Delta_i>0} \frac{4\beta\ln^2 T}{\Delta_i} \right\} + \frac{\tilde{\alpha}\ln K}{4K\beta^2}\ln T + 12K + 3, \tag{29}$$

Finally, summing equations (27), (28) and (29) finishes the proof. □

E Extension to Graphs that Change over Time

Proof of Theorem 4.

Adversarial Regime In the adversarial regime, the proof follows the analysis with a fixed feedback graph up to equation (12), which gives:

$$\mathcal{R}_T \leq 2K + 2\sqrt{2\ln K}\,\mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{T\ln K}, \tag{30}$$

for $\lambda = 1$. All that remains is to bound $\sum_{t=K+1}^{T}\theta_t$. For the first part of the bound, we use Lemma 8 and Proposition 4 to deduce that for all $t \geq K+1$:

$$\theta_t \leq \tilde{\alpha}_t,$$

and using this bound on θ in equation (30) gives:

$$\mathcal{R}_T \leq 4\sqrt{\ln K \sum_{t=1}^{T}\tilde{\alpha}_t} + 2K. \tag{31}$$

For the second part of the bound, we recall that for all $t \geq K+1$ the exploration parameter is lower bounded and fulfills $p_{t,i} \geq \varepsilon_{t,i} \geq \frac{4}{t^2} \geq \frac{4}{T^2}$. Thus we can apply Lemma 7 at each round $t \geq K+1$, which gives:

$$\theta_t = \sum_{i\in V}\frac{p_{t,i}}{P_{t,i}} \leq 8\alpha_t \ln\left(KT\right).$$

using this bound on θ in equation (30) gives:

$$\mathcal{R}_T \leq 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\sum_{t=1}^{T}\alpha_t} + 2K. \tag{32}$$

Taking the minimum over equations (31) and (32) finishes the proof.

Stochastic Regime

The structure of the proof follows from Theorem 3. We decompose the regret following equation (23), where t_{min} is chosen using $\lambda = 1$.

$$R_T \leq R_{t_{\min}} + \sum_{i\,:\,\Delta_i>0}\Delta_i \sum_{t=t_{\min}}^{T}\left(\mathbb{E}\left[q_{t,i}\right] + \mathbb{E}\left[\varepsilon_{t,i}\right]\right).$$

$R_{t_{min}}$ is bounded using the same approach as for Corollary 1, but using the time varying version of the bound given in equation (31). We bound $\tilde{\alpha}_t \leq K$ at each round, which gives:

$$R_{t_{min}} \leq \frac{160\beta K^{3/2}}{\Delta_{\min}^2}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K. \tag{33}$$

Then the second term is bounded by Proposition 7 with $\lambda = 1$, and using the derivation leading to equation (25), which gives:

$$\sum_{i\,:\,\Delta_i>0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \ln T \left(\frac{\ln K}{4K\beta^2} + 1\right) + \sum_{i:\Delta_i\geq 0} \frac{16K}{\Delta_i}, \tag{34}$$

The last term follows Proposition 8 with $\lambda = 1$, which gives:

$$\sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \Delta_i \mathbb{E}\left[\varepsilon_{t,i}\right]\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \frac{4\beta \ln t}{t\Delta_i}\right] + \frac{\ln T \ln K}{4K\beta^2} + 12K + 3.$$

We recall that because $\sum_{i\in S_t:\Delta_i\geq 0} \Delta_i \mathbb{E}\left[\varepsilon_{t,i}\right] \leq 1$ for all t, we can skip rounds that have the largest upper bound on S_t by upper bounding the contribution of such rounds by 1. Let $\tilde{A}_n$ be the nth largest element in the set containing the strong independence number of G_t, with $t \in [1, T]$. Then we can upper bound the first term in Proposition 8 as:

$$\sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \frac{4\beta \ln t}{t\Delta_i}\right] \leq \inf_{0\leq n\leq T} \left\{\max_{S\subset V:|S|=\tilde{A}_n} \left\{\sum_{i\in S:\Delta_i>0} \frac{4\beta \ln^2 T}{\Delta_i}\right\} + n\right\}.$$

This gives

$$\begin{aligned}\sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \Delta_i \mathbb{E}\left[\varepsilon_{t,i}\right]\right] \leq & \inf_{0\leq n\leq T} \left\{\max_{S\subset V:|S|=\tilde{\alpha}_n} \left\{\sum_{i\in S:\Delta_i>0} \frac{4\beta \ln^2 T}{\Delta_i}\right\} + n\right\} \\ & + \frac{\ln T \ln K}{4K\beta^2} + 12K + 3. \end{aligned} \tag{35}$$

We finish the proof by summing on equations (33), (34) and (35).

$$\begin{aligned} R_T \leq & \frac{160\beta K^{3/2}}{\Delta_{\min}^2} \ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K + \max_{S\subset V:|S|=\tilde{\alpha}} \left\{\sum_{i\in S:\Delta_i>0} \frac{4\beta \ln^2 T}{\Delta_i}\right\} \\ & + \ln T \left(\frac{\ln K}{2K\beta^2} + 1\right) + \sum_{i:\Delta_i>0} \frac{16K}{\Delta_i} + 12K + 3 \\ \leq & \inf_{0\leq n\leq T} \left\{\max_{S\subset V:|S|=\tilde{\alpha}_n} \left\{\sum_{i\in S:\Delta_i>0} \frac{4\beta \ln^2 T}{\Delta_i}\right\} + n\right\} \\ & + 2\ln T + \sum_{i:\Delta_i>0} \frac{16K}{\Delta_i} + \frac{161\beta K^{3/2}}{\Delta_{\min}^2} \ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right). \end{aligned}$$

□