UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Chloé Rouyer

Theoretical Foundations of Learning with Worst-Case and Easy Data

Advisor: Yevgeny Seldin

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on October 31st 2022.

Abstract

Online learning problems have received a lot of attention in the past couple of decades. In particular, multi-armed bandits have been extensively studied as they are a simple representation of the trade-off between exploration and exploitation. Originally, the literature made either the strong assumption that the environment was stochastic or no assumption about the nature of the environment and treated it as worst-case data. In practice however, it may be unreasonable to assume that the environment is stochastic and algorithms tailored for stochastic environments can fail arbitrarily badly as soon as the stochasticity assumptions are broken. Using algorithms tailored for worst-case data prevents this issue, but at the cost of a weaker performance. For those reasons, the past decade has seen the emergence of the study of algorithms that are efficient when faced with easy data while being simultaneously robust to worst-case sequences of data. These best-of-both-worlds algorithms have been extensively studied in the multi-armed bandit setting, where Zimmert and Seldin (2019) introduced the Tsallis-Inf algorithm, an algorithm capable of achieving an optimal rate against both adversarial and stochastically constrained adversarial environments.

Many online learning problems have only been studied in either the stochastic or the adversarial regime separately. Achieving best-of-both-worlds results for these problems may present extra challenges, as the conditions of the problem may impact the stochastic and the adversarial regimes differently. In this work, we consider different variations of the bandits problem where the trade-off between exploration and exploitation is affected differently in the adversarial and the stochastic regimes. We propose new algorithms and provide theoretical guarantees for their performance in both environments.

First, we consider a variant of the bandit problem where the learner can decouple exploration and exploitation by choosing one arm to play blindly and one arm to observe without suffering the associated loss in each round. We propose an algorithm based on Tsallis-Inf which recovers optimal regret guarantees in the adversarial regime and simultaneously provides a time independent regret bound in the stochastic regime, which is an improvement compared to the results for stochastic bandits whose regret scales with time.

We then consider another variation of the multi-armed bandits problem, where the learner has to pay a cost each time she decides to switch the action she plays. We propose an algorithm, based on Tsallis-Inf, that achieves the optimal rate in the adversarial regime, and an improved rate in the stochastically constrained adversarial regime. Furthermore, we generalize the analysis to sequences of switching costs that

change each time a switch is performed.

Finally, we consider the problem of online learning with feedback graphs. We propose an algorithm based on EXP3, which enjoys a near-optimal performance in both the adversarial and the stochastic regimes. We also generalize to sequences of graphs that change over time.

Resumé

Problemer indenfor online learning har fået meget opmærksomhed i de sidste årtier. Specielt har flerarmede tyveknægte været studeret eftersom de repræsenterer en afvejning mellem udforskning og udnytning. Oprindeligt lavede det meste af litteraturen den stærke antagelse at omgivelserne var stokastisk eller også lavede den ingen antagelser om omgivelserne og antog at det var worst-case data. I praksis er det imidlertid urimeligt at antage at omgivelserne er stokastisk og at algoritmer tilpasset stokastiske omgivelser fejler arbitrært dårligt lige så snart de stokastiske antagelser brydes. Ved brug af algoritmer tilpasset worst-case omgivelser undgår man dette, men på bekostning af en ringere ydeevne. Af disse årsager har det sidste årti set forskning i algoritmer der er effektive med nem data, samtidig med at de er robuste over for worst-case data. Disse algoritmer, som er det bedste af begge verdener, er blevet grundigt undersøgt i flerarmede tyveknægt-sammenhængen, hvor Zimmert et al. (2019) introducerede Tsallis-Inf-algoritmen, en algoritme i stand til at opnå optimal ydeevne over for både modstridende og stokastisk begrænsede modarbejdende omgivelser.

Mange problemer indenfor online learning er kun studeret separat i enten det stokastiske eller modarbejdende scenarie. Det kan være ekstra svært at opnå best-of-both-worlds ydeevne da problemets betingelser kan påvirke de stokastiske og modarbejdende scenarier forskelligt. Her kigger vi på forskellige variationer af tyveknægt-problemet hvor afvejningen mellem udforskning og udnyttelse påvirkes forskelligt i det modarbejdende og det stokastiske scenarie. Vi foreslår nye algoritmer og giver teoretiske garantier for deres ydeevne i både modarbejdende og stokastiske omgivelser.

Først ser vi på en variant af tyveknægt-problemet, hvor algoritmen kan afkoble udforskning og udnyttelse, ved at vælge én arm at spille blindt på og én arm til at observere, uden at lide det associerede tab hver runde. Vi foreslår en algoritme, baseret på Tsallis-Inf, som genvinder den optimale ydeevne i det modarbejdende scenarie og samtidig giver en tidsuafhængig regret-begrænsning i det stokastiske scenarie, hvilket er en forbedring sammenlignet med resultater for stokastiske tyveknægte.

Herefter betragter vi en anden variation af det flerarmede tyveknægt-problem, hvor der er en omkostning for algoritmen hver gang den beslutter sig for at skifte handling. Vi foreslår en algoritme der opnår optimal ydeevne i det modarbejdende scenarie og en forbedret ydeevne i det stokastisk begrænsede modarbejdende scenarie. Ydermere generaliserer vi analysen til sekvenser af omkostningsskift der ændrer sig hver gang der skiftes handling.

Til sidst kigger vi på online learning-problemet med feedbackgrafer. Vi fores-

lår en algoritme baseret på EXP3 som opnår en nær-optimal ydeevne i både det modarbejdende og det stokastiske regime. Vi generaliserer også disse resultater til sekvenser af grafer der ændrer sig over tid.

Acknowledgements

While I have pictured myself pursuing a Ph.D. for nearly as long as I can remember, I was far from imagining what this journey would be like. Luckily, I had the chance of being surrounded by amazing people without which this journey would not have been the same. My first word has to go to my supervisor Yevgeny Seldin, thanks to whom I discovered the fascinating world of bandits. Yevgeny has been a great support in these past few years and I am extremely grateful for everything that I have learned. My experience would not have been the same without my colleagues from the Delta group and in particular my fellow Ph.D. students Yi-Shan, Saeed, Yunlian, Hippolyte and Yijie. Both scientifically and personally, you have been an extremely valuable part of those years in Denmark. I am so proud of all the work we have accomplished and I cannot wait to see you all graduate. I am also very grateful to Nicolò Cesa-Bianchi for the chance to visit his group in Milan. I will never forget the warm welcome that I received from everyone, and I am beyond happy with the work that we produced with my awesome co-author Dirk van der Hoeven. I hope that we will see each other soon for some gelato al pistacchio salato.

Finally, this work is dedicated to my family and friends, close and afar. Your support has been invaluable and I hope that this work makes you as proud as I am.

Thank you all, merci à tous.
Chloé

Table of Contents

Chapter 1

Introduction

Online learning is a well-studied framework when it comes to sequential learning (see Slivkins (2019); Orabona (2019) and Lattimore and Svepesvári (2020) for recent surveys). This framework is often described as a game between a learner and the environment, wherein the learner is repeatedly faced with selecting an action in a list and suffering the associated loss generated by the environment. The objective of the learner is to minimize her cumulative loss, done by using feedback provided by the environment in order to improve the decisions she takes over time. This framework is simple enough to be thoroughly studied by theoreticians, and flexible enough to represent a wide range of practical problems. Nearly a century ago Thompson (1933) attempted to optimize the design of medical trials using this framework and in recent years it has become particularly well suited to handle the increasing amount of data available online. Applications include personalized advertisement, recommendation systems, investments and portfolio selection or routing problems. This framework can also be used as a building block in more complex problems such as Monte-Carlo tree search or even reinforcement learning and robotics.

There are several degrees of variation in this framework. One is the amount of feedback accessible to the learner in each round. If the learner is in a full information setting, meaning that she is allowed to observe the feedback related to all actions independently of what was played, the learner can exploit the previously gathered feedback to improve future predictions (Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006). On the opposite side of the spectrum, the learner may only be allowed to observe the feedback associated with the action that she played. In such a case, the optimal action for the learner depends on a trade-off between exploiting the previously gathered feedback, and choosing to explore a potentially less promising action in order to gather feedback about the

action. This setting is referred to as multi-armed bandits (Auer, 2002; Auer et al., 2002b).

While both full-information and bandit problems have been extensively studied in the literature, many practical problems do not simply fit into one of those categories. Constraints such as delays in observing feedback (Thune et al., 2019; Zimmert and Seldin, 2020; Masoudian et al., 2022), costs for switching between actions (Dekel et al., 2012) or access to side observations, possibly at a cost, (Seldin et al., 2014; Alon et al., 2015; Thune and Seldin, 2018) modify the balance between exploration and exploitation and require tailored solutions.

Online learning problems present another degree of variation in the assumptions made about the environment. The environment controls the choice of the loss vectors associated with the actions the learner has to choose from. Most machine learning models require the environment to behave stochastically in order to generalize results, as it is necessary for the training data to follow the same underlying distribution as future data. In online learning there is no such distinction between exploration and exploitation phases and thus it is possible to derive algorithms that are robust to sequences of losses that may change arbitrarily as the model continues to learn and evolve. The study of adversarial and stochastic online learning problems originally stayed rather separated, as they relied on drastically different approaches. However in real-life applications it may be difficult to ensure that sequentially acquired data fulfills stochasticity assumptions and, if one uses an algorithm tailored for stochastic environments with data that does not comply to that assumption, the algorithm can fail to learn anything. One solution may seem to rely on robust algorithms tailored for worst-case data. The downside to this approach is that robust algorithms may learn slowly even in the presence of easy data. Recently there has been a focus towards developing algorithms that adapt to both easy and worst-case data simultaneously and without requiring knowledge of the regime. Some of these algorithms can also have extra guarantees in intermediate regimes. This includes the stochastically constrained adversarial regime which is a generalization of the stochastic environment where the losses are generated according to underlying distributions that may change over time as long as they keep the gaps between the expected loss of each arm constant over time.

For the multi-armed bandits problem, different approaches attempted to merge algorithms for stochastic and adversarial regimes (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018), but this came at the cost of logarithmic factors in at least one of the regimes. Finally, Zimmert and Seldin (2019, 2021) introduced the Tsallis-Inf algorithm, which achieved an optimal rate for both the adversarial and the stochastically constrained

adversarial regimes simultaneously.

In the full information setting, Cesa-Bianchi et al. (2007); de Rooij et al. (2014); Gaillard et al. (2014); Sani et al. (2014); Koolen and van Erven (2015); Luo and Schapire (2015) investigated how to achieve best-of-both-worlds guarantees, and Mourtada and Gaïffas (2019) showed that the Hedge algorithm, which was optimal against adversarial sequences of losses, is also capable of achieving an optimal rate against stochastic sequences of losses when it is tuned with the appropriate learning rate.

The study of best-of-both-worlds problems now expands to many online learning problems that were previously only studied in the adversarial or the stochastic regimes separately. In the present work, we consider several online learning problems. Each of these problems is based on the multi-armed bandits framework and modifies one key aspect of the problem setting which affects the trade-off between exploration and exploitation. We provide and analyse best-of-both-worlds algorithms and gain a better understanding of the relation between the different regimes of losses and the exploration versus exploitation trade-off.

1.1 Outline of the Thesis

This thesis is structured with the following chapters.

In Chapter 2 we consider a variation of the multi-armed bandits problem where the learner is allowed to decouple exploration and exploitation by choosing two actions per round. The first action is played blindly and the loss associated with the second action is observed without being suffered by the learner. This framework can be used to represent problems where observations are queried independently of what the learner is playing. This problem was originally studied in Avner et al. (2012). The authors showed that the decoupling strategy does not help in the adversarial regime and the lower bound $\Omega(\sqrt{KT})$ of standard multi-armed bandits also holds in the decoupled setting. They provided an algorithm which enjoys near-optimal $\tilde{O}(\sqrt{KT})$ guarantees against adversarial sequences, while also achieving an improved rate of $\tilde{O}\big(\sqrt{K^{\max\{0,\frac{4}{3}-\frac{1}{3}\log_K(T)\}}T}\big)$ in the stochastic regime. We propose a new algorithm based on Follow The Regularized Leader (FTRL), which enjoys a rate-optimal $O(\sqrt{KT})$ pseudo-regret bound in the adversarial regime, while simultaneously enjoying a time-independent bound in the stochastically constrained adversarial regime, improving upon the results of Avner et al. in both regimes.

In Chapter 3 we consider another variation of the multi-armed bandits problem

where the learner has to pay an extra cost each time she decides to switch the action she plays. In the adversarial regime Dekel et al. (2013) proposed a lower bound of $\Omega\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ for all $\lambda \geq 0$, where λ is the switching cost. Dekel et al. (2012) proved an upper bound of $\mathcal{O}\big((K \ln K)^{1/3}T^{2/3}\big)$ in the case where $\lambda = 1$ for an algorithm derived from EXP3. In the stochastic regime Gao et al. (2019) and Esfandiari et al. (2021) also assumed that $\lambda = 1$ and achieved the optimal distribution-dependent regret of $\mathcal{O}\big((\ln T)\sum_{i:\,\Delta_i>0}\Delta_i^{-1}\big)$, which matches the lower bound for stochastic multi-armed bandits without switching costs (Lai and Robbins, 1985). We propose the first algorithm that adapts to both the adversarial and the stochastically constrained adversarial regimes and extend the analysis by removing the assumption that $\lambda = 1$. In the adversarial regime we derive an optimal bound $\Theta\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ for any value of $\lambda \geq 0$. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, we obtain the refined bound $\mathcal{O}\left(\big((\lambda K)^{2/3}T^{1/3} + \ln T\big)\sum_{i\neq i^*}\Delta_i^{-1}\right)$ where i^* is a unique optimal arm. Furthermore, we generalize the analysis by considering the version of the problem where the switching costs change each time a switch is taken.

In Chapter 4 we consider the problem of online learning with feedback graphs. This framework interpolates between full information and bandit problems by providing the learner with a graph that expresses what feedback the learner may observe when she plays a certain arm. This problem has been studied in the adversarial regime in Alon et al. (2015, 2017), wherein a lower bound in $\Omega(\sqrt{\alpha T})$ and a near matching $\tilde{O}(\sqrt{\alpha T})$ upper bound were derived, where α is the independence number of the graph. In the stochastic regime Buccapatnam et al. (2014, 2017) proposed a graph and problem dependent lower bound $\Omega(c^* \ln T)$ and almost matching $O(c^* \ln T) + O(K)$ upper bound. Erez and Koren (2021) propose the first best-of-both-worlds algorithm for this problem, which enjoys a $O\big(\sqrt{\chi T}\big(\ln(KT)\big)^2\big)$ 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 χ is the clique covering number of the feedback graph, which fulfills $\alpha \leq \chi$. We propose an algorithm that achieves an $\tilde{O}(\sqrt{\alpha T})$ upper bound against adversarial sequences of losses and $O\big((\ln T)^2 \max_{S\in\mathcal{I}(G)}\sum_{i\in S}\Delta_i^{-1}\big)$ against stochastic sequences of losses, improving upon the results of Erez and Koren (2021) in both regimes.

We finish this work with a discussion of these results in Chapter 5.

1.2 Main Contributions

The main contributions of this work are:

- We present an algorithm based on Tsallis-Inf for the problem of decoupled exploration and exploitation for multi-armed bandits. We analyse its regret in both the adversarial and the stochastically constrained regimes for different values of the regularization parameter α. We show that our algorithm achieves the optimal rate $\Theta\big(\sqrt{KT}\big)$ up to constants in the adversarial regime for any value of $\alpha \in (0, 1)$, while simultaneously achieving refined bounds in the stochastically constrained adversarial regime for any $\alpha \in \big(0; \frac{2}{3}\big]$.

- Crucially, we show that the algorithm requires $\alpha \in \big(\frac{1}{2}; \frac{2}{3}\big]$ to achieve an anytime bound in the stochastically constrained adversarial regime, which scales as $O\big(\sum_{i:\Delta_i>0} \frac{\sqrt{K}}{\Delta_i}\big)$. The tightest constants are obtained for $\alpha = \frac{2}{3}$.

- We consider a general version of multi-armed bandits with switching costs, where the switching cost is a constant $\lambda > 0$. We present an algorithm based on Tsallis-Inf which achieves an optimal rate $\Theta\big((\lambda K)^{1/3}T^{2/3}\big)$ up to constants in the adversarial regime, and refined guarantees scaling as $O\big(\big((\lambda K)^{2/3}T^{1/3} + \ln T\big)\sum_{i\neq i^*} \Delta_i^{-1}\big)$ in the stochastically constrained adversarial regime. To the best of our knowledge, this algorithm is the first to consider this problem in a best-of-both-worlds setting and to be analysed in the stochastically constrained adversarial regime.

- We generalize the notion of switching costs further by analysing our algorithm in the setting where switching costs may change each time a switch is taken.

- We generalize the analysis of the Tsallis-Inf algorithm to multi-armed bandits where the loss vectors are not restricted to the $[0, 1]$ interval. We consider a setting where the range of the losses is allowed to fluctuate over time as long as the learner knows the range of the loss vector at the beginning of the round.

- We propose an algorithm based on EXP3 for the problem of best-of-both-worlds online learning with feedback graphs, and show that it enjoys near-optimal regret guarantees against both adversarial and stochastic sequences of losses.

- We propose a novel exploration scheme. This greedy exploration scheme has a polynomial running time in the number of arms, while ensuring that it produces an exploration set that is both a dominating set and a strongly independent

set (which is a generalization of the independence set for directed graphs), even though computing graph related quantities such as the independence number are NP-hard problems.

- We derive a refined upper bound for undirected graphs in the adversarial regime which improve by a logarithm T factor .
- Some techniques used in our analysis improve upon the original analysis of the EXP3++ algorithm in the stochastic regime, refining a dependency on an additive $C_1 \sum_{i:\Delta_i>0} \frac{K}{\Delta_i^3}$ to a $C_2 \frac{K}{\Delta_{\min}^2}$, where C_1 and C_2 contain constants and logarithmic factors dependent on K and Δ.
- We generalize our results to sequences of graphs with a fixed set of vertices but sets of edges that change over time. We do so without affecting the run-time of the algorithm or requiring knowledge of the independence number of the graphs. This comes at the small cost of a multiplicative $\sqrt{K}$ factor in a time-independent additive factor in the bound in the stochastic regime and does not affect the bound in the adversarial regime.
- We use a skipping technique to improve the bound when a sub-logarithmic number of graphs in the sequence of feedback graphs have a significantly larger strong independence number than the rest of the graphs by handling those rounds separately in the analysis.

The presented algorithms require no prior knowledge of the nature of the environment or of the time horizon.

Chapter 2

Tsallis-INF for Decoupled Exploration and Exploitation in Multi-armed Bandits

The work presented in this chapter is based on a paper that has been published as:

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.

Abstract

We consider a variation of the multi-armed bandit problem, introduced by Avner et al. (2012), in which the forecaster is allowed to choose one arm to explore and one arm to exploit at every round. The loss of the exploited arm is blindly suffered by the forecaster, while the loss of the explored arm is observed without being suffered. The goal of the learner is to minimize the regret. We derive a new algorithm using regularization by Tsallis entropy to achieve best of both worlds guarantees. In the adversarial setting we show that the algorithm achieves the minimax optimal $O(\sqrt{KT})$ regret bound, slightly improving on the result of Avner et al.. In the stochastic regime the algorithm achieves a time-independent regret bound, significantly improving on the result of Avner et al.. The algorithm also achieves the same time-independent regret bound in the more general stochastically constrained adversarial regime introduced by Wei and Luo (2018).

2.1 Introduction

The multi-armed bandit problem is a central and most basic framework for studying the exploration-exploitation trade-off (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a,b; Slivkins, 2019; Lattimore and Svepesvári, 2020). In the multi-armed bandit game a player repeatedly chooses actions (also called arms) from a set of K actions and observes and suffers the loss of the selected action. This can be contrasted with the full information setting, where after selecting an action the player observes losses of all actions, not just the selected one (Cesa-Bianchi and Lugosi, 2006). The losses may be generated adversarially or stochastically, depending on problem setup. The goal of the learner is to find an action selection strategy minimizing the regret, which is the difference between the cumulative loss of the player and of the best fixed action in hindsight.

We focus on a variation of the multi-armed bandit problem introduced by Avner et al. (2012), in which at each round the learner is allowed to choose one action to play blindly and one action to observe without suffering its loss. The two actions may, but need not be identical. Thus, exploration is decoupled from exploitation. Practical settings having this structure are full information problems with restricted data access, where in principle the loss of any action could be accessed, but each observation, including the one of the selected action, is associated with a cost and the player can only afford one observation per round.

The decoupled setting takes an important place in the space of online learning

problems. On the one hand, it is a bridge between full information and bandit setups. In particular, as we discuss below, in the adversarial regime the problem is as hard as a bandit problem, but in the stochastic regime the regret scaling is time-independent, as in full information problems. Seldin et al. (2014) expand this bridge further by introducing multi-armed bandits with paid observations, where a learner can make an arbitrary number of observations at corresponding costs, which provides a continuous interpolation between full information and bandits. On the other hand, the decoupled setting is a bridge between exploration-exploitation and pure exploration problems (Even-Dar et al., 2006; Mannor and Tsitsiklis, 2004; Bubeck et al., 2011). In particular, one could think about doing pure exploration with the observations, but this is not an optimal strategy for the decoupled setting.

Avner et al. (2012) have shown that in the adversarial regime there is a lower bound of $\Omega(\sqrt{KT})$ for the regret in the decoupled setting. Thus, in the worst case the adversary can make the regret as large as in the standard multi-armed bandits. However, they have also shown that in some situations, in particular when one arm dominates all other arms, the regret can be reduced. More specifically, they have proposed an EXP3-style algorithm with the same exploitation strategy as EXP3, but modified exploration strategy, which achieves an $O(\sqrt{KT \ln K})$ regret bound in the worst case adversarial regime and an improved $O(\sqrt{T \ln K})$ regret bound in an adversarial regime with one dominating arm. A similar improvement in dependence on the number of arms was also shown by Seldin et al. (2014) for bandits with paid observations. Avner et al. have also analyzed their algorithm in the stochastic setting, showing that in a configuration with a single best arm (which would thus be dominating) the regret grows as $O(\sqrt{T \ln K})$. However, the analysis required a different tuning of the learning rate for the stochastic setting than for the adversarial one and, therefore, prior knowledge of the regime was essential. The stochastic regret bound was also highly suboptimal, since a simple approach of playing Follow the Leader for exploitation and uniform distribution for exploration leads to a time-independent expected regret bound of $O(\sum_{i:\Delta_i>0} \frac{K}{\Delta_i})$ in the stochastic setting.

Traditionally algorithms for multi-armed bandits and their variations, including the algorithm of Avner et al., were relying on prior knowledge of the nature of the environment, but following the work of Bubeck and Slivkins (2012) there has been a growing interest in algorithms that perform well in both settings without this knowledge (Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018). Eventually, Zimmert and Seldin (2019) have used Tsallis entropy regularizer with power $\frac{1}{2}$ to derive an algorithm that achieves the optimal regret bounds for multi-armed bandits in both settings with no prior knowledge of the regime. We follow this line of work and propose an algorithm for multi-armed bandits

with decoupled exploration and exploitation that achieves refined regret guarantees in both adversarial and stochastic regimes and requires no prior knowledge of the regime. Specifically, we make the following contributions:

- We propose a new algorithm for decoupled exploration and exploitation in multi-armed bandits based on Follow the Regularized Leader framework with regularization by Tsallis entropy.
- We show that in the adversarial regime the algorithm achieves $O(\sqrt{KT})$ regret upper bound, improving by a multiplicative factor of $\sqrt{\ln K}$ on the worst-case upper bound of Avner et al. and matching their worst-case lower bound within constants.
- We show that the same algorithm achieves a time independent $O(\sum_{i:\Delta_i>0} \frac{\sqrt{K}}{\Delta_i})$ regret bound in stochastic regime, considerably improving on the result of Avner et al.. (The result holds under a technical assumption that the best arm is unique.)
- The same regret bound is achieved in a more general stochastically constrained adversarial regime introduced by Wei and Luo (2018) (also under the assumption on uniqueness of the best arm).
- The algorithm requires no prior knowledge of the nature of the environment.
- Interestingly, the results are achieved with Tsallis entropy regulariser with power $\alpha = \frac{2}{3}$, whereas the optimal power for standard multi-armed bandits is $\alpha = \frac{1}{2}$. We show that power $\alpha = \frac{1}{2}$ does not achieve time-independent stochastic regret bounds in the decoupled setting and, therefore, inferior to $\alpha = \frac{2}{3}$.

The assumption on uniqueness of the best arm in the stochastic and stochastically constrained adversarial regimes underlies the prior work of Zimmert and Seldin (2019). We conjecture that it can be eliminated.

The paper is structured in the following way. In Section 2.3 we introduce the Follow the Regularized Leader framework and the approach used to decouple exploration and exploitation. The algorithm and main results are presented in Section 2.4 and their proofs can be found in Section 2.5. We conclude with discussion in Section 2.6.

2.2 Problem Setting and Notation

We consider a repeated game with K arms. At each round $t = 1, 2, \dots$ of the game the environment picks a loss vector $\ell_t \in [0,1]^K$ and the learner picks an action A_t to exploit and an action B_t to explore. The two actions may, but need not be identical. Then the learner blindly suffers ℓ_{t,A_t} and observes ℓ_{t,B_t} without suffering its loss.

In the adversarial setting, the environment chooses ℓ_t arbitrarily.

In the stochastic setting the losses are drawn from distributions with fixed means, i.e., for all i we have $\mathbb{E}[\ell_{t,i}] = \mu_i$ independently of t.

We also consider a more general stochastically constrained adversarial setting (Wei and Luo, 2018; Zimmert and Seldin, 2019). In this setting the losses are drawn from distributions with fixed gaps, while the baseline means are allowed to fluctuate, i.e., for all i, j we have $\mathbb{E}[\ell_{t,i} - \ell_{t,j}] = \Delta_{i,j}$ independently of t. The stochastic setting is a special case of the stochastically constrained adversary with $\Delta_{i,j} = \mu_i - \mu_j$. All results in the paper are presented for stochastically constrained adversaries and extend to stochastic environments as a special case.

We measure the performance of an algorithm in terms of pseudo-regret:

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_{t,A_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* = \arg\min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$ is the best action in hindsight. In the oblivious adversarial regime the losses are independent of the player's actions and the pseudo-regret coincides with the notion of expected regret (Bubeck and Cesa-Bianchi, 2012).

In the stochastically constrained adversarial setting we let $i^* = \arg\min_i \Delta_{i,1}$ denote an optimal arm (we can take any arm j as the second argument of $\Delta_{i,j}$ in the definition of i^*). Then we have $i_T^* = i^*$ for all T. We define $\Delta_i = \Delta_{i,i^*}$ to be the gaps to the best arm and rewrite the pseudo-regret in the stochastically constrained adversarial setting as

$$\mathcal{R}_T = \sum_{t=1}^{T} \sum_{i \neq i^*} \mathbb{E}[p_{t,i}]\, \Delta_i. \tag{2.1}$$

2.3 Decoupling Exploration and Exploitation in Follow the Regularized Leader Framework

The algorithm that we present is based on follow the regularized leader (FTRL) framework (Shalev-Shwartz, 2012). Following Zimmert and Seldin (2019), we use

the regularizer

$$\Psi_t(w) = -\frac{1}{\eta_t} \sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}, \tag{2.2}$$

which is a slight modification of the negative Tsallis entropy with power α defined by $H_\alpha(w) := \frac{1}{1-\alpha}(1 - \sum_i w_i^\alpha)$ (Tsallis, 1988). We focus on $\alpha \in (0,1)$, but one of the interesting properties of the above regularizer is that in the limits $\alpha \to 0$ and $\alpha \to 1$ it recovers the log-barrier and the negative entropy regularizers, respectively (Zimmert and Seldin, 2019). In particular, the EXP3 algorithm with losses (Auer et al., 2002b; Bubeck and Cesa-Bianchi, 2012) can be seen as a limit case of FTRL with regularization by Tsallis entropy with $\alpha \to 1$. As we are in a bandit setting and only observe one element of the loss vector at each round, we construct an unbiased estimate $\tilde{\ell}_t$ of the loss vector ℓ_t by using importance-weighted sampling

$$\forall t \in [T], i \in [K], \qquad \tilde{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\left[B_t = i\right]}{q_{t,i}},$$

where q_t is the distribution for sampling the exploratory action B_t and $\mathbb{1}$ is the indicator function.

We define the Decoupled-Tsallis-INF algorithm for an arbitrary exploration distribution q_t in Algorithm 1.

Algorithm 1: Decoupled-Tsallis-INF

Input : Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$.
Initialize: $\tilde{L}_0 = \mathbf{0}_K$
for $t = 1, 2, \ldots$ **do**
  $p_t = \arg\min_{p \in \Delta^{K-1}} \left\{ \left\langle p, \tilde{L}_{t-1} \right\rangle - \frac{1}{\eta_t} \sum_{i=1}^K \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)} \right\}$
  Construct exploration distribution q_t
  Sample A_t according to p_t, play it and suffer ℓ_{t,A_t}.
  Sample B_t according to q_t and observe ℓ_{t,B_t}.
  $\forall\, i \in [K]: \quad \tilde{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}\{B_t = i\}}{q_{t,i}} = \begin{cases} \frac{\ell_{t,i}}{q_{t,i}}, & \text{if } B_t = i, \\ 0, & \text{otherwise.} \end{cases}$
  $\forall\, i \in [K]: \quad \tilde{L}_t(i) = \tilde{L}_{t-1}(i) + \tilde{\ell}_{t,i}.$
end

In order to analyse the algorithm, we decompose the pseudo-regret into a stability and penalty components (Lattimore and Svepesvári, 2020; Zimmert and Seldin,

2019),

$$\mathcal{R}_T = \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right]}_{\text{stability}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right]}_{\text{penalty}}, \tag{2.3}$$

where i_T^* is the best action in hindsight, and the potential function is defined by

$$\Phi_t(-L) = \max_{w \in \Delta^{K-1}} \left\{ \langle w, -L \rangle + \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)} \right\}.$$

In order to achieve a tight bound on the pseudo-regret, one has to derive tight bounds on the stability and penalty. Recall that $\tilde{L}_t$ is an unbiased estimate of L_t and observe that the penalty term does not depend on the query distribution q_t. The stability term of the regret of Algorithm 1 satisfies the following lemma.

Lemma 2.1. *For any $\alpha \in (0,1)$ and any positive learning rate value, the stability term of the regret of Decoupled-Tsallis-INF with an arbitrary exploration distribution q_t satisfies:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2} \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right].$$

A proof of the lemma is provided in Section 2.7.2.2. We can see that the bound depends on the choice of exploration distribution q_t, and that picking $q_t = p_t$ recovers the bound for the stability term of the regret of Tsallis-INF for multi-armed bandits (Zimmert and Seldin, 2019). In the decoupled case we have the freedom of picking $q_t \neq p_t$, so we select the distribution q_t which minimizes the bound on the stability term in Lemma 2.1.

Lemma 2.2. *The right hand side of the bound in Lemma 2.1 is minimized by the distribution q_t defined by*

$$\forall t \in [T], i \in [K], \quad q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{j=1}^{K} (p_{t,j})^{1-\alpha/2}}.$$

We provide a proof of this Lemma in Section 2.7.3. In the previous section, we have mentioned that in the limit of $\alpha \to 1$ Tsallis-INF converges to the EXP3 algorithm. By taking $\alpha = 1$ in Lemma 2.2 we recover the exploration distribution used by Avner et al. (2012): $q_{t,i} = \sqrt{p_{t,i}/\|p_t\|_{1/2}}$.

2.4 Main Results

In the rest of the paper, we consider that the exploration distribution q_t of Decoupled-Tsallis-INF is the one defined in Lemma 2.2. In this section, we present bounds on pseudo-regret of that algorithm. The first theorem bounds the regret of Decoupled-Tsallis-INF in adversarial regime.

Theorem 2.1. *In the adversarial regime, for any $\alpha \in (0,1)$ the pseudo-regret of Decoupled-Tsallis-INF with learning rate $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$ and with q_t given by Lemma 2.2 satisfies:*

$$\mathcal{R}_T \leq \left(2 + \frac{1}{2\alpha(1-\alpha)}\right)\sqrt{KT} + 1.$$

We provide a proof of the theorem in Section 2.5.1. Avner et al. (2012) have derived a regret lower bound of $\Omega(\sqrt{KT})$ for the adversarial regime, which means that our algorithm is minimax optimal within constants. For comparison, the algorithm proposed by Avner et al. is suboptimal by a multiplicative factor of $\sqrt{\log K}$. In the next theorem we bound the regret of Decoupled-Tsallis-INF in the stochastically constrained adversarial setting.

Theorem 2.2. *In the stochastically constrained adversarial regime with a unique best action i^*, the pseudo-regret of Decoupled-Tsallis-INF with $\alpha \in (0; 2/3]$, $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$ and with q_t given by Lemma 2.2 satisfies*

$$\mathcal{R}_T \leq O\left(\left(\sum_{i\neq i^*}\sum_{t=T_0+1}^{T} \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{\frac{1}{2(\alpha-1)}}}\right) + \frac{\sqrt{K}}{\Delta_{\min}}\right),$$

where $T_0 = \max_{i\neq i^*}\left\lceil \left(\frac{8}{\Delta_i}\right)^2 \right\rceil$. This bound is time-independent if and only if $\alpha > 1/2$. When $\alpha \in (1/2, 2/3]$ the pseudo-regret satisfies,

$$\mathcal{R}_T \leq \sum_{i\neq i^*}\left(C(\alpha)\frac{\sqrt{K}}{\Delta_i}\right) + \frac{68\sqrt{K}}{\Delta_{\min}} + 11\sqrt{K},$$

where

$$C(\alpha) = \frac{2-2\alpha}{2\alpha-1}\left(\frac{1}{2\alpha(1-\alpha)} + \frac{\left(1-\frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}}+1}{2^{-1+\alpha/2}} + 2\right)^{\frac{1}{1-\alpha}}\left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right)8^{\frac{2\alpha-1}{\alpha-1}}.$$

A proof of the theorem is given in Section 2.5.2. The assumption on uniqueness of the best arm is a technical detail required in the analysis. The same assumption had to be used by Zimmert and Seldin (2019) in the analysis of Tsallis-INF. We conjecture that the assumption can be eliminated. The function $C(\alpha)$ is well defined on the interval $(1/2, 2/3]$, and numerical evaluation shows that it is monotonically decreasing on the interval and minimized by $\alpha = 2/3$. It is possible to derive a pseudo-regret bound for $\alpha \in (2/3, 1)$, however, for $\alpha > 2/3$ the dependency on K scales with $K^{2-1/\alpha} > \sqrt{K}$, without achieving a refined dependency on neither t nor Δ.

The following corollary combines adversarial and stochastically constrained adversarial analysis (and stochastic regime as a special case of the latter).

Corollary 2.1. *With $\alpha = 2/3$, $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}} = \frac{2K^{-1/6}}{\sqrt{t}}$, and q_t given by Lemma 2.2 Decoupled-Tsallis-INF achieves pseudo-regret bounds*

$$\mathcal{R}_T \leq 5\sqrt{KT} + 1$$

in the adversarial regime and

$$\mathcal{R}_T \leq \sum_{i \neq i^*} \frac{20\sqrt{K}}{\Delta_i} + \frac{68\sqrt{K}}{\Delta_{\min}} + 11\sqrt{K}$$

in stochastically constrained adversarial regimes with a unique best arm i^*. The two regret bounds hold simultaneously and with no need in prior knowledge of the regime.

The result is a direct application of Theorem 2.1 and the second part of Theorem 2.2. We note that unlike in the multi-armed bandit case, where $\alpha = 1/2$ is the optimal value both for adversarial and stochastically constrained adversarial regime, in the decoupled case there is a trade-off between the optimal values of α in the two regimes. However, the price of switching from the optimal $\alpha = 1/2$ to $\alpha = 2/3$ in the former is a minor multiplicative factor of $\frac{5}{4}$, whereas in the latter choosing $\alpha = 2/3$ allows to get rid of the dependency of the regret on the time horizon.

2.5 Proofs of the Theorems

Using the decomposition of the regret presented in Equation (2.3), we present bounds for the stability and penalty terms. We take advantage of the decoupling to refine the bound of Zimmert and Seldin (2019) on the stability term. The penalty term takes no advantage of the decoupling and we reuse the bound derived by Zimmert and Seldin.

Lemma 2.3. *For any $\alpha \in (0,1)$ and any positive learning rate value, the stability term of the regret bound of Decoupled-Tsallis-INF with q_t given by Lemma 2.2 satisfies:*

1. $$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \sum_{t=1}^{T} \frac{\eta_t}{2} K^{\alpha}.$$

2. *If further $\eta_t \leq \frac{1}{4}$, then for any j:*

$$\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \frac{\eta_t}{2}\left(K^{\alpha/2} + c(\alpha) + 1\right) \sum_{i \neq j} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2},$$

where $c(\alpha) = \left(1 - \frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}}$.

We present a proof of the lemma in Section 2.7.2.2. Note that for $\alpha \in (0,1)$, we have $c(\alpha) \in [1,2]$.

Lemma 2.4. *For any $\alpha \in (0,1)$, and non-increasing learning rate sequence η_t, the penalty term of the regret bound of Decoupled-Tsallis-INF with q_t given by Lemma 2.2 satisfies:*

1. $\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right] \leq \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{(1-\alpha)\alpha\eta_T} + 1.$

2. *Furthermore, if $\eta_t = \frac{2\beta}{\sqrt{t}}$ for some $\beta > 0$, then the penalty further satisfies:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right] \leq \frac{\sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}} + K^{1-\alpha}}{4\alpha(1-\alpha)\beta}.$$

A proof of the lemma is in Section 2.7.2.3.

2.5.1 Proof of Theorem 2.1

The proof of the theorem is based on application of the first parts of Lemmas 2.3 and 2.4.

Proof of Theorem 2.1. We use the first part of Lemma 2.3 to bound the stability term. We remind that the learning rate is $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$.

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \sum_{t=1}^{T} \frac{\eta_t}{2} K^\alpha = \sum_{t=1}^{T} \frac{K^{1/2-\alpha}}{\sqrt{t}} K^\alpha \leq 2\sqrt{KT}.$$

Similarly, we use the first part of Lemma 2.4 to bound the penalty term:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \Phi_t(-\tilde{L}_{t-1}) - \Phi_t(-\tilde{L}_t) - \ell_{t,i_T^*}\right] \leq & \frac{(K^{1-\alpha}-1)(1-T^{-\alpha})}{\alpha(1-\alpha)2K^{1/2-\alpha}\sqrt{\frac{1}{T}}} + 1 \\
\leq & \frac{1}{2\alpha(1-\alpha)} \frac{K^{1-\alpha}}{K^{1/2-\alpha}\sqrt{\frac{1}{T}}} + 1 \\
\leq & \frac{1}{2\alpha(1-\alpha)} \sqrt{KT} + 1.
\end{aligned}$$

Summing the stability and the penalty terms finishes the proof.

□

2.5.2 Proof of Theorem 2.2

The following two lemmas are needed in order to take advantage of the self-bounding technique and obtain a time-independent bound. They are proven in Section 2.7.1.

Lemma 2.5. *For $\alpha \in (0,1)$, $c > 0$ and $d \in (0,1]$, we have*

$$\max_{x \in [0,\infty)} cx^\alpha - dx = c^{\frac{1}{1-\alpha}} d^{\frac{\alpha}{\alpha-1}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right).$$

Lemma 2.6. *Let $T_0 = \max_{i \neq i^*} \left\lceil \left(\frac{8}{\Delta_i}\right)^2 \right\rceil$ and $S(T) = \frac{1}{\Delta_i^{-\frac{\alpha}{\alpha-1}}} \sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$ for any $i \neq i^*$. The series $S(T)$ converges for $T \to \infty$ if and only if $\alpha > \frac{1}{2}$. Furthermore, for $\alpha > \frac{1}{2}$, we have:*

$$\lim_{T\to\infty} S(T) \leq \frac{2-2\alpha}{2\alpha-1} \frac{8^{\frac{2\alpha-1}{\alpha-1}}}{\Delta_i}.$$

With the two lemmas at hand we move on to the proof. The proof strategy is the following. We define a time step T_0 from which we can achieve a refined upper bound for the instantaneous stability term introduced in Lemma 2.3. For $t \leq T_0$, the

proof is the same as in the adversarial setting, which gives a contribution of order $O(\sqrt{KT_0}) = O\left(\frac{\sqrt{K}}{\Delta_{\min}}\right)$.

Now, we focus on the part of the bound for $t > T_0$. Let B be an upper bound on the regret, $\mathcal{R}_T \leq B$. In the stochastically constrained adversarial regime we can use the alternative way of writing the regret given in equation (2.1) (what Zimmert and Seldin call the self-bounding property of the regret) to obtain

$$\mathcal{R}_T = 2B - \sum_{t=1}^{T} \sum_{i \neq i^*} \mathbb{E}[p_{t,i}]\Delta_i.$$

For $t > T_0$ we derive a refined bound for instantaneous contributions to the right hand side. By using the second parts of Lemmas 2.3 and 2.4 the instantaneous contributions to B can be bounded by $\sum_{i \neq i^*} C\mathbb{E}\left[p_{t,i}\right]^{\alpha} / \sqrt{t}$ for some constant C when $\alpha \leq 2/3$. The overall instantaneous contribution to the right hand side is then bounded by $\sum_{i \neq i^*} \left(2C\mathbb{E}\left[p_{t,i}\right]^{\alpha} / \sqrt{t} - \mathbb{E}[p_{t,i}]\Delta_i\right)$. By taking $x_i = \mathbb{E}\left[p_{t,i}\right]$ and using Lemma 2.5 we then bound the instantaneous contributions by

$$\sum_{i \neq i^*} \left(\frac{2Cx_i^{\alpha}}{\sqrt{t}} - \Delta_i x_i\right) \leq \sum_{i \neq i^*} \max_{x_i \in [0,\infty)} \left(\frac{2Cx_i^{\alpha}}{\sqrt{t}} - \Delta_i x_i\right) \leq \sum_{i \neq i^*} C' \Delta_i^{\frac{\alpha}{\alpha-1}} t^{\frac{1}{2(\alpha-1)}}$$

for some other constant C'. Note that the bound is meaningful only for $\Delta_i > 0$. This is why we need the assumption on uniqueness of the best arm. Summing the instantaneous contributions over t from T_0 to T completes the proof. The sum of the series of instantaneous contributions converges if and only if $\frac{1}{2(1-\alpha)} > 1$, which means that the bound is time independent if and only if $\alpha > \frac{1}{2}$.

Proof of Theorem 2.2. We bound the stability and the penalty terms. Concerning the stability term, we want to use the second part of Lemma 2.3 when t is large enough. We choose the threshold $T_0 = \max_{i \neq i^*} \left\lceil \left(\frac{8}{\Delta_i}\right)^2 \right\rceil \geq 64$. This choice allows us to use the second part of Lemma 2.3 because for all $t > T_0$ we have $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}} \leq \frac{2}{\sqrt{t}} \leq \frac{1}{4}$. We use the second part Lemma 2.3 where we select $j = i^*$ when $t > T_0$, and we use the first part of Lemma 2.3 when $t \leq T_0$.

Thus, we have:

$$\begin{aligned}
\text{stability} &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
&= \sum_{t=1}^{T_0} \mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
&\quad + \sum_{t=T_0+1}^{T} \mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
&\leq \sum_{t=1}^{T_0} \frac{K^{1/2-\alpha}}{\sqrt{t}} K^{\alpha} + \sum_{t=T_0+1}^{T} \frac{K^{1/2-\alpha}}{\sqrt{t}} \left(K^{\alpha/2} + c(\alpha) + 1\right) \left(\sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2}\right) \\
&\leq 2\sqrt{KT_0} + \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\left(\frac{c(\alpha)+1}{2^{\alpha/2}} + 1\right) K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2},
\end{aligned}$$

where we used that $\sum_{t=1}^{T_0} \frac{1}{\sqrt{t}} \leq 2\sqrt{T_0}$, and $K^{\alpha/2} \geq 2^{\alpha/2}$.

To bound the penalty term, we use the second part of Lemma 2.4 with $\beta = K^{1/2-\alpha}$.

$$\begin{aligned}
\text{penalty} &\leq \frac{1}{4\alpha(1-\alpha)K^{1/2-\alpha}} \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}} + \frac{\sqrt{K}}{4\alpha(1-\alpha)} \\
&\leq \frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \left(\sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}} + \sum_{i \neq i^*} \sum_{t=1}^{T_0} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}}\right) + \frac{\sqrt{K}}{4\alpha(1-\alpha)} \\
&\leq \left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}}\right) + \left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{t=1}^{T_0} \frac{K^{1-\alpha}}{\sqrt{t}}\right) + \frac{\sqrt{K}}{4\alpha(1-\alpha)} \\
&\leq \left(\frac{K^{\alpha-1/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}}\right) + \frac{\sqrt{KT_0}}{2\alpha(1-\alpha)} + \frac{\sqrt{K}}{4\alpha(1-\alpha)},
\end{aligned}$$

where we have used that $\sum_{i \neq i*} \mathbb{E}\left[p_{t,i}\right]^{\alpha} \leq (K-1)\left(\frac{1}{K-1}\right)^{\alpha} = (K-1)^{1-\alpha} \leq K^{1-\alpha}$ and $\sum_{t=1}^{T_0} \frac{1}{\sqrt{t}} \leq 2\sqrt{T_0}$.

We make two observations regarding the powers: for $\alpha \leq 2/3$ we have $1/2 - \alpha/2 \geq \alpha - 1/2$, so for all $K \geq 2$ it holds that $K^{\alpha-1/2} \leq K^{1/2-\alpha/2}$. Furthermore, for $\alpha \leq 2/3$ we have $\alpha \leq 1 - \alpha/2$, and as for all $t \in [T]$ and $i \in [K]$ we have $\mathbb{E}[p_{t,i}] \leq 1$ and

$\mathbb{E}[p_{t,i}]^{1-\alpha/2} \leq \mathbb{E}[p_{t,i}]^{\alpha}$. Thus, we have:

$$\text{stability} \leq 2\sqrt{KT_0} + \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\left(\frac{c(\alpha)+1}{2^{\alpha/2}} + 1\right) K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}\left[p_{t,i}\right]^{\alpha},$$

and

$$\text{penalty} \leq \left(\frac{K^{1/2-\alpha/2}}{4\alpha(1-\alpha)} \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}[p_{t,i}]^{\alpha}}{\sqrt{t}} \right) + \frac{\sqrt{KT_0}}{2\alpha(1-\alpha)} + \frac{\sqrt{K}}{4\alpha(1-\alpha)}.$$

We combine the two bounds and use alternative way of writing the regret in the stochastically constrained adversarial regime (the self-bounding technique) to obtain

$$\begin{aligned}
\mathcal{R}_T &= 2\mathcal{R}_T - \sum_{i \neq i*} \sum_{t=1}^{T} \mathbb{E}[p_{t,i}]\Delta_i \\
&\leq \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} \mathbb{E}[p_{t,i}]^{\alpha} - \mathbb{E}[p_{t,i}]\Delta_i \right) \\
&\quad + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{2\alpha(1-\alpha)} \\
&\leq \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \max_{x \in [0,\infty)^K} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} x_i^{\alpha} - x_i\Delta_i \right) \\
&\quad + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{2\alpha(1-\alpha)}.
\end{aligned}$$

In the last step we take $x_i = \mathbb{E}\left[p_{t,i}\right]$ and drop the constraint that p_t is a probability distribution. Using Lemma 2.5, for any $i \neq i^*$ and $t > T_0$, we have

$$\begin{aligned}
&\max_{x \in [0,\infty)} \left(\left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right) \frac{K^{1/2-\alpha/2}}{\sqrt{t}} x^{\alpha} - x\Delta_i \right) \\
&\qquad \leq \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{-\frac{1}{2(\alpha-1)}}} \left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right)^{\frac{1}{1-\alpha}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}} \right).
\end{aligned}$$

Using $\widetilde{C}(\alpha) = \left(\frac{1}{2\alpha(1-\alpha)} + \frac{c(\alpha)+1}{2^{(\alpha/2)-1}} + 2 \right)^{\frac{1}{1-\alpha}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}} \right)$, we can incorporate this result in the regret bound and deduce that:

$$\mathcal{R}_T \leq \sum_{i \neq i^*} \sum_{t=T_0+1}^{T} \left(\widetilde{C}(\alpha) \Delta_i^{\frac{\alpha}{\alpha-1}} \frac{\sqrt{K}}{t^{-\frac{1}{2(\alpha-1)}}} \right) + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{2\alpha(1-\alpha)},$$

which gives the first statement of the lemma. Finally, we use Lemma 2.6 to deduce that the bound is time-independent if and only if $\alpha > 1/2$. Furthermore, by definition of T_0 we have $\sqrt{T_0} \leq \frac{8}{\Delta_{\min}} + 1$.

We deduce that when $\alpha \in (1/2, 2/3]$, the pseudo-regret is upper bounded as:

$$\begin{aligned}
\mathcal{R}_T \leq & \sum_{i \neq i^*} \left(\widetilde{C}(\alpha) \frac{2-2\alpha}{2\alpha-1} 8^{\frac{2\alpha-1}{\alpha-1}} \frac{\sqrt{K}}{\Delta_i} \right) + \left(\frac{1}{\alpha(1-\alpha)} + 4 \right) \sqrt{KT_0} + \frac{\sqrt{K}}{2\alpha(1-\alpha)} \\
\leq & \sum_{i \neq i^*} \left(C(\alpha) \frac{\sqrt{K}}{\Delta_i} \right) + 68 \frac{\sqrt{K}}{\Delta_{\min}} + 11\sqrt{K},
\end{aligned}$$

where $C(\alpha) = \widetilde{C}(\alpha) \frac{2-2\alpha}{2\alpha-1} 8^{\frac{2\alpha-1}{\alpha-1}}$. This gives the second statement of the theorem. □

2.6 Discussion

We have derived an algorithm for the problem of decoupled exploration and exploitation in multi-armed bandits. We have shown that it achieves the minimax optimal $O(\sqrt{KT})$ regret bound in the adversarial regime and simultaneously a time-independent $O\left(\sum_{i \neq i^*} \frac{\sqrt{K}}{\Delta_i}\right)$ regret bound in the stochastically constrained adversarial regime. The results improve on the work of Avner et al. (2012) in both regimes without requiring prior knowledge of the regime.

As we have mentioned, the decoupled setting is an important bridge between full information and bandit problems, as well as a bridge between pure exploration and exploration-exploitation trade-off. An interesting direction for future research would be to use our techniques to improve results along these two directions. One possibility is to apply our regularization and exploration technique to tighten best of both worlds guarantees for prediction with limited advice (Seldin et al., 2014; Thune and Seldin, 2018). Another direction is to explore the relations with pure exploration problems. Abbasi-Yadkori et al. (2018) have shown that in the pure exploration setting it is impossible to achieve simultaneous optimality in both adversarial and stochastic settings. An interesting question is whether the decoupled formulation can be used to reformulate the objective and to achieve some alternative results there.

2.7 Appendix

2.7.1 Proofs of Auxiliary Lemmas

Proof of Lemma 2.5. Let $f(x) = cx^\alpha - dx$. We can calculate its first and second order derivatives and get $f'(x) = \alpha c x^{\alpha-1} - d$ and $f''(x) = \alpha(\alpha-1)cx^{\alpha-2} \leq 0$. Thus the solution of $f(x) = 0$ give the maximum of f.

$$f'(\tilde{x}) = 0 \Leftrightarrow \alpha c \tilde{x}^{\alpha-1} = d \Leftrightarrow \tilde{x} = \left(\frac{d}{\alpha c}\right)^{\frac{1}{\alpha-1}}.$$

Finally, we calculate $f(\tilde{x})$.

$$\max_{x\in[0,\infty)} cx^\alpha - dx = f(\tilde{x}) = c\left(\frac{d}{\alpha c}\right)^{\frac{\alpha}{\alpha-1}} - d\left(\frac{d}{\alpha c}\right)^{\frac{1}{\alpha-1}} = c^{\frac{1}{1-\alpha}} d^{\frac{\alpha}{\alpha-1}} \left(\alpha^{\frac{\alpha}{1-\alpha}} - \alpha^{\frac{1}{1-\alpha}}\right).$$

□

Proof of Lemma 2.6. Consider the series $s(T) = \sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$. We first show that when $\alpha > 1/2$, the series converges and upper bound its limit. Then, we show that the series diverges when $\alpha \leq 1/2$.

When $\alpha > 1/2$, we have $\frac{-1}{2(\alpha-1)} > 1$ so the Riemann's series $\sum_{t=1}^{\infty} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}}$ converges. This is an upper bound on $s(T)$, which converges as well. However, when we derive the upper bound on $\lim_{T\to\infty} s(T)$, we want to take advantage of the fact that we sum for $t \geq T_0 + 1$. We have:

$$\begin{aligned}
\lim_{T\to\infty} \sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} &\leq \lim_{T\to\infty} \int_{T_0}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} dt \\
&= \lim_{T\to\infty} \frac{T^{1+\frac{1}{2(\alpha-1)}} - T_0^{1+\frac{1}{2(\alpha-1)}}}{1+\frac{1}{2(\alpha-1)}} \\
&\leq \frac{-T_0^{1+\frac{1}{2(\alpha-1)}}}{1+\frac{1}{2(\alpha-1)}}
\end{aligned}$$

where in the last step we use the fact that $1 + \frac{1}{2(\alpha-1)} = \frac{2\alpha-1}{2\alpha-2}$ is negative. This also imply that as $T_0 \geq \left(\frac{8}{\Delta_{min}}\right)^2$, we can upper bound $T_0^{1+\frac{1}{2(\alpha-1)}}$ by $\left(\left(\frac{8}{\Delta_{min}}\right)^2\right)^{\frac{2\alpha-1}{2\alpha-2}} =$

$\left(\frac{8}{\Delta_{min}}\right)^{\frac{2\alpha-1}{\alpha-1}} \leq \left(\frac{8}{\Delta_i}\right)^{\frac{2\alpha-1}{\alpha-1}}$ for any $i \neq i^*$, because $\frac{2\alpha-1}{2(\alpha-1)} \leq 0$ when $\alpha > 1/2$. Incorporating this result in $S(T)$ finishes this part of the proof.

For the second part of the proof, when $\alpha \leq 1/2$, we have:

$$\begin{aligned}\sum_{t=T_0+1}^{T} \frac{1}{t^{-\frac{1}{2(\alpha-1)}}} &\geq \sum_{t=T_0+1}^{T} \frac{1}{t} \\ &\geq \int_{T_0+1}^{T+1} \frac{1}{t} dt \\ &= \log(T+1) - \log(T_0+1),\end{aligned}$$

which diverges because T_0 is a constant. Thus for any $\alpha \in (0, 1/2]$, the we cannot obtain a time independent upper bound on $S(T)$. □

2.7.2 Analysing the Follow the Regularized Leader framework

We first introduce some tools needed to work in this framework and then derive bounds on the stability and the penalty terms.

2.7.2.1 Follow the Regularized Leader and Tsallis Entropy

Follow the Regularized Leader (FTRL) has been widely used in online learning in the past few years. We use Tsallis entropy as our regularizer, defined as:

$$\Psi_t(w) = -\frac{1}{\eta_t} \sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)}, \tag{2.4}$$

and its convex conjugate is defined as:

$$\Psi_t^*(y) = \max_{x \in \mathbb{R}^K} \{\langle x, y\rangle - \Psi_t(x)\} = \max_{x \in \mathbb{R}^K} \left\{ \langle x, y\rangle + \frac{1}{\eta_t} \sum_i \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)} \right\}.$$

We let Δ^{K-1} denote a probability simplex over K vertices and define $\mathcal{I}_{\Delta^{K-1}}(x) = \begin{cases} 0, & \text{if } x \in \Delta^{K-1} \\ \infty, & \text{otherwise} \end{cases}$. Using results from convex analysis (Rockafellar, 1970), Ψ_t is a convex differentiable function with an invertible gradient $(\nabla\Psi)^{-1}$ so we have

$$\nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(y) = \operatorname*{argmax}_{x \in \Delta^{K-1}} \{\langle x, y\rangle - \Psi_t(x)\}.$$

Note that $\nabla(\Psi + \mathcal{I}_{\Delta^{K-1}})^*(y) \in \Delta^{K-1}$. We define the potential function Φ_t as

$$\Phi_t := (\Psi_t + \mathcal{I}_{\Delta^{K-1}})^* = \max_{w \in \Delta^{K-1}} \left\{ \langle w, -L \rangle + \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{w_i^\alpha - \alpha w_i}{\alpha(1-\alpha)} \right\}.$$

This means that Φ_t is the restriction of Ψ_t^* on the probability simplex. Furthermore, the weights p_t of the Decoupled-Tsallis-INF algorithm fulfil:

$$\begin{aligned} p_t = \nabla \Phi_t(-\tilde{L}_{t-1}) &= \arg\max_{p \in \Delta^{K-1}} \left\{ \left\langle p, -\tilde{L}_{t-1} \right\rangle + \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)} \right\} \\ &= \arg\min_{p \in \Delta^{K-1}} \left\{ \left\langle p, \tilde{L}_{t-1} \right\rangle - \frac{1}{\eta_t} \sum_{i=1}^{K} \frac{p_i^\alpha - \alpha p_i}{\alpha(1-\alpha)} \right\}. \end{aligned}$$

2.7.2.2 Analysing the Stability term

The analysis of the stability term follows from Zimmert and Seldin (2019, Lemma 11) once we have identified that even if the algorithm does not observe the loss of action A_t, we can use the fact that $\tilde{\ell}_t$ is an unbiased estimate of ℓ_t to deduce that:

$$\mathbb{E}\left[\ell_{t,A_t}\right] = \mathbb{E}\left[\mathop{\mathbb{E}}_{B_t \sim q_t}\left[\tilde{\ell}_{t,A_t}\right]\right] = \mathbb{E}\left[\mathop{\mathbb{E}}_{B_t \sim q_t}\left[\left\langle p_t, \tilde{\ell}_t \right\rangle\right]\right] = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t \right\rangle\right], \tag{2.5}$$

which follows from $\mathop{\mathbb{E}}_{B_t \sim q_t}\left[\tilde{\ell}_{t,A_t}\right] = \ell_{t,A_t}$ as it is an unbiased estimate, A_t being sampled according to p_t, and using the law of total expectation in the last step. Using this transformation, the analysis of the instantaneous stability using tools from convex analysis in Zimmert and Seldin (2019, Lemma 11) gives that:

Lemma 2.7. *Using that $p_t = \nabla\Phi_t(-\tilde{L}_{t-1})$ where $\tilde{L}_t = \tilde{L}_{t-1} + \tilde{\ell}_t$ for some $\tilde{\ell}_t$ unbiased estimate of ℓ_t. For any $x \in [0, \infty)$, the instantaneous stability of the pseudo-regret of Algorithm 1 satisfies*

$$\begin{aligned} &\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\ \leq\ &\mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p}_i \in [p_{t,i}, \nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x1_K)_i]} \frac{\eta_t}{2} (\tilde{\ell}_{t,i} - x)^2 (\tilde{p}_i)^{2-\alpha}\right]. \end{aligned}$$

Proof of Lemma 2.7.

$$\begin{aligned}
&\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t\right\rangle + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t\right\rangle + \Phi_t(\nabla\Psi_t(p_t) - \tilde{\ell}_t) - \Phi_t(\nabla\Psi_t(p_t))\right] \\
&\qquad = \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t - x\mathbf{1}_K\right\rangle + \Phi_t(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K) - \Phi_t(\nabla\Psi_t(p_t))\right] \\
&\qquad \leq \mathbb{E}\left[\left\langle p_t, \tilde{\ell}_t - x\mathbf{1}_K\right\rangle + \Psi_t^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K) - \Psi_t^*(\nabla\Psi_t(p_t))\right] \qquad (2.6) \\
&\qquad \leq \mathbb{E}\left[D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))\right]
\end{aligned}$$

where the first step follows from Equation (2.5). Then, we used the fact that $-\tilde{L}_{t-1} = \nabla\Psi_t(p_t)$ by definition, and the definition of Φ_t to add $x\mathbf{1}_K$. Finally, we recall that Ψ_t^* is the unrestricted version of Φ_t. On step 2.6, we recognize the Bregman divergence of Ψ_t, and using Taylor's expansion, there is some $z \in \text{conv}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))$ such that

$$D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t)) = \frac{1}{2}\|\tilde{\ell}_t - x\mathbf{1}_K\|^2_{\nabla^2\Psi_t(z)}.$$

We deduce that:

$$\begin{aligned}
&\mathbb{E}\left[D_{\Psi_t^*}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))\right] \\
&\qquad \leq \mathbb{E}\left[\max_{z \in \text{conv}(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x\mathbf{1}_K, \nabla\Psi_t(p_t))} \frac{1}{2}\|\tilde{\ell}_t - x\mathbf{1}_K\|^2_{\nabla^2\Psi_t(z)}\right] \\
&\qquad \leq \mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p}_i \in [p_{t,i}, \nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t + x1_K)_i]} \frac{\eta_t}{2}(\tilde{\ell}_{t,i} - x)^2(\tilde{p}_i)^{2-\alpha}\right].
\end{aligned}$$

where we used the fact that $\nabla\Psi_t(p_t)$ is in the probability simplex so $\Phi_t(\nabla\Psi_t(p_t)) = \Psi_t^*(\nabla\Psi_t(p_t))$, and finally the fact that $\nabla^2\Psi_t(p) = \text{diag}\left(\frac{p_i^{\alpha-2}}{\eta_t}\right)_{i=1,\dots,K}$. □

Using the base of the stability analysis in Lemma 2.7 which follows from Zimmert and Seldin (2019), we can move on to the stability bounds. First, we focus on bounding the stability term when the distribution to query the arm to observe is arbitrary.

Proof of Lemma 2.1. Let's start by bounding the instantaneous stability at a fixed round t. We start from Lemma 2.7 with $x = 0$. We recall that $\nabla\Psi^*(\nabla\Psi_t(p_t) - \tilde{\ell}_t) \leq \nabla\Psi^*(\nabla\Psi_t(p_t)) = p_t$ because the losses are non-negative, and $\nabla\Psi_t^*$ is monotonically increasing.

$$\begin{aligned}
\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}(\tilde{\ell}_{t,i})^2 (p_{t,i})^{2-\alpha}\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}\frac{(\ell_{t,i})^2}{q_{t,i}}(p_{t,i})^{2-\alpha}\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2}\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right],
\end{aligned}$$

where we used the definition of $\tilde{\ell}_t$ and that $\mathbb{E}\left[\mathbb{1}\left\{B_t = i\right\}\right] = q_{t,i}$. The last steps relies on the fact that the losses are bounded in the $[0, 1]$ interval. Finally, summing for $t = 1$ to T finishes the proof. □

In order to derive the proof of Lemma 2.3, we first need to bound the weights estimators $\tilde{p}$ from Lemma 2.7. The proof follows from Zimmert and Seldin (2019, Proof of Lemma 16), with a refinement when $\alpha = 2/3$.

Lemma 2.8. *Let $p \in \Delta^{K-1}$ and $\tilde{p} = \nabla\Psi_t^*(\nabla\Psi_t(p) - \ell)$. If $\eta_t \leq 1/4$, then for all $\ell_i \geq -1$ it holds that $\tilde{p}_i^{2-\alpha} \leq c(\alpha)p_i^{2-\alpha}$, where $c(\alpha) = (1 - \frac{(1-\alpha)}{4})^{-\frac{2-\alpha}{1-\alpha}}$.*

Proof of Lemma 2.8. $\nabla\Psi_t$ is the inverse of $\nabla\Psi_t^*$, which gives

$$\nabla\Psi_t(\tilde{p}) = \nabla\Psi_t(p) - \ell.$$

Using our lower bound on ℓ, we deduce that in each dimension:

$$\begin{aligned}
&\nabla\Psi_t(p)_i - \nabla\Psi_t(\tilde{p})_i = \ell_i \geq -1, \\
&\frac{p_i^{\alpha-1} - 1}{(1-\alpha)\eta_t} - \frac{\tilde{p}_i^{\alpha-1} - 1}{(1-\alpha)\eta_t} \leq 1, \\
&\tilde{p}_i^{1-\alpha} \leq \frac{p_i^{1-\alpha}}{1 - \eta_t(1-\alpha)p_i^{1-\alpha}} \leq \frac{p_i^{1-\alpha}}{1 - \eta_t(1-\alpha)}, \\
&\tilde{p}_i^{2-\alpha} \leq \frac{p_i^{2-\alpha}}{(1 - \eta_t(1-\alpha))^{\frac{2-\alpha}{1-\alpha}}}.
\end{aligned}$$

Now we need to upper bound $(1-\eta_t(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}}$. This function is monotonically decreasing when $\alpha \in [0,1]$, we get the upper bound:

$$(1-\eta_t(1-\alpha))^{-\frac{2-\alpha}{1-\alpha}} \leq \left(1-\frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}} = c(\alpha).$$

□

Using those results and Lemma 2.2, we can move on to the proof of Lemma 2.3. The first part of this lemma is a direct application of Lemma 2.1.

Proof of Lemma 2.3.

First statement of the Lemma Using Lemma 2.1 and the distribution given in Lemma 2.2, we can bound the instantaneous stability at round t by:

$$\begin{aligned}
\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2} \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K} \frac{\eta_t}{2} \frac{(p_{t,i})^{2-\alpha}}{(p_{t,i})^{1-\alpha/2}} \sum_{j=1}^{K} (p_{t,j})^{1-\alpha/2}\right] \\
&= \mathbb{E}\left[\frac{\eta_t}{2}\left(\sum_{i=1}^{K} (p_{t,i})^{1-\alpha/2}\right)^2\right].
\end{aligned}$$

We can upper bound the expectation by replacing p_t by the distribution which maximizes the expression. Because $f(x) = x^2$ is an increasing function for $x \in \mathbb{R}^+$, this expression is maximized when $\sum_{i=1}^{K} (p_{t,i})^{1-\alpha/2}$ is maximized. As $1-\alpha/2 \leq 1$, using the uniform distribution maximizes this term, and we get that:

$$\begin{aligned}
\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] &\leq \mathbb{E}\left[\frac{\eta_t}{2}\left(\sum_{i=1}^{K} K^{-1+\alpha/2}\right)^2\right] \\
&\leq \mathbb{E}\left[\frac{\eta_t}{2}\left(K^{\alpha/2}\right)^2\right] \\
&\leq \frac{\eta_t}{2} K^{\alpha}.
\end{aligned}$$

Finally, summing on t finishes this part of the proof.

Second statement of the Lemma Now we move on to the second part of the lemma. This time, we start from Lemma 2.7, where we choose $x = \mathbb{1}_t [B_t = j] \ell_{t,j}$. Now the analysis depends on whether $B_t \neq j$ or $B_t = j$. In the first case, we have $x = 0$, and the expression is maximized when $\tilde{p}_i = p_{t,i}$, because the losses are non-negative, and $\nabla \Psi_t^*$ is monotonically increasing. In the second case, we have $B_t = j$, which means that for $i \neq j$, $\tilde{\ell}_{t,i} - x = 0 - x \geq -1$ and we can apply Lemma 2.8 and bound $\tilde{p}_i^{2-\alpha}$ by $c(\alpha) p_i^{2-\alpha}$, where $c(\alpha) = \left(1 - \frac{(1-\alpha)}{4}\right)^{-\frac{2-\alpha}{1-\alpha}}$. Otherwise we have $\tilde{\ell}_{t,j} - x \geq 0$, and using the fact that $\nabla \Psi_t^*$ is monotonically increasing to deduce that $\tilde{p}_j = p_{t,j}$. Combining all cases, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \max_{\tilde{p} \in [p_{t,i}, \nabla \Psi^*(\nabla \Psi_t(p_t) - \tilde{\ell}_t + x 1_K)_i]} \frac{\eta_t}{2} (\tilde{\ell}_{t,i} - x)^2 (\tilde{p}_i)^{2-\alpha}\right] \\
&\quad \leq \sum_{i \neq j} \frac{\eta_t}{2} \mathbb{E}\left[\mathbb{1}\left[B_t = i\right] \sum_{k=1}^{K} (\tilde{\ell}_{t,k})^2 (p_{t,k})^{2-\alpha}\right] \\
&\quad + \mathbb{E}\left[\mathbb{1}\left[B_t = j\right] \left(\left(\sum_{i \neq j} \frac{\eta_t}{2} (\tilde{\ell}_{t,i} - \ell_{t,j})^2 c(\alpha) (p_{t,i})^{2-\alpha}\right) + \frac{\eta_t}{2} (\tilde{\ell}_{t,j} - \ell_{t,j})^2 (p_{t,j})^{2-\alpha}\right)\right] \\
&\quad \leq \sum_{i \neq j} \frac{\eta_t}{2} \mathbb{E}\left[q_{t,i} \left(\frac{\ell_{t,i}}{q_{t,i}}\right)^2 \frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \\
&\quad + \mathbb{E}\left[\left(\sum_{i \neq j} \frac{\eta_t c(\alpha)}{2} (\ell_{t,j})^2 (p_{t,i})^{2-\alpha} q_{t,j}\right) + \frac{\eta_t}{2} \left(\frac{\ell_{t,j}}{q_{t,j}} - \ell_{t,j}\right)^2 (p_{t,j})^{2-\alpha} q_{t,j}\right],
\end{aligned}
$$

where in both equations, the first term concerns the cases where $B_t \neq j$, and the second term the case where $B_t = j$. In the second term, the index j is taken out of the sum as it is treated differently. Continuing the derivation above we have

$$
\begin{aligned}
&\leq \sum_{i \neq j} \frac{\eta_t}{2} \mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] \\
&\quad + \mathbb{E}\left[\left(\sum_{i \neq j} \frac{\eta_t c(\alpha)}{2} (\ell_{t,j})^2 (p_{t,i})^{2-\alpha} q_{t,j}\right) + \frac{\eta_t}{2} \frac{1}{(q_{t,j})^2} (\ell_{t,j})^2 (1 - q_{t,j})^2 (p_{t,j})^{2-\alpha} q_{t,j}\right] \\
&\leq \sum_{i \neq j} \frac{\eta_t}{2} \mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] + \mathbb{E}\left[\left(\sum_{i \neq j} \frac{\eta_t c(\alpha)}{2} (p_{t,i})^{2-\alpha} q_{t,j}\right) + \frac{\eta_t}{2} \frac{(p_{t,j})^{2-\alpha}}{q_{t,j}} (1 - q_{t,j})^2\right].
\end{aligned}
$$

Now, let's consider each term separately. We can replace q_t by its definition from Lemma 2.2, $\forall i \in [K], q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}}$. Note that $1-\alpha/2 \leq 1$ so $\sum_{j=1}^{K}(p_{t,j})^{1-\alpha/2} \leq K\left(\frac{1}{K}\right)^{1-\alpha/2} = K^{\alpha/2}$. In the previous expression, the first term is bounded as:

$$\begin{aligned}\sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\right] &\leq \sum_{i\neq j}\frac{\eta_t}{2}\mathbb{E}\left[(p_{t,i})^{1-\alpha/2}\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right] \\ &\leq \frac{\eta_t}{2}K^{\alpha/2}\sum_{i\neq j}\mathbb{E}\left[(p_{t,i})^{1-\alpha/2}\right].\end{aligned}$$

In order to bound the third term, we observe that $(1-q_{t,j})^2 \leq 1-q_{t,j} = \sum_{i\neq j} q_{t,i}$. This gives:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta_t}{2}\frac{(p_{t,j})^{2-\alpha}}{q_{t,j}}(1-q_{t,j})^2\right] &= \mathbb{E}\left[\frac{\eta_t}{2}(1-q_{t,j})^2(p_{t,j})^{1-\alpha/2}\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right] \\ &\leq \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}q_{t,i}\left(\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right)\right] \\ &\leq \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}\frac{(p_{t,i})^{1-\alpha/2}}{\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}}\left(\sum_{k=1}^{K}(p_{t,k})^{1-\alpha/2}\right)\right] \\ &\leq \mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}(p_{t,i})^{1-\alpha/2}\right].\end{aligned}$$

Finally, the second term can be bounded as:

$$\mathbb{E}\left[\sum_{i\neq j}\frac{\eta_t c(\alpha)}{2}(p_{t,i})^{2-\alpha}q_{t,j}\right] \leq c(\alpha)\mathbb{E}\left[\frac{\eta_t}{2}\sum_{i\neq j}(p_{t,i})^{1-\alpha/2}\right].$$

The final bound relies on the fact that $\forall t, i\ : 0 \leq p_{t,i} \leq 1$, so $(p_{t,i})^{2-\alpha} \leq \sqrt{(p_{t,i})^{2-\alpha}} = (p_{t,i})^{1-\alpha/2}$.

Combining those terms and using Jensen's inequality finishes the proof of the lemma.

$$\mathbb{E}\left[\ell_{t,A_t} + \Phi_t(-\tilde{L}_t) - \Phi_t(-\tilde{L}_{t-1})\right] \leq \frac{\eta_t}{2}\left(K^{\alpha/2} + c(\alpha) + 1\right)\sum_{i\neq j}\mathbb{E}\left[p_{t,i}\right]^{1-\alpha/2}.$$

□

2.7.2.3 Analysing the Penalty Term

The analysis of the penalty term follows from the work of Zimmert and Seldin (2019, Lemma 12).

Proof of Lemma 2.4. Algorithm 1 is part of the TSALLIS-INF framework introduced by Zimmert and Seldin (2019). Furthermore, by assumption the learning rate η_t is non increasing, and we have $\forall i \in [K],\ \xi_i = 1$, which implies that the regularizer Ψ_t that we use is symmetric. This means that both statements ofZimmert and Seldin (2019, Lemma 12) apply. The first statement of Lemma 2.4 follows directly from the first statement of Zimmert and Seldin (2019, Lemma 12).

Concerning the second statement, we consider the second statement from Zimmert and Seldin (2019, Lemma 12) for $x = \infty$. By assumption, we have $\eta_t = \frac{2\beta}{\sqrt{t}}$ for some constant $\beta > 0$.

$$
\begin{aligned}
\text{penalty} \leq & \frac{1}{\alpha(1-\alpha)} \sum_{i \neq i^*} \left(\frac{\mathbb{E}\left[p_{1,i}\right]^\alpha}{\eta_1} + \sum_{t=2}^{T} (\eta_t^{-1} - \eta_{t-1}^{-1}) \mathbb{E}\left[p_{t,i}\right]^\alpha \right) \\
= & \frac{1}{\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\frac{1}{2} \mathbb{E}\left[p_{1,i}\right]^\alpha + \sum_{t=2}^{T} \frac{1}{2} (\sqrt{t} - \sqrt{t-1})) \mathbb{E}\left[p_{t,i}\right]^\alpha \right) \\
= & \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(2\mathbb{E}\left[p_{1,i}\right]^\alpha + \sum_{t=2}^{T} (2\sqrt{t} - 2\sqrt{t-1})) \mathbb{E}\left[p_{t,i}\right]^\alpha \right) \\
& + \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} - \sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} \right) \\
= & \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\sum_{t=1}^{T} \frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}} + 2\mathbb{E}\left[p_{1,i}\right]^\alpha \right) \\
& + \frac{1}{4\alpha(1-\alpha)\beta} \sum_{i \neq i^*} \left(\sum_{t=2}^{T} \left(2\sqrt{t} - 2\sqrt{t-1} - \frac{1}{\sqrt{t}} \right) \mathbb{E}\left[p_{t,i}\right]^\alpha \right)
\end{aligned}
$$

Now we can simplify the last parts of the expression by using the uniform distribution on $[K] \backslash \{i^*\}$ which upper bounds $\sum_{i \neq i^*} \mathbb{E}\left[p_{t,i}\right]^\alpha \leq (K-1) \left(\frac{1}{K-1}\right)^\alpha = (K-1)^{1-\alpha} \leq K^{1-\alpha}$, and using that p_1 is the uniform distribution. The previous equation is upper

bounded by

$$\leq \frac{1}{4\alpha(1-\alpha)\beta}\left(\left(\sum_{i\neq i^*}\sum_{t=1}^{T}\frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right)+\left(2+\sum_{t=2}^{T}\left(2\sqrt{t}-2\sqrt{t-1}-\frac{1}{\sqrt{t}}\right)\right)K^{1-\alpha}\right)$$
$$\leq \frac{1}{4\alpha(1-\alpha)\beta}\left(\left(\sum_{i\neq i^*}\sum_{t=1}^{T}\frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}\right)+K^{1-\alpha}\right)$$
$$= \frac{1}{4\alpha(1-\alpha)\beta}\sum_{i\neq i^*}\sum_{t=1}^{T}\frac{\mathbb{E}[p_{t,i}]^\alpha}{\sqrt{t}}+\frac{K^{1-\alpha}}{4\alpha(1-\alpha)\beta},$$

where we telescoped the sum, and used the fact that $2\sqrt{T}\leq 1+\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$. □

2.7.3 Choosing the Distribution for Exploration

Given the bound on the stability derived in Lemma 2.1, choosing q_t is an optimization problem.

Proof of Lemma 2.2. Note that for each round t, we derive a new distribution q_t, so we can focus on each round separately. Furthermore, $\frac{\eta_t}{2}$ is a constant and cannot influence the choice of the distribution q_t. For any round t, q_t is the solution of the constrained optimization problem:

$$\begin{aligned}\text{minimize } & \sum_{i=1}^{K}\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}\\ \text{subject to } & \sum_{i=1}^{K}q_{t,i}=1,\\ & \forall i\in[K], 0\leq q_{t,i}\leq 1.\end{aligned}$$

We drop the second constraint, solve the minimization problem using a Lagrangian, and verify that we obtain a solution that fulfils the second constraint. Consider the Lagrangian function $\mathcal{L}(q_t,\lambda)=\sum_{i=1}^{K}\frac{(p_{t,i})^{2-\alpha}}{q_{t,i}}+\lambda\left(\sum_{i=1}^{K}q_{t,i}-1\right)$. Minimizing the Lagrangian gives solutions of the shape:

$$\forall i\in[K], q_{t,i}=\pm\frac{\sqrt{p_{t,i}^{2-\alpha}}}{\sqrt{\lambda}}.$$

All that remains is to pick the positive values of $q_{t,i}$, and select $\sqrt{\lambda} = \sum_{i=1}^{K} \sqrt{p_{t,i}^{2-\alpha}}$. This ensures that both constraints are fulfilled. The distributions that minimize the bound in Lemma 2.1 are given by

$$\forall t \in [T], i \in [K], \quad q_{t,i} = \frac{(p_{t,i})^{1-\alpha/2}}{\sum_{j=1}^{K} (p_{t,j})^{1-\alpha/2}}.$$

□

Chapter 3

An Algorithm for Stochastic and Adversarial Bandits with Switching Costs

The work presented in this chapter is based on a paper that has been published as:

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.

Abstract

We propose an algorithm for stochastic and adversarial multiarmed bandits with switching costs, where the algorithm pays a price λ every time it switches the arm being played. Our algorithm is based on adaptation of the Tsallis-INF algorithm of Zimmert and Seldin (2021) and requires no prior knowledge of the regime or time horizon. In the oblivious adversarial setting it achieves the minimax optimal regret bound of $\mathcal{O}\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$, where T is the time horizon and K is the number of arms. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, it achieves a regret bound of $\mathcal{O}\left(\big((\lambda K)^{2/3}T^{1/3} + \ln T\big)\sum_{i\neq i^*}\Delta_i^{-1}\right)$, where Δ_i are the suboptimality gaps and i^* is a unique optimal arm. In the special case of $\lambda = 0$ (no switching costs), this bound is also minimax optimal within constants. We also explore variants of the problem, where switching cost is allowed to change over time. We provide experimental evaluation showing competitiveness of our algorithm with the relevant baselines in the stochastic, stochastically constrained adversarial, and adversarial regimes with fixed switching cost.

3.1 Introduction

Multiarmed bandits are the reference framework for the study of a wide range of sequential decision-making problems, including recommendation, dynamic content optimization, digital auctions, clinical trials, and more. In this framework the algorithm repeatedly picks actions, a.k.a. arms, and, after each selection, observes the loss or reward of the corresponding action. In many application domains, algorithms have to pay a penalty $\lambda > 0$ each time they play an arm different from the one played in the previous round. Such switching cost may occur in the form of a transaction cost in financial trading, or a reconfiguration cost in industrial environments.

So far, the problem of bandits with switching costs has been studied using algorithms whose optimality depends on the nature of the source of losses (or, equivalently, rewards) for the K arms. In the oblivious adversarial case, when losses are generated by an arbitrary deterministic source, Dekel et al. (2012) used a simple variant of the Exp3 algorithm to prove an upper bound of $\mathcal{O}\big((K\ln K)^{1/3}T^{2/3}\big)$ for $\lambda = 1$ (i.e., unit switching cost) — see also (Blum and Mansour, 2007) for an earlier, slightly weaker result. A result by Dekel et al. (2013) implies a lower bound of $\Omega\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ for all $\lambda \geq 0$. Note the phase transition: if $\lambda > 0$, then the regret asymptotically grows as $T^{2/3}$, as opposed to $\sqrt{T}$ when there is no switching

cost.

In the stochastic case, where losses of each arm are generated by an i.i.d. process, Gao et al. (2019) and Esfandiari et al. (2021) used arm elimination algorithms to prove that $\mathcal{O}(\ln T)$ switches are sufficient to achieve the optimal distribution-dependent regret of $\mathcal{O}\big((\ln T)\sum_{i:\Delta_i>0}\Delta_i^{-1}\big)$, where Δ_i is the suboptimality gap of arm i. Hence, in the stochastic case the introduction of switching costs does not lead to a qualitative change of the minimax regret rate.

In practical applications, it is desirable to have algorithms that require no prior knowledge about the nature of the loss generation process and maintain robustness in the adversarial regime simultaneously with the ability to achieve lower regret in the stochastic case. A number of such algorithms have been developed for the standard multiarmed bandits (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018; Zimmert and Seldin, 2019, 2021; Masoudian and Seldin, 2021) and the ideas have been extended to several other domains, including combinatorial bandits (Zimmert et al., 2019), decoupled exploration and exploitation (Rouyer and Seldin, 2020), and episodic MDPs (Jin and Luo, 2020). We aim at designing algorithms with similar properties for bandits with switching costs.

Main contributions

Our starting point is the Tsallis-INF algorithm of Zimmert and Seldin (2021), which was shown to achieve minimax regret rates in both stochastic and adversarial regimes for standard bandits. We introduce a modification of this algorithm, which we call Tsallis-Switch, to take care of the switching costs. In the adversarial regime, the regret bound of Tsallis-Switch matches (within constants) the minimax optimal regret bound $\Theta\big((\lambda K)^{1/3}T^{2/3}+\sqrt{KT}\big)$ for any value of $\lambda \geq 0$. In the stochastically constrained adversarial regime, which includes the stochastic regime as a special case, we prove a bound $\mathcal{O}\left(\big((\lambda K)^{2/3}T^{1/3}+\ln T\big)\sum_{i\neq i^*}\Delta_i^{-1}\right)$, where i^* is a unique optimal arm. Note that, in the special case of $\lambda = 0$ (no switching costs), we recover (up to constant factors) the minimax optimal bounds of Tsallis-INF for both regimes. Similarly to Tsallis-INF, our algorithm is fully oblivious to both the regime and the time horizon T.

Tsallis-Switch, which runs Tsallis-INF as a subroutine, uses the standard tool to control the frequency of arm switching: game rounds are grouped into consecutive blocks $B_1, B_2, \ldots$, and Tsallis-Switch runs Tsallis-INF over the blocks, preventing it from switching arms within each block. The number of switches is thus bounded by the number of blocks. Since T is unknown, we use block sizes of increasing length.

As a new arm is drawn only at the beginning of each block, the effective range of the losses experienced by Tsallis-INF grows with time. Therefore, we modify the analysis of Tsallis-INF to accommodate losses of varying range. This extension may potentially be of independent interest.

3.2 Problem Setting and Notations

We consider a repeated game with K arms and a switching cost $\lambda \geq 0$. At each round $t = 1, 2, \ldots$ of the game, the environment picks a loss vector $\ell_t \in [0,1]^K$, and the algorithm chooses an arm $J_t \in [K]$ to play. The learner then incurs the loss ℓ_{t,J_t}, which is observed. If $J_t \neq J_{t-1}$, then the learner also suffers an extra penalty of λ. The penalty λ is known to the learner. We use the same setting as Dekel et al. (2013), and assume that $J_0 = 0$, which means that there is always a switch at the first round.

We consider two regimes for the losses. In the oblivious adversarial regime, the loss vectors ℓ_t are arbitrarily generated by the environment and do not depend on the actions taken by the learner. We also work in the stochastically constrained adversarial regime. This setting, introduced by Wei and Luo (2018), generalizes the widely studied stochastic regime by allowing losses to be drawn from distributions with fixed gaps. It means that at for all i, $\mathbb{E}[\ell_{t,i}]$ can fluctuate with t, but $\mathbb{E}[\ell_{t,i} - \ell_{t,j}] = \Delta_{i,j}$ remains constant over time for all pairs i, j. The suboptimality gaps are then defined as $\Delta_i = \Delta_{i,1} - \min_j \Delta_{j,1}$.

We define the pseudo-regret with switching costs as follows,

$$\begin{aligned} \mathrm{RS}(T,\lambda) &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,J_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] + \lambda \sum_{t=1}^{T} \mathbb{P}(J_{t-1} \neq J_t) \\ &= R_T + \lambda\, S_T. \end{aligned} \tag{3.1}$$

We recognize that $R_T = \mathrm{RS}(T,0)$ is the classical definition of the pseudo regret (without switching costs), while S_T counts the expected number of switches. Furthermore, we recall that in the stochastically constrained adversarial regime, the pseudo-regret can be rewritten in terms of the sub-optimality gaps, as:

$$\mathrm{R}_T = \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[p_{t,i}]\, \Delta_i, \tag{3.2}$$

where $p_{t,i}$ is the probability of playing action i at round t.

Algorithm 2: Tsallis-Switch

Input: Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$.
Block lengths $|B_1|, |B_2|, \ldots$.
Initialize: $\tilde{C}_0 = \mathbf{0}_K$
for $n = 1, 2, \ldots$ **do**
$p_n = \underset{p \in \Delta^{K-1}}{\arg\min} \left\{ \left\langle p, \tilde{C}_{n-1} \right\rangle - \sum_{i=1}^{K} \frac{4\sqrt{p_i} - 2p_i}{\eta_n} \right\}$.
Sample $I_n \sim p_n$ and play it for all rounds $t \in B_n$.
Observe and suffer $c_{n,I_n} = \sum_{t \in B_n} \ell_{t,I_n}$.
$\forall i \in [K] : \tilde{c}_{n,i} = \begin{cases} \frac{c_{n,i}}{p_{n,i}}, & \text{if } I_n = i, \\ 0, & \text{otherwise.} \end{cases}$
$\forall\, i \in [K] : \quad \tilde{C}_n(i) = \tilde{C}_{n-1}(i) + \tilde{c}_{n,i}$.
end for

3.3 Using Blocks to Control Switching Frequency

In order to control S_T, we limit the number of action switches that the algorithm makes by dividing the game rounds into blocks and forcing the algorithm to play the same action for all the rounds within a block. Given a sequence of blocks $(B_n)_{n \geq 1}$ of lengths $|B_n|$, and a time horizon T, we define N as the smallest integer, such that $\sum_{n=1}^{N} |B_n| \geq T$, and we truncate the last block, such that the cumulative length of the first N blocks sum up to T.

As $S_T \leq N$, we bound N and the pseudo-regret R_T (without the switching costs) over the N blocks. Let $c_{n,i} = \sum_{s \in B_n} \ell_{s,i}$ be the cumulative loss of action i in block n. Since $\ell_{t,i} \in [0, 1]$, we have $c_{n,i} \in \big[0, |B_n|\big]$. We use I_n to refer to the action played by the algorithm in block n. Then, for all $t \in B_n$, we have $J_t = I_n$ and

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n}\right] - \min_j \mathbb{E}\left[\sum_{n=1}^{N} c_{n,j}\right].$$

3.4 The Algorithm

Our Tsallis-Switch algorithm (see Algorithm 2) calls Tsallis-INF at the beginning of each block to obtain an action, plays the proposed action in each round within the

block, and then feeds back to Tsallis-INF the total loss suffered by the action over the block. As blocks have varying lengths, we adapt the Tsallis-INF algorithm and its analysis to losses of varying range.

3.5 Main Results

We start by considering the case where the switching cost λ is a fixed parameter given to the algorithm. Since λ is known in advance, it can be used to tune the block lengths.

Theorem 3.1. *Let $\lambda \geq 0$ be the switching cost. Define blocks with lengths $|B_n| = \max\{\lceil a_n \rceil, 1\}$, where $a_n = \frac{3\lambda}{2}\sqrt{\frac{n}{K}}$. The pseudo-regret of* Tsallis-Switch *with learning rate $\eta_n = \frac{2}{a_n+1}\sqrt{\frac{2}{n}}$ executed over the blocks in any adversarial environment satisfies:*

$$R(T,\lambda) \leq 5.25(\lambda K)^{1/3}T^{2/3} + 6.4\sqrt{KT} + 3\sqrt{2K} + 5.25\lambda + 6.25.$$

Furthermore, in any stochastically constrained adversarial regime with a unique best arm i^*, the pseudo-regret additionally satisfies:

$$\begin{aligned} R(T,\lambda) \leq \big(66(\lambda K)^{2/3}T^{1/3} + 32\ln T\big)\sum_{i\neq i^*}\frac{1}{\Delta_i} \\ + \big(160\lambda^{2/3}T^{1/3}K^{1/6} + 160\lambda + 49\lambda^2 + 32\big)\sum_{i\neq i^*}\frac{1}{\Delta_i} \\ + \frac{544\lambda}{\sqrt{K}} + \lambda + 66. \end{aligned}$$

A proof is provided in Section 3.6. For $\lambda = 0$ (no switching costs) both regret bounds match within constants the corresponding bounds of Tsallis-INF for multi-armed bandits with no switching costs. Furthermore, in the adversarial regime the algorithm achieves the optimal regret rate for all values of λ. In the stochastically constrained adversarial regime, for $\lambda > 0$ the regret grows as $T^{1/3}$ rather than logarithmically in T. This is also the case for the stochastic regime, which is a special case. While the algorithm does not achieve the logarithmic regret rate in the stochastic regime, as do the algorithms of Gao et al. (2019) and Esfandiari et al. (2021), it still exploits the simplicity of the regime and reduces the regret rate from $T^{2/3}$ to $T^{1/3}$. Additionally, in contrast to the work of Gao et al. (2019) and Esfandiari et al. (2021), the stochastic regret guarantee holds simultaneously with the adversarial regret guarantee, and the algorithm requires no knowledge of the time horizon.

We also note that we are unaware of specialized lower bounds for the more general stochastically constrained adversarial regime with switching costs, and it is unknown whether the corresponding regret guarantee is minimax optimal.

Theorem 3.1 is based on the following generalized analysis of the Tsallis-INF algorithm that accommodates losses of varying range. The result may be of independent interest.

Theorem 3.2. *Consider a multi-armed bandit problem where the loss vector at round t belongs to $[0, b_t]^K$ and b_t is revealed to the algorithm before round t. Then the pseudo-regret of* Tsallis-Switch *in any adversarial environment for any positive and non-decreasing sequence of learning rates $(\eta_t)_{t\geq 1}$ satisfies*

$$R_T \leq \sqrt{K}\left(\sum_{t=1}^{T} \frac{\eta_t}{2} b_t^2 + \frac{4}{\eta_T}\right) + 1. \tag{3.3}$$

Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret also satisfies

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{\left(\frac{7}{2}\eta_t b_t^2 + 2c\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\right)^2}{4\Delta_i b_t} + \sum_{t=1}^{T_0} \eta_t b_t^2 + 2, \tag{3.4}$$

where $c = \begin{cases} 2, & \text{if } \forall t: \ \frac{5\eta_t}{4} b_t^2 \geq 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right), \\ 4, & \text{otherwise.} \end{cases}$

In particular, if $b_t = B$ for all rounds t, we have the following more interpretable result.

Corollary 3.1. *Consider a multi-armed bandit problem with loss vectors belonging to $[0, B]^K$. Then the pseudo-regret of Tsallis-INF with $\eta_t = \frac{2}{B\sqrt{t}}$ satisfies $R_T \leq 4B\sqrt{KT}+1$ in any adversarial regime. Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret additionally satisfies*

$$R_T \leq 21B(\ln T + 1)\sum_{i \neq i^*} \frac{1}{\Delta_i} + 8\sqrt{B} + 2.$$

3.5.1 Varying Switching Cost

Now we consider a setting, where the switching cost may change after each switch. The learner is given the n-th switching cost λ_n right after the $n-1$-th switch is

taken, and we allow the length of the block $|B_n|$ to depend on it. In this setting, the cumulative expected switching cost becomes

$$S\left(T, (\lambda_n)_{n\geq 1}\right) = \sum_{n=1}^{N} \lambda_n \mathbb{P}(I_n \neq I_{n-1}),$$

where, as before, N is the smallest number of blocks to cover T rounds. We construct blocks, such that the terms R_T and $S\left(T, (\lambda_n)_{n\geq 1}\right)$ remain balanced.

Theorem 3.3. *Let $(\lambda_n)_{n\geq 1}$ be a sequence of non-negative switching costs. The pseudo-regret with switching costs of* Tsallis-Switch *executed with block lengths $|B_n| = \max\left\{\left\lceil \frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n}\lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}} \right\rceil, 1\right\}$ and $\eta_n = \frac{2\sqrt{2K}}{3a_n}$, where $a_n = \left(\sum_{s=1}^{n}\lambda_s + \sqrt{K/s}\right)$, satisfies:*

$$R(T, \lambda) \leq \sum_{n=1}^{N} 7\lambda_n + 12\sqrt{KN} + 2, \tag{3.5}$$

where N is the smallest integer such that $\sum_{n=1}^{N}|B_n| \geq T$. Furthermore, in the stochastically constrained adversarial regime with a unique best arm i^*, the pseudo regret additionally satisfies

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \sum_{n=1}^{N}\sum_{i\neq i^*} \frac{\left(11\lambda_n + \lambda_{n+1} + \frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i |B_n|} + \sum_{n=1}^{N_0}\left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + \lambda_1 + 2,$$

where N_0 is the smallest $n \leq N$ such that for all $n \geq N_0$, $\eta_n|B_n| \leq \frac{1}{4}$. If such an integer does not exist, then $N_0 = N$.

A proof is provided in Appendix 3.9.4. Note that for $\lambda_n = \lambda$, the bound (3.5) for the adversarial setting is of the same order as the corresponding bound in Theorem 3.1.

If λ_n is not monotone, then controlling the first term in the regret bound for the stochastically constrained adversarial regime is challenging, because the block length $|B_n|$ in the denominator does not depend on λ_{n+1} in the numerator. Below, we provide a specialization of the regret bound assuming that the switching costs increase as $\lambda_n = n^\alpha$ for some $\alpha > 0$.

Corollary 3.2. *Assume that for $n \geq 1$, $\lambda_n = n^\alpha$ for some $\alpha > 0$. Then the regret bound for the stochastically constrained adversarial regime with a unique best arm i^* in Theorem 3.3 satisfies*

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \mathcal{O}\left(\sum_{i\neq i^*} \frac{K^{\frac{2\alpha+2}{2\alpha+3}} T^{\frac{2\alpha+1}{2\alpha+3}} + K^{\frac{2\alpha}{2\alpha+3}} T^{\frac{4\alpha}{2\alpha+3}}}{\Delta_i}\right).$$

A proof is provided in Appendix 3.9.4. At the limit $\alpha \to 0$, the bound scales as $\mathcal{O}\big(K^{2/3}T^{1/3}\sum_{i\neq i^*}\frac{1}{\Delta_i}\big)$, which matches the pseudo-regret bound in the stochastically constrained adversarial regime in Theorem 3.1 with $\lambda = 1$. Note also that the bound remains sublinear in T, as long as $\alpha < \frac{3}{2}$. In other words, with a switching cost as high as $\lambda_n = n^{3/2-\varepsilon}$, for any $\varepsilon > 0$, Tsallis-Switch still has sublinear regret.

3.6 Proofs

We start by introducing some preliminary definitions and results. Recall that the pseudo-regret can be decomposed into a sum of stability and penalty terms (Lattimore and Svepesvári, 2020; Zimmert and Seldin, 2021). Let Φ_n be defined as:

$$\Phi_n(C) = \max_{p\in\Delta^{K-1}} \left\{ \langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n} \right\}.$$

Note that the distribution p_n used by Tsallis-Switch to draw action I_n for block B_n satisfies $p_n = \nabla\Phi_n(-\tilde{C}_{n-1})$. We can write:

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^N c_{n,I_n}\right] - \min_j \mathbb{E}\left[\sum_{n=1}^N c_{n,j}\right] = &\underbrace{\mathbb{E}\left[\sum_{n=1}^N c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right]}_{\text{stability}} \\ &+ \underbrace{\mathbb{E}\left[\sum_{n=1}^N \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i_N^*}\right]}_{\text{penalty}},\end{aligned} \tag{3.6}$$

where i_N^* is any arm with smallest cumulative loss over the N blocks (i.e., a best arm in hindsight).

We start by introducing bounds on the stability and the penalty parts of the regret. The results generalize the corresponding results of Zimmert and Seldin (2021)

to handle losses that take values in varying ranges and may be larger than 1. The proofs are provided in Section 3.9.2. Note the multiplicative factor b_n^2 in the stability term.

Lemma 3.1. *For any sequence of positive learning rates $(\eta_n)_{n\geq 1}$ and any sequence of bounds $(b_n)_{n\geq 1}$ on the losses at round n, the* stability *term of the regret bound of* Tsallis-Switch *satisfies:*

$$\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \leq \sum_{n=1}^{N} \frac{\eta_n}{2} b_n^2 \sum_{i=1}^{K} \sqrt{\mathbb{E}\left[p_{n,i}\right]}.$$

Furthermore, if $\eta_n b_n \leq \frac{1}{4}$, then for any fixed j:

$$\mathbb{E}\left[c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \leq \frac{\eta_n}{2} b_n^2 \sum_{i\neq j} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right]\right).$$

In particular, if there exists N_0 such that for all $n \geq N_0$, $\eta_n b_n \leq \frac{1}{4}$, then:

$$\begin{aligned}\mathbb{E}\left[\sum_{n=1}^{N} c_{n,I_n} + \Phi_n(-\tilde{C}_n) - \Phi_n(-\tilde{C}_{n-1})\right] \leq & \sum_{n=1}^{N} \frac{\eta_n}{2} b_n^2 \sum_{i\neq j} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right]\right) \\ & + \sum_{n=1}^{N_0} \frac{\eta_n}{2} b_n^2.\end{aligned}$$

The penalty term is not affected by the change of the range of the losses.

Lemma 3.2. *For any non-increasing positive learning rate sequence $(\eta_n)_{n\geq 1}$, the* penalty *term of the regret bound of* Tsallis-Switch *satisfies:*

$$\mathbb{E}\left[\sum_{n=1}^{N} \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i_N^*}\right] \leq \frac{4\sqrt{K}}{\eta_N} + 1.$$

Furthermore, if we define η_0, such that $\eta_0^{-1} = 0$, then

$$\begin{aligned}&\mathbb{E}\left[\sum_{n=1}^{N} \Phi_n(-\tilde{C}_{n-1}) - \Phi_n(-\tilde{C}_n) - c_{n,i_N^*}\right] \\ &\leq 4\sum_{n=1}^{N} (\eta_n^{-1} - \eta_{n-1}^{-1}) \sum_{i\neq i_N^*} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \frac{1}{2}\mathbb{E}\left[p_{n,i}\right]\right) + 1.\end{aligned}$$

We also present a bound for the cumulative switching cost, which is the key to obtain refined guarantees in the stochastically constrained adversarial regime.

Lemma 3.3. *Consider a sequence of switching costs $(\lambda_n)_{n\geq 1}$. Then for any fixed j, the cumulative switching cost satisfies*

$$S\left(T, (\lambda_n)_{n\geq 1}\right) \leq \lambda_1 + \sum_{n=1}^{N} (\lambda_n + \lambda_{n+1}) \sum_{i\neq j} \mathbb{P}(I_n = i).$$

Proof of Lemma 3.3. By convention, there is always a switch at round 1. For subsequent rounds, when there is a switch at round n at least one of I_{n-1} or I_n is not equal to j. Thus, we have:

$$\mathbb{P}(I_{n-1} \neq I_n) \leq \sum_{i\neq j} \mathbb{P}(I_{n-1} = i) + \mathbb{P}(I_n = i),$$

and the cumulative switching cost satisfies

$$\begin{aligned} S\left(T, (\lambda_n)_{n\geq 1}\right) &= \lambda_1 + \sum_{n=2}^{N} \lambda_n \mathbb{P}(I_{n-1} \neq I_n) \\ &\leq \lambda_1 + \sum_{n=2}^{N} \lambda_n \left(\sum_{i\neq j} \mathbb{P}(I_{n-1} = i) + \mathbb{P}(I_n = i) \right) \\ &\leq \lambda_1 + \sum_{n=1}^{N} \sum_{i\neq j} (\lambda_n + \lambda_{n+1}) \, \mathbb{P}(I_n = i), \end{aligned}$$

which concludes the proof. □

Armed with these results, we can move on to the proof of Theorem 3.1.

Proof of Theorem 3.1. In order to apply our results to blocks, we first calculate an upper bound on the number of blocks N. The length of the n-th block is defined as $|B_n| = \max\left\{ \left\lceil \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \right\rceil, 1 \right\}$. The sequence $(B_n)_{n\geq 1}$ satisfies $|B_n| \geq b(n)$ for $b(n) = \frac{3\lambda\sqrt{n}}{2\sqrt{K}}$ and is non-decreasing. Let $N^* = K^{1/3}(T/\lambda)^{2/3}$ and observe that:

$$\sum_{n=1}^{\lfloor N^* \rfloor + 1} |B_n| \geq \sum_{n=1}^{\lfloor N^* \rfloor + 1} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \geq \int_0^{\lfloor N^* \rfloor + 1} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} \geq \int_0^{N^*} \frac{3\lambda\sqrt{n}}{2\sqrt{K}} = \frac{\lambda}{\sqrt{K}} (N^*)^{3/2} \geq T.$$

Thus, we can upper bound N by $K^{1/3}(T/\lambda)^{2/3} + 1$.

Proof of the adversarial bound. We start by focusing on the bound in the adversarial regime. To do so, we need to control the stability and penalty terms in (3.6), and also the number of switches. As we already said, the number of switches is bounded by the number of blocks, $S_T \leq N \leq K^{1/3}(T/\lambda)^{2/3} + 1$, and thus the cumulative switching cost satisfies $\lambda S_T \leq \lambda N \leq K^{1/3}T^{2/3}\lambda^{1/3} + \lambda$.

Next, we bound the quantity $\eta_n |B_n|^2$ for all $n \leq N$:

$$\frac{\eta_n}{2}|B_n|^2 \leq \frac{\sqrt{2}}{\sqrt{n}}\left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}} + 1\right) \leq \frac{3\lambda}{\sqrt{2K}} + \frac{\sqrt{2}}{\sqrt{n}}. \tag{3.7}$$

Note that even though the last block B_N may be truncated, we can upper bound its length by the non-truncated length of that block.

Then, we bound the inverse of the learning rate at round N,

$$\frac{1}{\eta_N} \leq \frac{\sqrt{N}}{2\sqrt{2}}\left(\frac{3\lambda\sqrt{N}}{2\sqrt{K}} + 1\right) \leq \frac{3\sqrt{2}}{8}\frac{\lambda N}{\sqrt{K}} + \frac{\sqrt{2}}{4}\sqrt{N}.$$

In order to bound the pseudo-regret over the N blocks, we apply inequality (3.3) from Theorem 3.2. We then add the cumulative switching cost and use the upper bound on N derived earlier,

$$\begin{aligned} R(T,\lambda) &\leq 3\sqrt{2}\lambda N + 3\sqrt{2KN} + \lambda N + 1 \\ &= (3\sqrt{2}+1)\lambda N + 3\sqrt{2KN} + 1 \\ &\leq 5.25\lambda^{1/3}K^{1/3}T^{2/3} + 3\sqrt{2}\frac{K^{2/3}T^{1/3}}{\lambda^{1/3}} + 3\sqrt{2K} + 5.25\lambda + 6.25. \end{aligned}$$

For small λ the term $K^{2/3}(T/\lambda)^{1/3}$ dominates the expression. However, when $\lambda \leq \frac{2}{3}\sqrt{\frac{K}{T}}$, then for all $n \leq T$ we have $\frac{3\lambda\sqrt{n}}{2\sqrt{K}} \leq \sqrt{\frac{n}{T}} \leq 1$, which means that $|B_n| = 1$. In this case the algorithm is not using blocks and we have $\lambda S_T \leq \lambda T \leq \frac{2}{3}\sqrt{KT}$. As we also have $a_n \leq 1$, we get $\frac{\sqrt{2}}{\sqrt{n}} \leq \eta_n \leq \frac{2\sqrt{2}}{\sqrt{n}}$. In this case we use Lemmas 3.1 and 3.2 to bound the stability and the penalty terms and obtain that stability and penalty are both bounded by $2\sqrt{2KN}$. Thus, overall, for $\lambda \leq \frac{2}{3}\sqrt{\frac{K}{T}}$ we have $R(T,\lambda) \leq 6.4\sqrt{KT}$, and for $\lambda \geq \frac{2}{3}\sqrt{\frac{K}{T}}$ we have $K^{2/3}(T/\lambda)^{1/3} \leq 1.15\sqrt{KT}$.

Piecing together all parts of the bound finishes the proof.

Proof of the stochastically constrained adversarial bound. We now derive refined guarantees in the stochastically constrained adversarial regime with a unique best arm i^*. We start by deriving bounds for the stability and penalty terms in (3.6).

Let N_0 be a constant, such that for $n \geq N_0$ we have $\eta_n |B_n| \leq \frac{1}{4}$. We note that $\eta_n |B_n| \leq \frac{2\sqrt{2}}{\sqrt{n}}$, so picking $N_0 = 128$ works. For the stability term we use the second part of Lemma 3.1 with $j = i^*$. Using (3.7) to bound $\frac{\eta_n}{2}|B_n|^2$ we obtain that the stability term is upper bounded by

$$\text{stab} \leq \sum_{n=1}^{N} \left(\frac{3\sqrt{2}\lambda}{2\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}} \right) \sum_{i \neq i^*} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} + 2.5\mathbb{E}\left[p_{n,i}\right] \right) + \sum_{n=1}^{N_0} \left(\frac{3\sqrt{2}}{2} \frac{\lambda}{\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}} \right).$$

For the penalty term, we first bound the difference between the inverse of two consecutive learning rates.

$$\begin{aligned}
\eta_n^{-1} - \eta_{n-1}^{-1} &= \left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}} + 1 \right) \frac{\sqrt{n}}{2\sqrt{2}} - \left(\frac{3\lambda\sqrt{n-1}}{2\sqrt{K}} + 1 \right) \frac{\sqrt{n-1}}{2\sqrt{2}} \\
&= \frac{3\sqrt{2}\lambda}{8\sqrt{K}} + \frac{\sqrt{n} - \sqrt{n-1}}{2\sqrt{2}} \\
&\leq \frac{3\sqrt{2}\lambda}{8\sqrt{K}} + \frac{\sqrt{2}}{4\sqrt{n}}.
\end{aligned}$$

Now we use the second part of Lemma 3.2 to bound the penalty term as follows

$$\sum_{n=1}^{N} \left(\frac{3\sqrt{2}\lambda}{2\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}} \right) \sum_{i \neq i^*} \left(\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \frac{1}{2}\mathbb{E}\left[p_{n,i}\right] \right) + 1.$$

Summing the two bounds, and using that for all n, i, $\mathbb{E}\left[p_{n,i}\right] \leq \sqrt{\mathbb{E}\left[p_{n,i}\right]}$, we have:

$$R_T \leq \sum_{n=1}^{N} \left(\left(\frac{6\sqrt{2}\lambda}{\sqrt{K}} + \frac{4\sqrt{2}}{\sqrt{n}} \right) \sum_{i \neq i^*} \sqrt{\mathbb{E}\left[p_{n,i}\right]} \right) + \frac{3\sqrt{2}\lambda}{2\sqrt{K}} N_0 + 2\sqrt{2N_0} + 1.$$

Now we use the self-bounding technique (Zimmert and Seldin, 2021), which states that if L and U are such that $L \leq R \leq U$, then $R \leq 2U - L$. For the lower bound L, we use the following identity for the regret

$$R_T = \sum_{n=1}^{N} |B_n| \sum_{i \neq i^*} \Delta_i \mathbb{E}\left[p_{n,i}\right],$$

where B_N is truncated, so that $|B_1| + \cdots + |B_N| = T$. Using the previous expression for the upper bound U, we get:

$$R_T \leq \sum_{n=1}^{N} \left(\frac{12\sqrt{2}\lambda}{\sqrt{K}} + \frac{8\sqrt{2}}{\sqrt{n}} \right) \sum_{i \neq i^*} \sqrt{\mathbb{E}\left[p_{n,i}\right]} - \sum_{n=1}^{N} |B_n| \sum_{i \neq i^*} \Delta_i \mathbb{E}\left[p_{n,i}\right] + \frac{544\lambda}{\sqrt{K}} + 66.$$

We bound the cumulative switching cost using Lemma 3.3:

$$\lambda S_T \leq \lambda + \sum_{n=1}^{N} \sum_{i \neq i^*} 2\lambda \mathbb{E}\left[p_{n,i}\right].$$

We add those two bounds together to obtain a bound on the regret with switching costs. Note (again) that $\mathbb{E}\left[p_{n,i}\right] \leq \sqrt{\mathbb{E}\left[p_{n,i}\right]}$ for all n and i, and that $\frac{\sqrt{2}}{\sqrt{K}} \leq 1$. Thus, we can upper bound the pseudo-regret with switching costs as:

$$R(T,\lambda) \leq \sum_{n=1}^{N} \sum_{i \neq i^*} \left(\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}} \right) \sqrt{\mathbb{E}\left[p_{n,i}\right]} - \Delta_i |B_n| \mathbb{E}\left[p_{n,i}\right] \right) + \frac{544\lambda}{\sqrt{K}} + \lambda + 66.$$

Now we note that each term in the inner sum is an expression of the form $a\sqrt{x} - bx$, which for $x \in [0, \infty]$ is maximized at $x = \frac{a^2}{4b}$. Put attention that the cumulative switching cost is part of the optimization problem. So, for any i and any $n < N$, we have:

$$\begin{aligned}
\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}} \right) \sqrt{\mathbb{E}\left[p_{n,i}\right]} - \Delta_i |B_n| \mathbb{E}\left[p_{n,i}\right] &\leq \frac{\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{n}} \right)^2}{4\Delta_i |B_n|} \\
&\leq \frac{(14\lambda)^2}{4\Delta_i \left(\frac{3\lambda\sqrt{n}}{2\sqrt{K}} \right)} + 2\frac{14\lambda \left(\frac{8\sqrt{2}}{\sqrt{n}} \right)}{4\Delta_i} + \frac{\left(\frac{8\sqrt{2}}{\sqrt{n}} \right)^2}{4\Delta_i} \qquad (3.8) \\
&\leq \frac{33\lambda\sqrt{K}}{\Delta_i \sqrt{n}} + \frac{80\lambda}{\Delta_i \sqrt{n}} + \frac{32}{\Delta_i n}, \qquad (3.9)
\end{aligned}$$

where in the first term of (3.8) we have lower bounded $|B_n|$ by b_n and in the last two terms by 1. As the last block may be truncated, for $n = N$ we bound $|B_N|$ in the first term in (3.9) by 1, leading to

$$\left(14\lambda + \frac{8\sqrt{2}}{\sqrt{N}} \right) \sqrt{\mathbb{E}\left[p_{N,i}\right]} - \Delta_i |B_N| \mathbb{E}\left[p_{N,i}\right] \leq \frac{49\lambda^2}{\Delta_i} + \frac{80\lambda}{\Delta_i \sqrt{n}} + \frac{32}{\Delta_i n},$$

All that remains is to sum over n. For the first term in (3.9) we have:

$$\begin{aligned}
\frac{49\lambda^2}{\Delta_i} + \sum_{n=1}^{N-1} \frac{33\lambda\sqrt{K}}{\Delta_i \sqrt{n}} &\leq 66\frac{\lambda\sqrt{K(N-1)}}{\Delta_i} + \frac{49\lambda^2}{\Delta_i} \\
&\leq 66\frac{\lambda^{2/3} T^{1/3} K^{2/3}}{\Delta_i} + \frac{49\lambda^2}{\Delta_i}.
\end{aligned}$$

Similarly, the second term in (3.9) gives:

$$\sum_{n=1}^{N} \frac{80\lambda}{\Delta_i \sqrt{n}} \leq 160 \frac{\lambda \sqrt{N}}{\Delta_i} \leq 160 \frac{\lambda^{2/3} T^{1/3} K^{1/6} + \lambda}{\Delta_i}.$$

For the last term in (3.9), we use the fact that $N \leq T$ and we have:

$$\sum_{n=1}^{N} \frac{32}{\Delta_i n} \leq \frac{32 \ln T}{\Delta_i} + \frac{32}{\Delta_i}.$$

Putting everything together finishes the proof. □

3.7 Experiments

We compare the performance of Tsallis-Switch to different baselines, both in the stochastic and in the stochastically constrained adversarial regime. We compare Tsallis-Switch with block lengths chosen as in Theorem 3.1 against Tsallis-INF without blocks, and against the BaSE algorithm of Gao et al. (2019), which achieves a regret of $\mathcal{O}\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right)$ with $\mathcal{O}(\log T)$ switches in the stochastic regime. We use T to tune the parameters of BaSE, and we consider both arithmetic and geometric blocks —see (Gao et al., 2019) for details.

We also include in our baselines the EXP3 algorithm with a time-varying learning rate, and the block version of EXP3, where the blocks have length $\lambda^{2/3} \frac{T^{1/3}}{K^{1/3}}$. Both block length and learning rate are chosen according to the analysis of EXP3 in the adversarial regime.

In the experiments, we fix the number of arms $K = 8$, and set the expected loss of a suboptimal arm to 0.5. We generate binary losses using two sets of parameters: an "easy" setting, where the gaps $\Delta = 0.2$ are large and the switching costs $\lambda = 0.025$ are small. A "hard" setting, where the gaps $\Delta = 0.05$ are small and the switching costs $\lambda = 1$ are large. For each experiment, we plot the pseudo-regret, the number of switches, and the pseudo-regret with switching cost. This allows us to observe the trade-off between the pseudo-regret and the number of switches.

In the first experiment (Figure 3.1) we use stochastic i.i.d. data with the easy setting ($\Delta = 0.2$ and $\lambda = 0.025$). As the gaps are large, even the methods that do not use blocks are not making many switches, and the best performance is achieved by Tsallis-INF without blocks. In Figure 3.2 we use the hard setting ($\Delta = 0.05$ and $\lambda = 1$). In this case, we see a trade-off between achieving a small pseudo-regret and

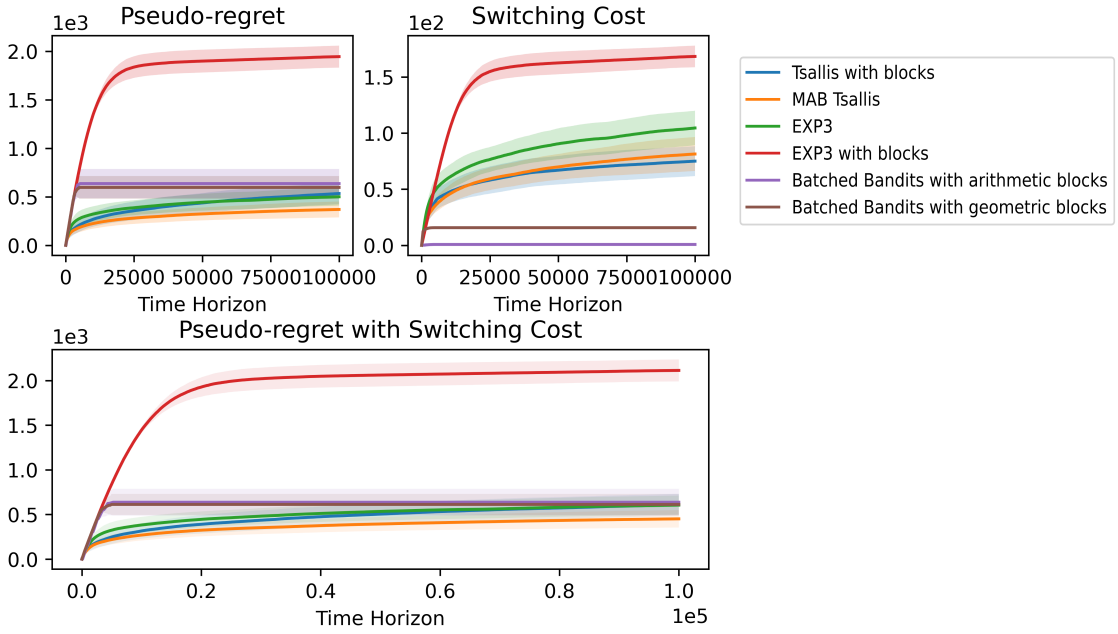

Figure 3.1: Stochastic losses, $\Delta = 0.2$ and $\lambda = 0.025$ (easy setting).

limiting the cumulative switching cost. The small value of Δ forces a larger number of switches, and because the cost of switching is now large, the cumulative switching cost dominates the pseudo-regret with switching cost.

In Figure 3.3, we test a stochastic setting with small gaps and zero switching cost. In this case, we observe that Tsallis-Inf and Tsallis-Switch outperform both EXP3 and the BaSE algorithms. Note that here Tsallis-Switch and Tsallis-Inf have very similar performances, though not identical due to a slight difference in the tuning of learning rates.

We present a wider range of experiments in Appendix 3.9.5. We show that our algorithm outperforms the BaSE algorithm in the stochastically constrained adversarial regime. Being an elimination-based algorithm, BaSE also fails in the adversarial regime.

3.8 Discussion

We introduced Tsallis-Switch, the first algorithm for multiarmed bandits with switching costs that provides adversarial pseudo-regret guarantees simultaneously with im-

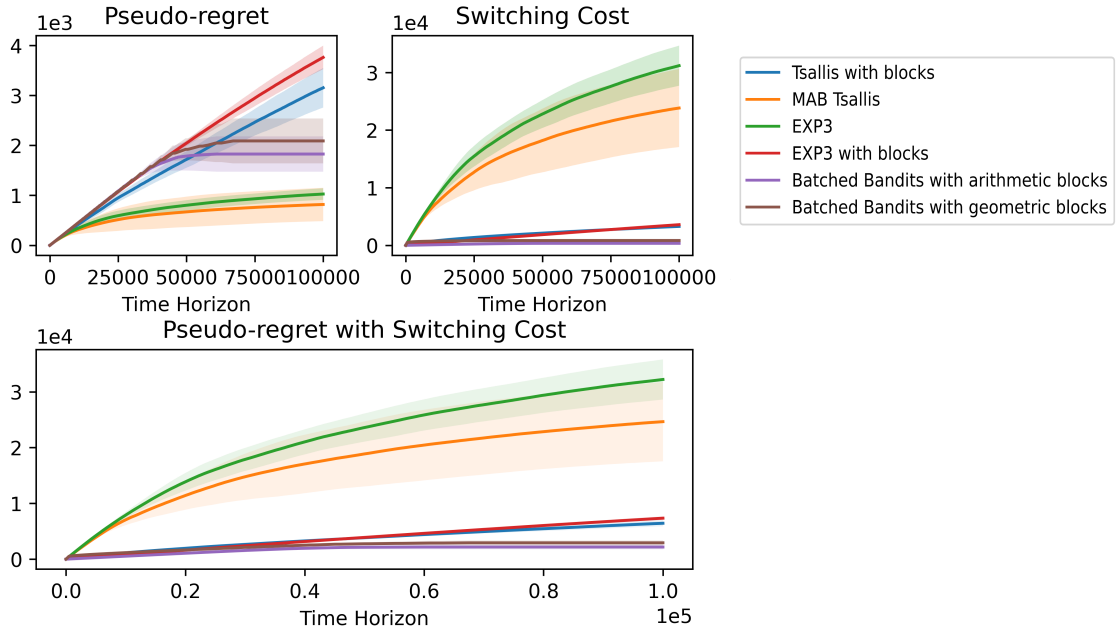

Figure 3.2: Stochastic losses, $\Delta = 0.05$ and $\lambda = 1$ (hard setting).

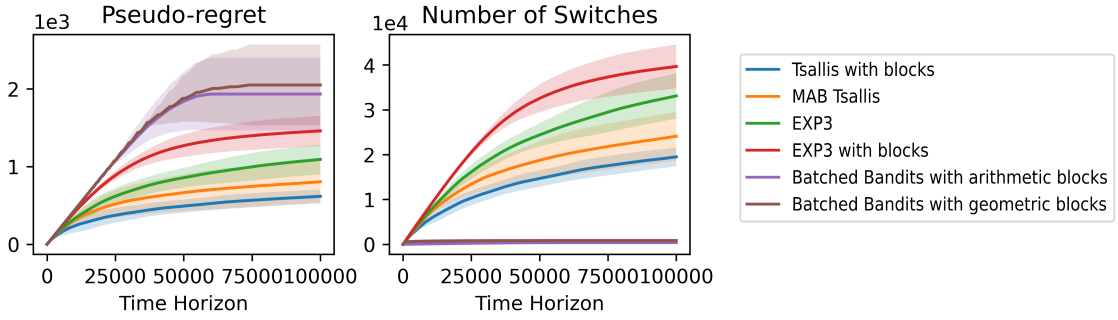

Figure 3.3: Stochastic losses and no switching cost, $\lambda = 0$ and $\Delta = 0.05$. As the switching costs are 0, the pseudo-regret and the pseudo-regret with switching costs are equal.

proved pseudo-regret guarantees in the stochastic regime, as well as the more general stochastically constrained adversarial regime. The adversarial regret bound matches the minimax lower bound within constants, and guarantees $T^{2/3}$ scaling of the regret in time. The stochastic and stochastically constrained adversarial bounds reduce the dependence of the regret on time down to $T^{1/3}$. Our experiments demonstrate that Tsallis-Switch is competitive with the relevant benchmarks over a range of settings: in the stochastic setting, it is competitive with state-of-the-art algorithms for stochastic bandits with switching costs, and outperforms state-of-the-art adversarial algorithms. In the adversarial setting, it is competitive with state-of-the-art adversarial algorithms and significantly outperforms the stochastic ones.

Our work opens multiple directions for future research. For example, it is known that in the stochastic setting with switching costs it is possible to achieve logarithmic regret scaling, but it is unknown whether it is achievable simultaneously with the adversarial regret guarantee. It is also unknown whether logarithmic regret scaling is achievable for the more general stochastically constrained adversarial regime with switching costs (even with no simultaneous requirement of an adversarial regret guarantee). Elimination of the assumption on uniqueness of the best arm in the stochastically constrained adversarial regime is another challenging direction to work on. Unfortunately, for now it is unknown how to eliminate this assumption even in the analysis of the Tsallis-INF algorithm for multiarmed bandits without switching costs. But while in the setting without switching costs the assumption has been empirically shown to be an artifact of the analysis having no negative impact on the regret (Zimmert and Seldin, 2021), in the setting with switching costs treating multiple best arms is more challenging, because switching between best arms is costly.

3.9 Appendix

3.9.1 Properties of the Potential Function

We recall several properties of the potential function provided by Zimmert and Seldin (2021, Appendix C), which we use in our proofs. We use $v = (v_i)_{i=1,\dots,K}$ to denote a column vector $v \in \mathcal{R}^K$ with elements $v_1, \dots, v_K$, and $\mathrm{diag}(v)$ to denote a $K \times K$ matrix with $v_1, \dots, v_K$ on the diagonal and 0 elsewhere. For a positive semidefinite matrix M we use $||\cdot||_M = \sqrt{\langle \cdot, M\cdot \rangle}$ to denote the canonical norm with respect to

M. The potential function is defined as

$$\Psi_n(p) = -\sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}$$

and we have

$$\nabla\Psi_n(p) = -\left(\frac{2p_i^{-1/2} - 2}{\eta_n}\right)_{i=1,\dots,K}$$

and

$$\nabla^2\Psi_n(p) = \operatorname{diag}\left(\left(\frac{p_i^{-3/2}}{\eta_n}\right)_{i=1,\dots,K}\right).$$

For $C \leq 0$, the convex conjugate and the gradient of the convex conjugate are

$$\Psi_n^*(C) = \max_p \left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}\right\}, \tag{3.10}$$

$$\nabla\Psi_n^*(C) = \arg\max_p \left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}\right\} = \left(\left(-\frac{\eta_n}{2}C_i + 1\right)^{-2}\right)_{i=1,\dots,K}. \tag{3.11}$$

We use Δ^{K-1} to denote the probability simplex over K points and $\mathcal{I}_{\Delta^{K-1}}(x) = \begin{cases} 0 & \text{if } x \in \Delta^{K-1} \\ -\infty & \text{otherwise} \end{cases}$. We also use:

$$\Phi_n(C) = \left(\Psi_n + \mathcal{I}_{\Delta^{K-1}}\right)^*(C) = \max_{p\in\Delta^{K-1}} \left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}\right\},$$

and

$$\nabla\Phi_n(C) = \arg\max_{p\in\Delta^{K-1}} \left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i} - 2p_i}{\eta_n}\right\}.$$

Φ_n is a constrained version of Ψ_n^*, where p is restricted to the probability simplex. Following Zimmert and Seldin (2021, Section 4.3), there exists a Lagrange multiplier ν such that:

$$p_n = \nabla\Phi_n(-\tilde{C}_{n-1}) = \nabla\Psi_n^*(-\tilde{C}_{n-1} + \nu\mathbf{1}_K). \tag{3.12}$$

It is important to note that Ψ_n is a Legendre function, which implies that its gradient is invertible and $\nabla(\Psi_n)^{-1} = \nabla(\Psi_n^*)$. By the Inverse Function theorem

$$\nabla^2\Psi_n^*\big(\nabla\Psi_n(w)\big) = \big(\nabla^2\Psi_n(w)\big)^{-1}. \tag{3.13}$$

The Bregman divergence associated with a Legendre function f is defined by:

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle . \tag{3.14}$$

By Taylor's theorem,

$$D_f(x, y) \leq \frac{1}{2} \|x - y\|^2_{\nabla^2 f(z)} \tag{3.15}$$

for some $z \in \text{conv}(x, y)$.

3.9.2 Proofs of the Lemmas

Here we provide a proof of the bound on the stability term in Lemma 3.1. The scaling of the stability term directly depends on the bound on the losses, so we adapt the bound for sequences of losses that are not in the $[0, 1]$ interval. Lemma 3.2 follows directly from Zimmert and Seldin (2021, Lemma 12). We focus on the case where $\alpha = \frac{1}{2}$ and in the second part of the lemma we pick $x = \infty$.

3.9.2.1 Bounding the Stability

The proof of Lemma 3.1 closely follows the proof of the corresponding result by Zimmert and Seldin (2021, Lemma 11). The main adaptation that we make is to take care of the losses that take values in $[0, b_n]$ intervals rather than $[0, 1]$ intervals.

In order to prove Lemma 3.1, we first need to adapt Zimmert and Seldin (2021, Lemma 17) to properly scale with the range b_n. Furthermore, we take advantage of the fact that $\alpha = \frac{1}{2}$ in order to derive a tighter multiplicative constant.

Lemma 3.4. *Let $p \in \Delta^{K-1}$ and $\tilde{p} = \nabla \Psi_n^*(\nabla \Psi_n(p) - c)$. If $\eta_n b_n \leq \frac{1}{4}$ and $\alpha = \frac{1}{2}$, then for all $c \in \mathbb{R}^K$ with $c_i \leq -b_n$ for all i, it holds that $\tilde{p}_i^{3/2} \leq 1.5 p_i^{3/2}$ for all i.*

Note that we obtain a slightly better constant factor 1.5 rather than factor 2 in the more general analysis by Zimmert and Seldin (2021, Lemma 17).

Proof. Since $\nabla \Psi_n$ is the inverse of $\nabla \Psi_n^*$, we have:

$$
\begin{aligned}
&\nabla\Psi_n(p)_i - \nabla\Psi_n(\tilde{p})_i = c_i \geq -b_n,\\
&\frac{p_i^{-1/2}-1}{\frac{1}{2}\eta_n} - \frac{\tilde{p}_i^{-1/2}-1}{\frac{1}{2}\eta_n} \leq b_n,\\
&\frac{p_i^{-1/2}-1}{(\frac{1}{2}\eta_n} - \frac{\tilde{p}_i^{-1/2}-1}{\frac{1}{2}\eta_n} \leq b_n,\\
&\frac{p_i^{-1/2}-1}{\frac{1}{2}\eta_n b_n} - \frac{\tilde{p}_i^{-1/2}-1}{\frac{1}{2}\eta_n b_n} \leq 1,\\
&\tilde{p}_i^{1/2} \leq \frac{p_i^{1/2}}{1-\eta_n b_n \frac{1}{2} p_i^{1/2}} \leq \frac{p_i^{1/2}}{1-\eta_n b_n \frac{1}{2}},\\
&\tilde{p}_i^{3/2} \leq \frac{p_i^{3/2}}{\left(1-\frac{1}{2}\eta_n b_n\right)^3}.
\end{aligned}
$$

It remains to bound $\left(1-\frac{1}{2}\eta_n b_n\right)^{-3}$. Using the fact that $\eta_n b_n \leq \frac{1}{4}$, we have:

$$
\left(1-\frac{1}{2}\eta_n b_n\right)^{-3} \leq \left(1-\frac{1}{8}\right)^{-3} \leq \frac{8^3}{7^3} \leq 1.5.
$$

□

With this Lemma at hand, we can move on to the proof of Lemma 3.1. We first verify that the bound still holds for losses outside of the $[0,1]$ interval, and then we observe how the bound scales in terms of the bounds b_n.

Proof of Lemma 3.1. The beginning of the proof is useful for both statements of the Lemma.

By definition, we have $p_n = \nabla\Phi_n(-\tilde{C}_{n-1})$ and $c_{n,I_n} = \langle p_n, \tilde{c}_n\rangle$. We also have $\Phi_n(C + x\mathbf{1}_K) = \Phi_n(C) + x$, because

$$
\begin{aligned}
\Phi_n(C+x\mathbf{1}_K) &= \max_{p\in\Delta^{K-1}} \left\{ \langle p, C+x\mathbf{1}_K\rangle + \sum_i \frac{4\sqrt{p_i}-2p_i}{\eta_n} \right\}\\
&= \max_{p\in\Delta^{K-1}} \left\{ \langle p, C\rangle + \langle p, x\mathbf{1}_K\rangle + \sum_i \frac{4\sqrt{p_i}-2p_i}{\eta_n} \right\}\\
&= \max_{p\in\Delta^{K-1}} \left\{ \langle p, C\rangle + x + \sum_i \frac{4\sqrt{p_i}-2p_i}{\eta_n} \right\} = \Phi_n(C) + x.
\end{aligned}
$$

Using Equation 3.12, there exists a constant λ_n, such that $\nabla\Psi_n(p_n) = -\tilde{C}_{n-1}+\lambda_n \mathbf{1}_K$. Hence, for any $x \in \mathbb{R}$:

$$
\begin{aligned}
&\mathbb{E}\left[c_{n,I_n} + \Phi_n(-\tilde{C}_n) + \Phi_n(-\tilde{C}_{n-1})\right] \\
&= \mathbb{E}\left[\langle p_n, \tilde{c}_n\rangle + \Phi_n(-\tilde{C}_n) + \Phi_n(-\tilde{C}_{n-1})\right] \\
&= \mathbb{E}\left[\langle p_n, \tilde{c}_n\rangle + \Phi_n(\nabla\Psi_n(p_n) - \tilde{c}_n) + \Phi_n(\nabla\Psi_n(p_n))\right] \\
&= \mathbb{E}\left[\langle p_n, \tilde{c}_n - x\mathbf{1}_K\rangle + \Phi_n(\nabla\Psi_n(p_n) - \tilde{c}_n + x\mathbf{1}_K) + \Phi_n(\nabla\Psi_n(p_n))\right] \\
&\leq \mathbb{E}\left[\langle p_n, \tilde{c}_n - x\mathbf{1}_K\rangle + \Psi_n^*(\nabla\Psi_n(p_n) - \tilde{c}_n + x\mathbf{1}_K) + \Psi_n^*(\nabla\Psi_n(p_n))\right] \qquad (3.16) \\
&= \mathbb{E}\left[D_{\Psi_n^*}\left(\nabla\Psi_n(p_n) - \tilde{c}_n + x\mathbf{1}_K, \nabla\Psi_n(p_n)\right)\right] \\
&\leq \mathbb{E}\left[\max_{z\in\mathrm{conv}(\nabla\Phi_n(p_n),\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K)} \frac{1}{2}\|\tilde{c}_n - x\mathbf{1}_K\|^2_{\nabla^2\Psi_n^*(z)}\right] \qquad (3.17) \\
&= \mathbb{E}\left[\max_{p\in\mathrm{conv}(p_n,\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K))} \frac{1}{2}\|\tilde{c}_n - x\mathbf{1}_K\|^2_{\nabla^2\Psi_n(p)^{-1}}\right] \qquad (3.18) \\
&\leq \mathbb{E}\left[\sum_{i=1}^{K} \max_{p\in[p_{n,i},\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K))_i]} \frac{\eta_n}{2}\left(\tilde{c}_{n,i} - x\right)^2 p_i^{3/2}\right],
\end{aligned}
$$

where Equation (3.16) uses that $\Phi_n(x) \leq \Psi_n^*(x)$, because Φ_n is a constrained version of Ψ_n^*, and $\Phi_n(\nabla\Psi_n(p_n)) = \Psi_n^*(\nabla\Psi_n(p_n))$, because $\arg\max_{p\in\mathbb{R}^K}\langle p, \nabla\Psi_n(p_n)\rangle - \Psi_n(p) = p_n$ and p_n is in the probability simplex, so the constraint in Φ_n is inactive. Equation (3.17) follows from Equation (3.15), and Equation (3.18) from Equation (3.13).

First part of the Lemma In order to prove the first part of the Lemma, we set $x = 0$ and observe that $\nabla\Psi_n^*\left(\nabla\Psi_n(p_n) - \tilde{c}_n\right)_i \leq \nabla\Psi_n^*\left(\nabla\Psi_n(p_n)\right)_i = p_{n,i}$, because the losses are non-negative and $\nabla\Psi_n^*(C) = \arg\max_p\left\{\langle p, C\rangle + \sum_i \frac{4\sqrt{p_i}-2p_i}{\eta_n}\right\}$ is a monotonically increasing function of C. This observation implies that the highest value of $[p_{n,i}, \nabla\Psi_n^*(\nabla\Psi_n(p_n) - \tilde{c}_n + x\mathbf{1}_K))_i]$ is $p_{n,i}$. Since the importance weighted losses are

0 for the arms that were not played, we have:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{K} \max_{p\in[p_{n,i},\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K))_i]} \frac{\eta_n}{2}\tilde{c}_{n,i}^2 p_i^{3/2}\right] &= \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_n}{2}\tilde{c}_{n,i}^2 p_{n,i}^{3/2}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_n}{2}\frac{c_{n,i}^2}{p_{n,i}^2}\mathbb{1}(I_n=i)p_{n,i}^{3/2}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\frac{\eta_n}{2}b_n^2 p_{n,i}^{1/2}\right] \\
&\leq \frac{\eta_n}{2}b_n^2\sum_{i=1}^{K}\mathbb{E}\left[p_{n,i}\right]^{1/2},
\end{aligned}$$

where we use the fact that $c_{n,i}^2 \leq b_n^2$, and that $\mathbb{E}_n\big[\mathbb{1}(I_n=i)\big] = p_{n,i}$, where $\mathbb{1}(I_n=i)$ is the indicator function of the event $\{I_n=i\}$ and the expectation is taken with respect to all randomness prior to round n. We use Jensen's inequality in the last inequality. Finally, summing on n finishes this part of the proof.

Second part of the Lemma We now set $x = \mathbb{1}_n[I_n=j]c_{n,j}$, where $1_n[\cdot]$ is conditioned on all randomness previous to block n. In the calculation below, for the events $I_n \in [K]\backslash\{j\}$, we have $x=0$ and use the same derivation as in the previous case. When $I_n=j$, for $i\neq j$ we have $\tilde{c}_{n,i}-x = -x \geq -b_n$, and for j we have $\tilde{c}_{n,j}-x \geq 0$. For $i\neq j$ we use Lemma 3.4 to bound $\left(\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K)\right)_i\big)^{3/2} \leq 1.5p_{n,i}^{3/2}$ and for j we use $\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n)_j \leq \nabla\Psi_n^*(\nabla\Psi_n(p_n))_j = p_{n,j}$. Therefore, we can

write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{K} \max_{p\in[p_{n,i},\nabla\Psi_n^*(\nabla\Psi_n(p_n)-\tilde{c}_n+x\mathbf{1}_K))_i]} \frac{\eta_n}{2}\left(\tilde{c}_{n,i}-x\right)^2 p_i^{3/2}\right] \\
&\leq \sum_{i\neq j} \frac{\eta_n b_n^2}{2}\mathbb{E}\left[p_{n,i}\right]^{1/2} \\
&\quad + \mathbb{E}\left[\mathbb{1}_n[I_n=j](j)\left(\frac{\eta_n}{2}\left(\frac{c_{n,j}}{p_{n,j}}-c_{n,j}\right)^2 p_{n,j}^{3/2} + \sum_{i\neq j}\frac{\eta_n}{2}c_{n,j}^2 1.5 p_{n,i}^{3/2}\right)\right] \\
&\leq \sum_{i\neq j} \frac{\eta_n b_n^2}{2}\mathbb{E}\left[p_{n,i}\right]^{1/2} + \mathbb{E}\left[\frac{\eta_n b_n^2}{2}\left(1-p_{n,j}\right)^2 p_{n,j}^{1/2} + \sum_{i\neq j}\frac{1.5}{2}\eta_n b_n^2 p_{n,i}^{3/2} p_{n,j}\right] \\
&\leq \frac{\eta_n b_n^2}{2}\sum_{i\neq j}\left(\mathbb{E}\left[p_{n,i}\right]^{1/2} + 2.5\mathbb{E}\left[p_{n,i}\right]\right),
\end{aligned}$$

where in the last step we used the fact that $\left(1-p_{n,j}\right)^2 p_{n,j}^{1/2} \leq \left(1-p_{n,j}\right) = \sum_{i\neq j} p_{n,i}$ for the middle term and $p_{n,i}^{1/2} p_{n,j} \leq 1$ for the last term. □

3.9.3 Proof of Theorem 3.2 and its Corollary

A side result of our analysis generalizes the analysis of Tsallis-INF (Zimmert and Seldin, 2021) to loss sequences that are not in the $[0,1]^K$ range.

We start with the proof of Theorem 3.2.

Proof of Theorem 3.2.

The Adversarial Regime The sequence of learning rates $(\eta_t)_{t\geq 1}$ is positive and non decreasing. Therefore, we can apply the first parts of Lemmas 3.1 and 3.2, and since $\sum_{i=1}^K \sqrt{\mathbb{E}\left[p_{n,i}\right]} \leq \sqrt{K}$, we directly obtain the result:

$$R_T = \text{stability} + \text{penalty} \leq \sum_{t=1}^{T}\frac{\eta_t}{2}b_t^2\sqrt{K} + \frac{4\sqrt{K}}{\eta_T} + 1.$$

The Stochastically Constrained Adversarial Regime Now we derive refined guarantees in the stochastically constrained adversarial regime with a unique best arm i^*. We start by deriving bounds for the stability and the penalty.

Let T_0 be a constant such that for all $t \geq T_0$ we have $\eta_t b_t \leq \frac{1}{4}$. Then by the last part of Lemma 3.1 with $j = i^*$ we have:

$$\text{stab} \leq \sum_{t=1}^{T} \frac{\eta_t}{2} b_t^2 \sum_{i \neq i^*} \left(\sqrt{\mathbb{E}\left[p_{t,i}\right]} + 2.5\mathbb{E}\left[p_{t,i}\right] \right) + \sum_{t=1}^{T_0} \frac{\eta_t}{2} b_t^2.$$

For the penalty, we use the second part of Lemma 3.2:

$$\text{pen} \leq \sum_{i \neq i_T^*} \sum_{t=1}^{T} 4\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \left(\sqrt{\mathbb{E}\left[p_{t,i}\right]} - \frac{1}{2}\mathbb{E}\left[p_{t,i}\right] \right) + 1.$$

We put the two bounds together and first group the $\sqrt{\mathbb{E}\left[p_{t,i}\right]}$ terms and $\mathbb{E}\left[p_{t,i}\right]$ terms.

$$\begin{aligned} R_T \leq & \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{\eta_t}{2} b_t^2 + 4\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \sqrt{\mathbb{E}\left[p_{t,i}\right]} \\ & + \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{5}{4}\eta_t b_t^2 - 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \mathbb{E}\left[p_{t,i}\right] \\ & + \sum_{t=1}^{T_0} \frac{\eta_t}{2} b_t^2 + 1. \end{aligned}$$

If $\frac{5\eta_t}{4} b_t^2 \geq 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)$ for all t, then the factor in front of $\mathbb{E}\left[p_{t,i}\right]$ is positive and by upper bounding $\mathbb{E}\left[p_{t,i}\right]$ by $\sqrt{\mathbb{E}\left[p_{t,i}\right]}$ and grouping the first and the second summations we obtain

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{7}{4}\eta_t b_t^2 + 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \sqrt{\mathbb{E}\left[p_{t,i}\right]} + \sum_{t=1}^{T_0} \frac{\eta_t}{2} b_t^2 + 1.$$

Otherwise, we upper bound the negative contribution $-2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\mathbb{E}\left[p_{t,i}\right]$ by zero and $\mathbb{E}\left[p_{t,i}\right]$ by $\sqrt{\mathbb{E}\left[p_{t,i}\right]}$ and obtain

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{7}{4}\eta_t b_t^2 + 4\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \sqrt{\mathbb{E}\left[p_{t,i}\right]} + \sum_{t=1}^{T_0} \frac{\eta_t}{2} b_t^2 + 1.$$

Overall, we have

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{7}{4}\eta_t b_t^2 + c\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \sqrt{\mathbb{E}\left[p_{t,i}\right]} + \sum_{t=1}^{T_0} \frac{\eta_t}{2} b_t^2 + 1,$$

where

$$c = \begin{cases} 2, & \text{if } \frac{5\eta_t}{4} b_t^2 \geq 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \text{ for all } t, \\ 4, & \text{otherwise.} \end{cases}$$

Now we use the self-bounding technique (Zimmert and Seldin, 2021). The self-bounding technique states that if L and U are such that $L \leq R \leq U$, then $R \leq 2U - L$. We use the lower bound stated in the theorem, and the upper bound from the previous expression, and we get:

$$\begin{aligned} R_T \leq & \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\left(\frac{7}{2} \eta_t b_t^2 + 2c\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \right) \sqrt{\mathbb{E}\left[p_{t,i}\right]} - \Delta_i b_t \mathbb{E}\left[p_{t,i}\right] \right) + \sum_{t=1}^{T_0} \eta_t b_t^2 + 2 \\ \leq & \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{\left(\frac{7}{2}\eta_t b_t^2 + 2c\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right)\right)^2}{4\Delta_i b_t} + \sum_{t=1}^{T_0} \eta_t b_t^2 + 2, \end{aligned}$$

where we used the fact that each term in the first summation is an expression of the form $a\sqrt{x} - bx$, which for $x \geq 0$ is bounded by $\frac{a^2}{4b}$. □

In Corollary 3.1 we consider a special case, where the losses at each round are bounded by a constant B.

Proof of Corollary 3.1. The learning rate $\eta_t = \frac{2}{B\sqrt{t}}$ is a positive and non-increasing sequence, which allows us to use the results of Theorem 3.2.

The Adversarial Regime In the adversarial regime, we can directly use the learning rate in the first part of Theorem 3.2 and get:

$$R_T \leq \sum_{t=1}^{T} \frac{\eta_t}{2} B^2 \sqrt{K} + \frac{4\sqrt{K}}{\eta_T} + 1 \leq 4B\sqrt{KT} + 1.$$

The Stochastically Constrained Adversarial Regime In order to use the second part of Theorem 3.2, we need to bound the difference between two successive learning rates.

$$\eta_t^{-1} - \eta_{t-1}^{-1} = \frac{B}{2}\left(\sqrt{t} - \sqrt{t-1}\right) \leq \frac{B}{2\sqrt{t}}.$$

We pick $T_0 = 64$, which satisfies that for all $t \geq T_0$, we have $\eta_t B = \frac{2}{\sqrt{T}} \leq \frac{1}{4}$. We note that

$$\frac{5\eta_t}{4} B^2 - 2\left(\eta_t^{-1} - \eta_{t-1}^{-1}\right) \geq \frac{5B}{2\sqrt{t}} - \frac{2B}{2\sqrt{t}} \geq 0.$$

Thus, we have:

$$R_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{81B}{4\Delta_i t} + \sqrt{BT_0} + 2 \leq \sum_{i \neq i^*} \frac{21B\left((\ln T) + 1\right)}{\Delta_i} + 8\sqrt{B} + 2.$$

□

3.9.4 Proofs of Results with Time-Varying Switching Cost

In this regime, the block lengths and the learning rates depend on the sequence of switching costs $(\lambda_n)_{n=1,2,\ldots}$.

Proof of Theorem 3.3. The switching costs are positive, which means that the learning rate $\eta_n = \frac{2\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)}$ is positive and non-decreasing. Thus, we can apply Theorem 3.2 and Lemmas 3.1 and 3.2 through the rest of the proof. We recall that the length of the n-th block is defined as $|B_n| = \max\left\{\left\lceil \frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}} \right\rceil, 1\right\}$.

Thus, we can bound $\frac{\eta_n}{2}|B_n|^2$ as:

$$\begin{aligned}
\frac{\eta_n}{2}|B_n|^2 \leq & \frac{\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)} \left(\frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}} + 1\right)^2 \\
\leq & \frac{\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)} \left(\frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}}\right)^2 \\
& + \frac{2\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)} \left(\frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}}\right) \\
& + \frac{\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^{n} \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)} \\
\leq & \frac{\sqrt{2}\lambda_n}{3\sqrt{K}} + \frac{2\sqrt{2}\sqrt{\lambda_n}}{3\left(K^{1/4}n^{1/4}\right)} + \frac{\sqrt{2}}{3\sqrt{n}} \\
\leq & \frac{\sqrt{2}\lambda_n}{\sqrt{K}} + \frac{\sqrt{2}}{\sqrt{n}},
\end{aligned}$$

where we use that $\frac{1}{\sum_{s=1}^{n}\lambda_s+\frac{\sqrt{K}}{\sqrt{s}}} \leq \frac{1}{\sqrt{Kn}}$ and we deduce that $\frac{\sqrt{\lambda_n}}{\sqrt{\sum_{s=1}^{n}\lambda_s+\frac{\sqrt{K}}{\sqrt{s}}}} \leq \frac{\lambda_n}{\sqrt{K}}+\frac{1}{\sqrt{n}}$ by considering the cases $\lambda_n \geq \frac{\sqrt{K}}{\sqrt{n}}$ and $\lambda_n \leq \frac{\sqrt{K}}{\sqrt{n}}$.

The Adversarial Regime The weighted switching cost on N blocks is upper bounded by $\sum_{n=1}^{N}\lambda_n$. To bound the pseudo-regret, we can directly apply Theorem 3.2 and get that:

$$\begin{aligned}
R_T &\leq \sum_{n=1}^{N}\frac{\eta_n}{2}|B_n|^2\sqrt{K}+\frac{4\sqrt{K}}{\eta_N}+1 \\
&\leq \sum_{n=1}^{N}\sqrt{2}\left(\lambda_n+\frac{\sqrt{K}}{\sqrt{n}}\right)+4\sqrt{K}\frac{3\left(\sum_{s=1}^{n}\lambda_s+\frac{\sqrt{K}}{\sqrt{s}}\right)}{2\sqrt{2}\sqrt{K}}+1 \\
&\leq 4\sqrt{2}\sum_{n=1}^{N}\lambda_n+8\sqrt{2}\sqrt{KN}+1.
\end{aligned}$$

Combining the pseudo regret and the weighted switching cost finishes this part of the proof.

The Stochastically Constrained Adversarial Regime We start by deriving a bound for the stability term. Let N_0 be the smallest number, such that for all $n \geq N_0$, we have $\eta_n|B_n| \leq \frac{1}{4}$. Then, using the last part of Lemma 3.1, we have:

$$\text{stab} \leq \sum_{n=1}^{N}\left(\frac{\sqrt{2}\lambda_n}{\sqrt{K}}+\frac{\sqrt{2}}{\sqrt{n}}\right)\sum_{i\neq i^*}\left(\sqrt{\mathbb{E}\left[p_{n,i}\right]}+2.5\mathbb{E}\left[p_{n,i}\right]\right)+\sum_{n=1}^{N_0}\left(\frac{\sqrt{2}\lambda_n}{\sqrt{K}}+\frac{\sqrt{2}}{\sqrt{n}}\right).$$

We now bound the penalty term. We first need to bound the difference between two successive learning rates.

$$\begin{aligned}
\eta_n^{-1}-\eta_{n-1}^{-1} &= \frac{3}{2\sqrt{2}\sqrt{K}}\left(\sum_{s=1}^{n}\lambda_s+\frac{\sqrt{K}}{\sqrt{s}}-\sum_{s=1}^{n-1}\lambda_s+\frac{\sqrt{K}}{\sqrt{s}}\right) \\
&= \frac{3}{2\sqrt{2}\sqrt{K}}\left(\lambda_n+\frac{\sqrt{K}}{\sqrt{n}}\right).
\end{aligned}$$

Then, we apply the second part of Lemma 3.2 and get:

$$\text{penalty} \leq \sum_{i\neq i^*}\left(\frac{3\sqrt{2}}{\sqrt{K}}\left(\lambda_n+\frac{\sqrt{K}}{\sqrt{n}}\right)\right)\left(\sqrt{\mathbb{E}\left[p_{n,i}\right]}-0.5\mathbb{E}\left[p_{n,i}\right]\right)+1.$$

Adding these expressions together and using the self-bounding technique, we have:

$$R_T \leq \sum_{n=1}^{N} 10\sqrt{2}\left(\frac{\lambda_n}{\sqrt{K}} + \frac{1}{\sqrt{n}}\right)\sum_{i\neq i^*}\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \sum_{i\neq i^*}\sum_{n=1}^{N}\Delta_i|B_n|\mathbb{E}\left[p_{n,i}\right] + \sum_{n=1}^{N_0}\left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + 2.$$

Finally, we use Lemma 3.3 to bound the number of switches, and the fact that $\sqrt{\mathbb{E}\left[p_{n,i}\right]} \geq \mathbb{E}\left[p_{n,i}\right]$, which gives:

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \sum_{n=1}^{N}\sum_{i\neq i^*}\left(\left(11\lambda_n + \lambda_{n+1} + \frac{10\sqrt{2}}{\sqrt{n}}\right)\sqrt{\mathbb{E}\left[p_{n,i}\right]} - \Delta_i|B_n|\mathbb{E}\left[p_{n,i}\right]\right) + \sum_{n=1}^{N_0}\left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + \lambda_1 + 2.$$

We then observe that for each term in the first summation, we can upper bound the expression by replacing $\mathbb{E}\left[p_{n,i}\right]$ by $x_{n,i} \in [0, \infty)$ and maximizing each term independently on $[0, \infty)$.

Thus, we have:

$$R\left(T, (\lambda_n)_{n\geq 1}\right) \leq \sum_{n=1}^{N}\sum_{i\neq i^*}\frac{\left(11\lambda_n + \lambda_{n+1} + \frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i|B_n|} + \sum_{n=1}^{N_0}\left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + \lambda_1 + 2.$$

□

We move on to the corollary with a parametric form of switching costs.

Proof of Corollary 3.2. In this setting, we assume that the sequence of switching

costs satisfies $\lambda_n = n^\alpha$ for $\alpha > 0$. We start by upper bounding N_0.

$$\begin{aligned}
\eta_n|B_n| &\leq \frac{2\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^n \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}\right)}\left(\frac{\sqrt{\lambda_n}\sqrt{\sum_{s=1}^n \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}}{\sqrt{K}} + 1\right) \\
&\leq \frac{2\sqrt{2}\sqrt{\lambda_n}}{3\sqrt{\sum_{s=1}^n \lambda_s + \frac{\sqrt{K}}{\sqrt{s}}}} + \frac{2\sqrt{2}\sqrt{K}}{3\left(\sum_{s=1}^n \lambda_s\right)} \\
&\leq \frac{2\sqrt{2}n^{\alpha/2}}{3\sqrt{\frac{n^{\alpha+1}}{\alpha+1}}} + \frac{2\sqrt{2}}{3\sqrt{n}} \\
&\leq \frac{2\sqrt{2}}{3\sqrt{n}}\left(\sqrt{\alpha+1}+1\right),
\end{aligned}$$

which is a decreasing sequence of n. For all $n \geq \frac{32}{9}\left(\sqrt{\alpha+1}+1\right)^2$, we have $\eta_n|B_n| \leq \frac{1}{4}$, thus we pick $N_0 = \left\lceil \frac{32}{9}\left(\sqrt{\alpha+1}+1\right)^2 \right\rceil$.

Now we move on to bounding the terms $\frac{\left(11\lambda_n+\lambda_{n+1}+\frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i|B_n|}$. Here, the switching costs are increasing, $\lambda_{n+1} \geq \lambda_n$, and we have:

$$\frac{\left(11\lambda_n+\lambda_{n+1}+\frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i|B_n|} \leq \frac{\left(12\lambda_{n+1}+\frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i|B_n|} \leq \frac{\left(12^2\lambda_{n+1}+240\lambda_{n+1}\frac{\sqrt{2}}{\sqrt{n}}+\frac{200}{n}\right)}{4\Delta_i|B_n|}.$$

When $n < N$ the block has not been truncated and the first term is upper bounded as:

$$\frac{12^2\lambda_{n+1}^2}{4\Delta_i|B_n|} \leq \frac{36\sqrt{\alpha+1}\,(n+1)^{2\alpha}\sqrt{K}}{\Delta_i n^{\alpha+1/2}} \leq \frac{36\cdot 4^\alpha\sqrt{\alpha+1}\,(n)^{\alpha-1/2}\sqrt{K}}{\Delta_i},$$

where $|B_n| \geq \frac{n^{\alpha+1/2}}{\sqrt{K}\sqrt{\alpha+1}}$, and for all $n \geq 1$, we have $\frac{(n+1)^2}{n} \leq 4n$. For the case where $n = N$, we can only lower bound $|B_N|$ by 1, and we get:

$$\frac{12^2\lambda_{N+1}^2}{4\Delta_i|B_N|} \leq \frac{36(N+1)^{2\alpha}}{\Delta_i}.$$

The second and third terms are directly upper bounded by lower bounding the block length by 1:

$$\frac{240\frac{\lambda_{n+1}\sqrt{2}}{\sqrt{n}}}{4\Delta_i|B_n|} \leq 60\sqrt{2}\frac{(n+1)^\alpha}{\sqrt{n}\Delta_i} \leq 60\sqrt{2}\cdot 2^\alpha\frac{(n)^{\alpha-1/2}}{\Delta_i},$$

and

$$\frac{\frac{200}{n}}{4\Delta_i|B_n|} \leq \frac{50}{n\Delta_i}.$$

We now sum over n, from 1 to $N-1$, and get:

$$\sum_{n=1}^{N-1} n^{\alpha-1/2} \leq 1 + \int_1^N n^{\alpha-1/2} \leq 1 + (\alpha+1/2)\, N^{\alpha+1/2},$$

which is an upper bound which considers the case where $\alpha - \frac{1}{2} \leq 0$ and where $\alpha - \frac{1}{2} \geq 0$. We finish the proof by combining these results, and we get:

$$\begin{aligned}
&R\left(T, (\lambda_n)_{n\geq 1}\right) \\
&\leq \sum_{n=1}^{N} \sum_{i\neq i^*} \frac{\left(11\lambda_n + \lambda_{n+1} + \frac{10\sqrt{2}}{\sqrt{n}}\right)^2}{4\Delta_i|B_n|} + \sum_{n=1}^{N_0} \left(\frac{2\sqrt{2}\lambda_n}{\sqrt{K}}\right) + 4\sqrt{2N_0} + \lambda_1 + 2 \\
&\leq \sum_{i\neq i^*} \left(\frac{36\cdot 4^\alpha \sqrt{\alpha+1}\left(\alpha+\frac{1}{2}\right) N^{\alpha+1/2}\sqrt{K}}{\Delta_i} + \frac{36\cdot 4^\alpha\sqrt{\alpha+1}\sqrt{K}}{\Delta_i}\right) + \frac{36\,(N+1)^{2\alpha}}{\Delta_i} \\
&\quad + \sum_{i\neq i^*} \left(\frac{60\sqrt{2}\cdot 2^\alpha\left(\alpha+\frac{1}{2}\right)N^{\alpha+1/2}}{\Delta_i} + \frac{60\sqrt{2}\cdot 2^\alpha}{\Delta_i}\right) + \frac{60\sqrt{2}\cdot 2^\alpha N^{\alpha-1/2}}{\Delta_i} \\
&\quad + \sum_{i\neq i^*} \frac{50\ln N}{\Delta_i} + \frac{50}{\Delta_i} + \frac{2\sqrt{2}\left(\frac{32}{9}\left(\sqrt{\alpha+1}+1\right)^2+2\right)^{\alpha+1}}{(\alpha+1)\sqrt{K}} + 11\left(\sqrt{\alpha+1}\right) + 20.
\end{aligned}$$

We now upper bound N. We first note that the length of the n-th block is lower bounded by $\frac{n^{\alpha/2}\sqrt{\sum_{s=1}^n s^\alpha}}{\sqrt{K}}$. Using the fact that $\alpha > 0$, we can lower bound $\sum_{s=1}^n s^\alpha \geq \int_0^n s^\alpha ds = \frac{n^{\alpha+1}}{\alpha+1}$. Thus, we have $|B_n| \geq \frac{n^{\alpha+1/2}}{\sqrt{(\alpha+1)K}}$. Since the block length is an increasing function, for any $\bar{N}$ we have:

$$\sum_{n=1}^{\bar{N}} |B_n| \geq \int_0^{\bar{N}} \frac{n^{\alpha+1/2}}{\sqrt{(\alpha+1)K}} = \frac{\bar{N}^{\alpha+3/2}}{\left(\alpha+\frac{3}{2}\right)\sqrt{(\alpha+1)K}}.$$

We observe that $\bar{N} = \left(\alpha+\frac{3}{2}\right)^{\frac{2}{2\alpha+3}} (\alpha+1)^{\frac{1}{2\alpha+3}} K^{\frac{1}{2\alpha+3}} T^{\frac{2}{2\alpha+3}}$ satisfies $\frac{\bar{N}^{\alpha+3/2}}{(\alpha+\frac{3}{2})\sqrt{(\alpha+1)K}} = T$. Thus, we are sure that N is upper bounded by

$\left(\alpha+\frac{3}{2}\right)^{\frac{2}{2\alpha+3}}(\alpha+1)^{\frac{1}{2\alpha+3}}K^{\frac{1}{2\alpha+3}}T^{\frac{2}{2\alpha+3}}+1$. All that remains is to upper bound N in the pseudo-regret bound.

$$
\begin{aligned}
R\left(T,(\lambda_n)_{n\geq 1}\right) \leq & \sum_{i\neq i^*}\frac{36\cdot 4^{\alpha}\sqrt{\alpha+1}\left(\alpha+\frac{1}{2}\right)\left(\alpha+\frac{3}{2}\right)^{\frac{2\alpha+1}{2\alpha+3}}(\alpha+1)^{\frac{\alpha+1/2}{2\alpha+3}}T^{\frac{2\alpha+1}{2\alpha+3}}K^{\frac{2\alpha+2}{2\alpha+3}}}{\Delta_i} \\
& + \sum_{i\neq i^*}\frac{36\left(\alpha+2\right)\cdot 4^{\alpha}\sqrt{\alpha+1}\sqrt{K}}{\Delta_i}+\frac{36\left(N+1\right)^{2\alpha}}{\Delta_i} \\
& + \sum_{i\neq i^*}\frac{60\sqrt{2}\cdot 2^{\alpha}\left(\alpha+\frac{1}{2}\right)N^{\alpha+1/2}}{\Delta_i}+\frac{60\sqrt{2}\cdot 2^{\alpha}}{\Delta_i}+\frac{60\sqrt{2}\cdot 2^{\alpha}N^{\alpha-1/2}}{\Delta_i} \\
& + \sum_{i\neq i^*}\frac{50\ln T}{\Delta_i}+\frac{50}{\Delta_i}+\frac{2\sqrt{2}\left(\frac{32}{9}\left(\sqrt{\alpha+1}+1\right)^2+2\right)^{\alpha+1}}{(\alpha+1)\sqrt{K}} \\
& + 11\left(\sqrt{\alpha+1}\right)+20.
\end{aligned}
$$

□

3.9.5 Supplementary Experiments

In this section, we present additional experiments highlighting the robustness of Tsallis-Switch. In all the experiments we take $K=8$. Similar results were observed for other values of K.

First, we consider stochastically constrained adversarial sequences. We take a setting, inspired by Zimmert and Seldin (2021), where the environment alternates between two phases. In the first one, the expected loss of the best arm is 0, and the expected loss of the suboptimal arms is Δ. In the second phase, the expected loss of the best arm is $1-\Delta$, and the expected loss of suboptimal arms is 1. At all rounds, the gap between the expected loss of the best arm and any other arm remains constant. In this experiment, the environment generates phases of exponentially increasing length with he i^{th} phase starting at index 1.6^i. We observe in Figures 3.4 and 3.5 that the BaSE algorithm with arithmetic blocks is not robust in this regime. BaSE algorithm with geometric blocks performs really well against this sequence. Tsallis-Switch performs well in both experiments, achieving a regret with switching costs similar to algorithms without blocks when the switching cost is small, and a much better performance when switching becomes costly.

In the second experiment we construct an adversarial sequence that easily breaks BaSE with both arithmetic and geometric grids. We also observe the behavior of

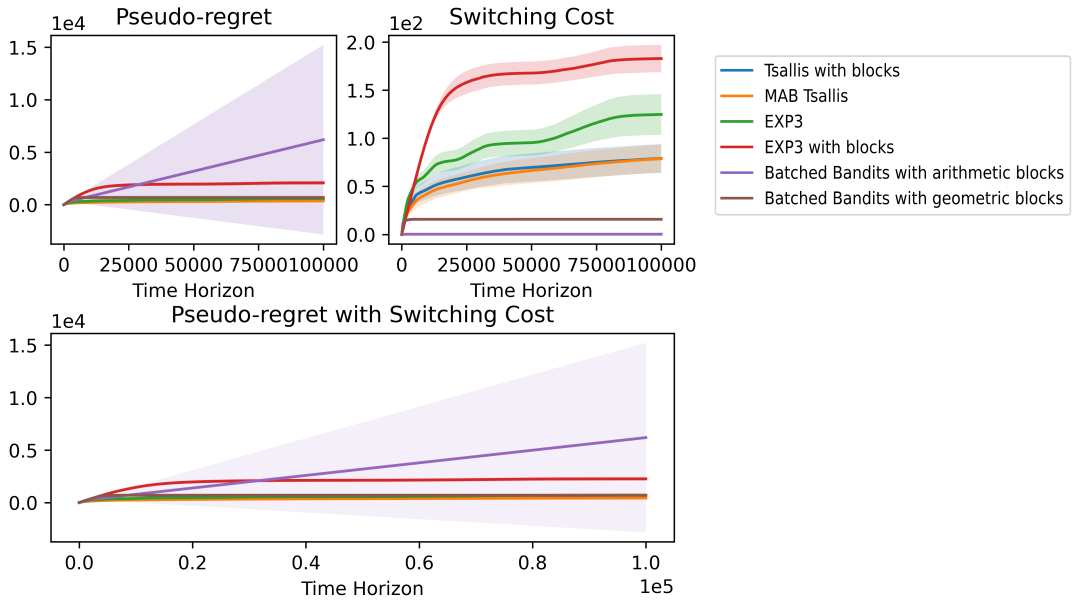

Figure 3.4: Stochastically constrained adversarial losses, $\Delta = 0.2$ and $\lambda = 0.025$.

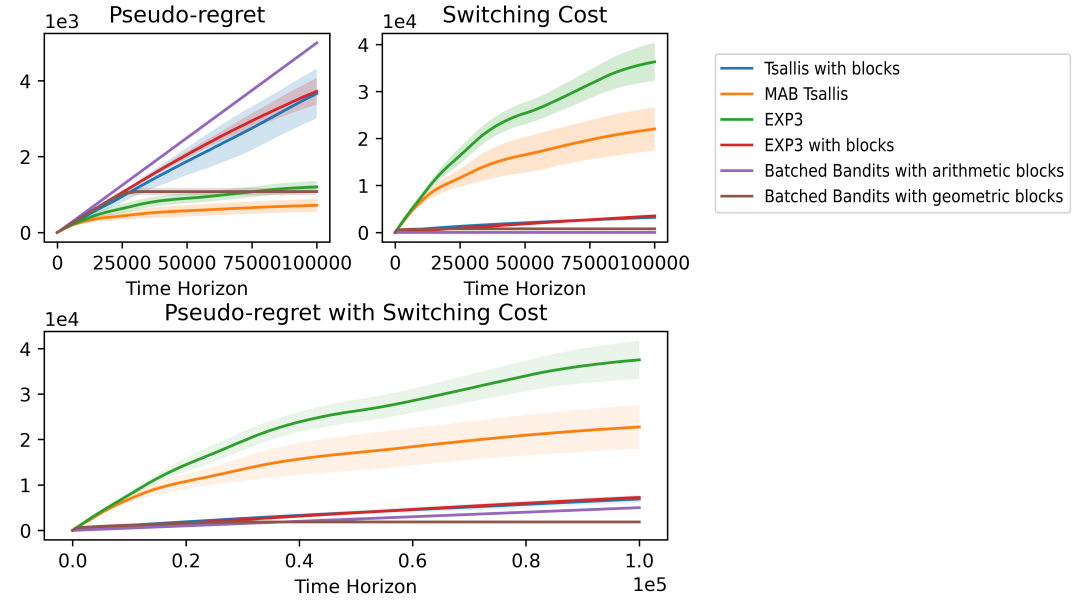

Figure 3.5: Stochastically constrained adversarial losses. $\Delta = 0.05$ and $\lambda = 1$.

Remark 3.1. *The shaded area represents one standard deviation above and below the average measured on 10 repetitions of the experiment. On Figure 3.4, the standard deviation of Batched Bandits is large because the algorithm eliminates the optimal arms in some of the runs of the experiment, but not all of them.*

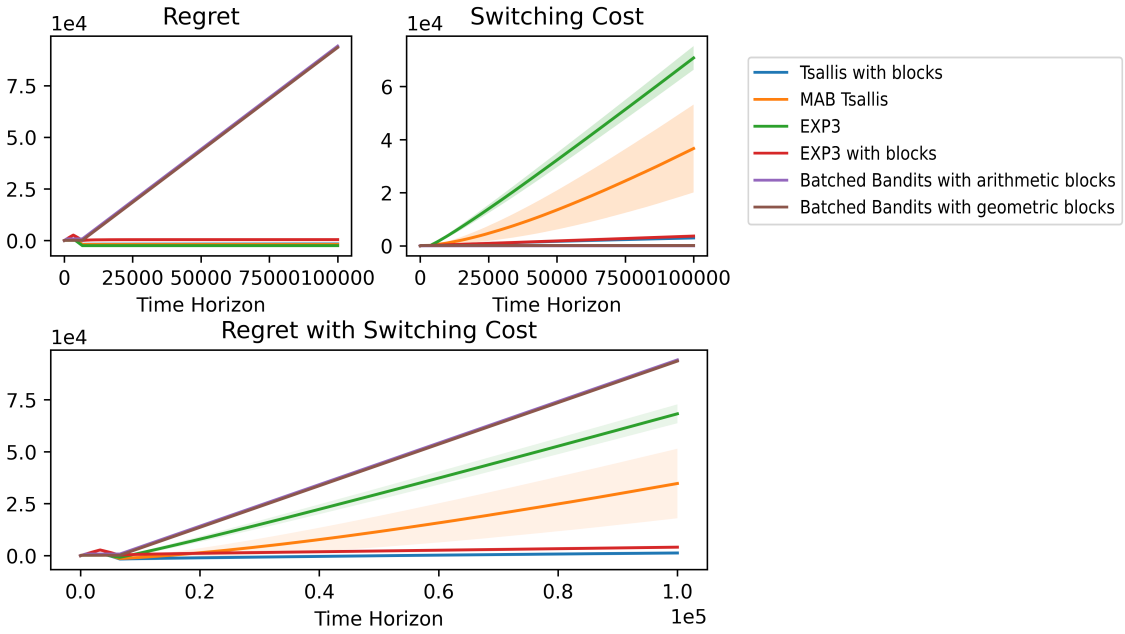

Figure 3.6: Regret against a deterministic adversarial sequence described in the text with $\lambda = 1$. The shaded area represents one standard deviation above and below the average measured on 10 repetitions of the experiment. The curves for batched bandits with arithmetic blocks and batched bandits with geometric blocks almost coincide and they are the highest ones.

the other algorithms in this context. The sequence of losses is constructed in the following way: in the first $\sqrt{KT\ln(KT)}$ rounds, one arm suffers a loss of 0, while all the other arms suffer a loss of 1. After the $\sqrt{KT\ln(KT)}$ rounds the losses are reversed, so the first arm suffers a loss of 1 and all other arms suffer a loss of 0. In Figure 3.6 we observe that the BaSE algorithm with both arithmetic and geometric grid suffers linear regret, as it, with high probability, eliminates the best arm based on the first rounds. We can see that with this sequence, Tsallis-Switch achieves both a very low regret and a low number of switches, even though at the end of the game, $K-1$ arms have the same performance, and only one is suboptimal.

In the last experiment we test robustness of Tsallis-Switch in a stochastic setting with several best arms and a stochastically constrained adversarial setting with several best arms. We take $\Delta = 0.2$ and $\lambda = 1$ and change the number of optimal arms from 1 to 7 while keeping the total number of arms $K = 8$. We recall that Zimmert and Seldin (2021) experimentally observed that in a stochastic setting without

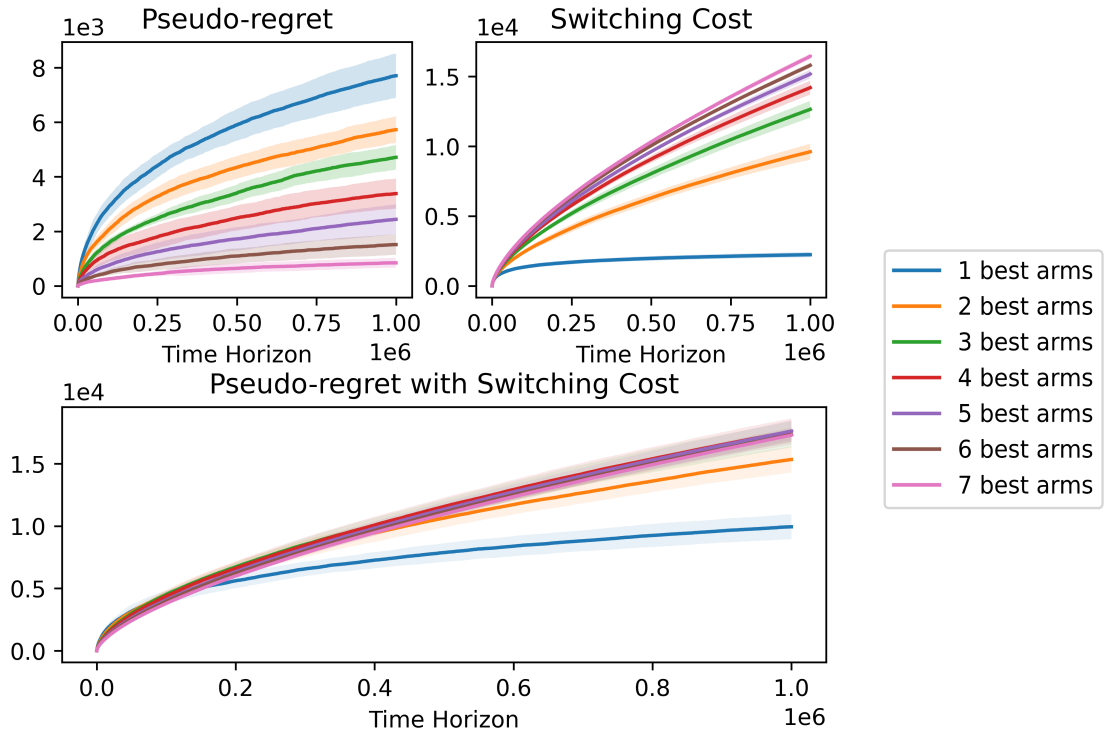

Figure 3.7: The performance of Tsallis-Switch under stochastic losses and several optimal arms. $K = 8$, $\Delta = 0.2$ and $\lambda = 1$. The shaded area represents one standard deviation above and below the average measured on 10 repetitions of the experiment.

switching costs the regret of Tsallis-INF decreases with the increase of the number of best arms, suggesting that the requirement on uniqueness of the best arm is an artifact of the analysis, rather than a real limitation of the algorithm. In Figures 3.7 and 3.8 we observe that in the setting with switching costs the picture is different, because switching between best arms is costly. We note that Tsallis-Switch still has the adversarial regret guarantee of $\mathcal{O}\big((\lambda K)^{1/3}T^{2/3} + \sqrt{KT}\big)$ in both settings, so the regret is still under control, but there is a clear increase in the regret as the number of optimal arms grows beyond 1. Therefore, the experiments seem to suggest that the improved regret scaling with $T^{1/3}$ only holds under the assumption on uniqueness of the best arm and elimination of this assumption will require modification of the algorithm.

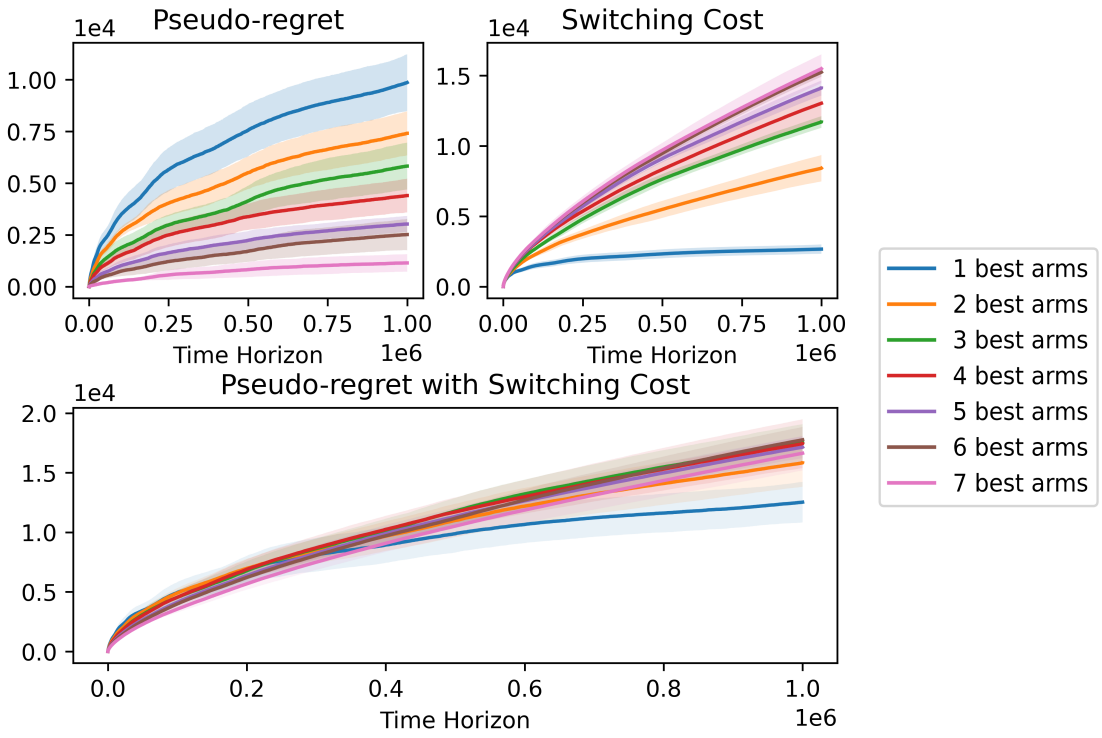

Figure 3.8: The performance of Tsallis-Switch under stochastically constrained adversarial losses and several optimal arms. $K = 8$, $\Delta = 0.2$ and $\lambda = 1$. The shaded area represents one standard deviation above and below the average measured on 10 repetitions of the experiment.

Chapter 4

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

The work presented in this chapter is based on a paper that has been published as:

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

4.1 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.

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 a $\min\big\{O\big(\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T}\big), O(\sqrt{\tilde{\alpha} T \ln K})\big\}$ 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 4.1 (Informal)**.** *Given a directed feedback graph $G = (V, E)$ with independence number α and strong independence number $\tilde{\alpha}$, there exists an algorithm (Algorithm 3) whose pseudo-regret can simultaneously be bounded by $\min\left\{O(\sqrt{\tilde{\alpha}T\ln K}), O\big(\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T}\big)\right\}$ 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 3 requires neither prior knowledge of the type of the environment (adversarial or stochastic), nor the time horizon.

4.1.1 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., 2002b) 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., 2002b). 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.

4.2 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 $\left\{(i, \ell_{t,i}) : i \in N^{\text{out}}(I_t)\right\}$, where $N^{\text{out}}(i) = \{j \in V : (i, j) \in E\}$ is the out-neighbourhood of i. Similarly, we define $N^{\text{in}}(i) = \{j \in V : (j, i) \in E\}$ 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. $i^* = \arg\min_{i \in V} \mathbb{E}[\ell_{1,i}]$. The pseudo-regret can then be expressed in terms of the suboptimality gaps $\Delta_i = \mathbb{E}[\ell_{1,i} - \ell_{1,i^*}]$,

$$\mathcal{R}_T = \sum_{t=1}^{T} \sum_{i \in V} \mathbb{E}\left[p_{t,i}\right] \Delta_i, \tag{4.1}$$

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.

Definition 4.1. *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.

4.3 Algorithm

We present the EXP3.G++ algorithm (Algorithm 3), 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 4. 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 4, which is used by EXP3.G++ to update the exploration rates $\varepsilon_{t,i}$ according to Equation (4.2). Algorithm 4 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

Algorithm 3: 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 (4.2)
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**

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

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

$$\forall i \in V : \hat{\Delta}_{t,i} = \max\left\{0, \mathrm{LCB}_{t,i} - \min_j \mathrm{UCB}_{t,j}\right\}$$

$\forall i \in V$: 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}.

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

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

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

end for

of that arm. The exploration set can be constructed in $O(K^3)$ time, but note that 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 4.1 below.

Proposition 4.1. *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 4 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 $i \in N^{\mathrm{out}}(j)$ 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 4. Since G contains self-loops, if $i \in S$, then

Algorithm 4: 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

$i \in N^{\text{out}}(i)$ 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,

$$\varepsilon_{t,i} = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}, \xi_{t,i}\right\}, \tag{4.2}$$

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

$$\xi_{t,i} = \begin{cases} (\beta \ln t)/(t\hat{\Delta}_{t,i}^2), & \text{if } i \in S_t, \\ 4/t^2, & \text{otherwise,} \end{cases} \tag{4.3}$$

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 $P_{t,i} = \mathbb{P}\left[i \in N^{\text{out}}(I_t)\right]$ 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,

$$P_{t,i} \geq o_{t,i} = \min\left\{\frac{1}{2K}, \frac{1}{2}\sqrt{\frac{\lambda \ln K}{tK^2}}, \frac{\beta \ln t}{t\hat{\Delta}_{t,i}^2}\right\}. \tag{4.4}$$

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 3) 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 3. Concentration of the gap estimates around the true gaps is crucial for bounding the regret in the stochastic setting.

Lemma 4.1. *If Algorithm 3 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 (4.4) 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 $t \geq t_{\min}(i)$,

$$\mathbb{P}\left[\hat{\Delta}_{t,i} \leq \frac{1}{2}\Delta_i\right] \leq \left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2} + \frac{2}{Kt^{\gamma-1}} + 2\left(\frac{1}{t}\right)^{\frac{\beta}{10}}. \tag{4.5}$$

A proof of the lemma is provided in Section 4.8.3.

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 4.1, if $\gamma = 4$ and $t \geq t_{\min}(i)$, we have that $\frac{(\ln t)^2}{t} \leq \frac{\lambda \Delta_i^4 \ln K}{4K^2\beta^2}$, implying

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

4.4 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:

$$\theta_t := \sum_{i \in V} \frac{p_{t,i}}{P_{t,i}}, \tag{4.6}$$

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 (4.6) in terms of the strong independence number (Lemma 4.8) and the independence number (Lemma 4.7) 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 4.2. *Assume that Algorithm 3 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 (4.2)–(4.3) 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 4.8.2.

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.

4.5 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 4.2 holds.

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

Theorem 4.3. *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 4.2 and choosing $\lambda = \alpha$, the pseudo-regret of Algorithm 3 against any stochastic stochastic loss sequence, satisfies:*

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

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 4.8.4.

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.

$$R_T = R_{t_{\min}} + \sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \mathbb{E}\left[p_{t,i}\right] \leq R_{t_{\min}} + \sum_{t=t_{\min}}^{T} \sum_{i\,:\,\Delta_i>0} \Delta_i \big(\mathbb{E}\left[q_{t,i}\right] + \mathbb{E}\left[\varepsilon_{t,i}\right]\big). \quad (4.7)$$

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,

$$R_{t_{\min}} = C_0\sqrt{\alpha\ t_{\min}} \log\left(t_{\min}\right) = C_1 \frac{K}{\Delta_{\min}^2}\left(\ln\left(\frac{K}{\Delta_{\min}}\right)\right), \quad (4.8)$$

where the first equality follows from the second part of the bound presented in Theorem 4.2 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 4.1. These concentration inequalities allow us to show that the two following propositions hold.

Proposition 4.2 (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 4.3 (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 4.8.4. 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 (4.7) 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 4.2, 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 4.1. *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 4.2, the pseudo-regret of Algorithm 3 against any stochastic stochastic loss sequence, satisfies:*

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

4.6 Extension to Time Varying Feedback Graphs

The results presented in Theorem 4.3 and Corollary 4.1 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 (4.2) with $\lambda = 1$ to ensure that the exploration is never too large. This exploration rate allows us to apply Lemma 4.1 with $\lambda = 1$, and derive the following result.

Theorem 4.4. *Assume that Algorithm 3 is run on a sequence of arbitrarily generated feedback graphs $G_1, G_2, \ldots$ with learning rate $\eta_t = \sqrt{\frac{\ln K}{2\sum_{s=K}^{t-1}\theta_s}}$ and exploration rates defined in* (4.2) *and* (4.3) *with $\lambda = 1$, $\gamma = 4$ and $\beta = 320$. Then the pseudo-regret against any oblivious loss sequence satisfies*

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

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 n^{th} 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 4.8.5.

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 4.2. In the stochastic regime, using $\lambda = 1$ allows us to obtain the same tight constants as in Corollary 4.1, 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.

4.7 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.

4.8 Appendix

4.8.1 Tools to Bound Series

We use the following lemmas to bound series.

Lemma 4.2 (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 4.3 (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 4.4. *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 4.3 in the last step. □

Lemma 4.5 (Lemma 3 (Thune and Seldin, 2018))**.** *For $c > 0$ we have*

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

4.8.2 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 4.6 (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 4.7 (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 4.2. *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 4.4. *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 $\mathrm{mas}(G) \leq \tilde{\alpha}(G)$ which finishes the proof. □

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

Lemma 4.8 (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:*

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

With those results, we can move on to the proof of Theorem 4.2.

Proof of Theorem 4.2. 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:

$$\begin{aligned}
\mathcal{R}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] \\
&\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T} \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=K+1}^{T} \ell_{t,i}\right] \\
&= K + \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} p_{t,i} \mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] \\
&\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} q_{t,i} \mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T} \sum_{i=1}^{K} \varepsilon_{t,i} \mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right],
\end{aligned} \tag{4.9}$$

where $i^* = \arg\min\left\{\sum_{t=K+1}^{T} \mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right\}$, and $\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] = \ell_{t,i}$. Equation (4.9) 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:

$$\begin{aligned}\mathcal{R}_T &\leq K + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K} q_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right] - \sum_{t=K+1}^{T}\mathbb{E}_t\left[\tilde{\ell}_{t,i^*}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K}\varepsilon_{t,i}\mathbb{E}_t\left[\tilde{\ell}_{t,i}\right]\right] \\ &\leq K + \mathbb{E}\left[\mathbb{E}_t\left[\frac{\ln K}{\eta_T}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}_t\left[\sum_{i\in V}\frac{\eta_t}{2}\frac{q_{t,i}}{P_{t,i}}\right]\right] + \mathbb{E}\left[\sum_{t=K+1}^{T}\sum_{i=1}^{K}\mathbb{E}_t\left[\varepsilon_{t,i}\tilde{\ell}_{t,i}\right]\right].\end{aligned} \tag{4.10}$$

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:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}_t\left[\sum_{i\in V}\frac{\eta_t}{2}\frac{q_{t,i}}{P_{t,i}}\right]\right] &\leq \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}_t\left[\eta_t\sum_{i\in V}\frac{p_{t,i}}{P_{t,i}}\right]\right] \\ &\leq \sqrt{\ln K}\ \mathbb{E}\left[\sum_{t=K+1}^{T}\mathbb{E}_t\left[\frac{\theta_t}{\sqrt{2\sum_{s=K}^{t-1}\theta_s}}\right]\right] \\ &\leq \sqrt{2\ln K}\ \mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{K},\end{aligned} \tag{4.11}$$

where equation (4.11) follows from Lemma 4.4.

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 4.10 gives:

$$\mathcal{R}_T \leq 2K + 2\sqrt{2\ln K}\ \mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{\lambda T \ln K}. \tag{4.12}$$

All that remains is to bound θ and λ. To obtain the first part of the bound, we use Proposition 4.4 and Lemma 4.8, and deduce that for all $t \leq K+1$:

$$\theta_t \leq \tilde{\alpha}\ \text{, which gives } \sqrt{\sum_{t=K+1}^{T}\theta_t\ } \leq \sqrt{\tilde{\alpha}T}. \tag{4.13}$$

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

$$\begin{aligned}\mathcal{R}_T &\leq 2K + 2\sqrt{2\ln K}\ \sqrt{\tilde{\alpha}T} + \sqrt{\tilde{\alpha}T\ln K} \\ &\leq 4\sqrt{\tilde{\alpha}T\ln K} + 2K. \end{aligned}\tag{4.14}$$

For the second part of the bound, we use Lemma 4.7 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:

$$\begin{aligned}\mathcal{R}_T &\leq 2K + 2\sqrt{2\ln K}\ \sqrt{8\alpha T\ln(KT)} + \sqrt{\ln K}\sqrt{\ln T}\sqrt{\alpha T} \\ &\leq 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\alpha T} + 2K, \end{aligned}\tag{4.15}$$

Taking the minimum between equations (4.14) and (4.15) finishes the proof.

□

4.8.3 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 4.1 in two parts.

4.8.3.1 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 4.9. *The confidence intervals satisfy:*

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

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

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

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

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

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

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

□

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

Proof of the first part of Lemma 4.1.
We recall that $\hat{\Delta}_{t,i} = \max\{0, \mathrm{LCB}_{t,i} - \min_{j\neq i} \mathrm{UCB}_{t,j}\}$. Then using Lemma 4.9, we have:

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

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 $\Delta_i \leq \bar{\Delta}_i$, and we finish the proof with a union bound and applying Lemma 4.9. □

4.8.3.2 Lower bound

To derive a lower bound on the gap estimates and prove the second part of Lemma 4.1, 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 4.3. *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 4.10. *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 4.10. 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:

$$\begin{aligned}
\mathbb{P}\left[\overline{\mathcal{E}(i,t)}\right] &= \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : o_{s,i} \leq \frac{\beta \ln t}{t\Delta_i^2}\right] \\
&\leq \mathbb{P}\left[\exists s \in \left[\frac{t\Delta_i^2}{\ln t}, t\right] : \hat{\Delta}_{s,i} \geq \Delta_i\right] \qquad (4.16) \\
&\leq \sum_{s=\frac{t\Delta_i^2}{\ln t}} \frac{1}{s^{\gamma-1}} \\
&\leq \frac{1}{2}\left(\frac{\ln t}{t\Delta_i^2}\right)^{\gamma-2},
\end{aligned}$$

where the last summation follows from Lemma 4.2. The proof of the second inequality is similar, Equation (4.16) 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 4.5 (Theorem 8 (Seldin and Lugosi, 2017))**.** *Let $X_1, \dots, X_n$ be Bernoulli random variables adapted to filtration $\mathcal{F}_1, \dots, \mathcal{F}_n$ (in particular, X_s may depend on $X_1, \dots, X_{s-1}$). 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 4.5 on the $t-K$ random variables $\mathbb{1}[i \in N^{\text{out}}(I_s)]$ 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 4.9, we know that the upper and lower confidence bounds satisfy: $\mathbb{P}\left[\mathrm{UCB}_{t,i^*} \leq \mu_{i^*} \vee \mathrm{LCB}_{t,i} \geq \mu_i\right] \leq \frac{2}{Kt^{\gamma-1}}$. Then assuming that $\mathrm{UCB}_{t,i^*} \geq \mu_{i^*}$ and $\mathrm{LCB}_{t,i} \leq \mu_i$, we have:

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

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:

$$\begin{aligned}
\hat{\Delta}_{t,i} \geq & \Delta_i - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i}}} - 2\sqrt{\frac{\gamma \ln\left(tK^{1/\gamma}\right)}{2O_{t-1,i^*}}} \\
\geq & \Delta_i - 4\sqrt{\frac{2\Delta_i^2 \gamma \ln\left(tK^{1/\gamma}\right)}{2\beta \ln t}} \\
\geq & \Delta_i - 4\sqrt{\frac{\Delta_i^2 (\gamma+1) \ln\left(tK^{1/\gamma}\right)}{\beta \ln t}} \qquad (4.17) \\
= & \Delta_i \left(1 - 4\sqrt{\frac{\gamma+1}{\beta}}\right),
\end{aligned}$$

where equation (4.17) 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:

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

4.8.4 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.

4.8.4.1 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 4.5. *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, $\max_t\{t \leq c(\ln t)^2\}$ 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 $b \geq e$:

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

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

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

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

Proposition 4.6. *Under the conditions of Lemma 4.1 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 4.5, we have that $t_{\min} \leq 25d\,(\ln d)^2$, where $d = \frac{4\beta^2K^2}{\lambda\ln K\Delta_{\min}^4}$.

Then, we can use the first half Theorem 4.2 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 4.2, and we deduce:

$$
\begin{aligned}
& R_{t_{\min}} \\
& \leq 9\sqrt{\alpha t_{\min}}\sqrt{\ln(Kt_{\min})}\sqrt{\ln K} + 2K \\
& \leq 9\sqrt{\alpha \ln K \ 25d}\ln(d)\sqrt{\ln\left(25Kd^2\right)} + 2K && (4.18) \\
& \leq 9\sqrt{\alpha \ln K \ 25\frac{4\beta^2K^2}{\lambda \ln K\Delta_{\min}^4}}\ln\left(\frac{4\beta^2K^2}{\lambda \ln K\Delta_{\min}^4}\right)\sqrt{\ln\left(25K\left(\frac{4\beta^2K^2}{\lambda \ln K\Delta_{\min}^4}\right)^2\right)} + 2K \\
& \leq 9 * 5 * 2\frac{\beta K}{\Delta_{\min}^2}\sqrt{\frac{\alpha}{\lambda}}\ln\left(\frac{4\beta^2K^2}{\lambda \ln K\Delta_{\min}^4}\right)\sqrt{\ln\left(\frac{400\beta^4K^5}{\Delta_{\min}^8}\right)} + 2K && (4.19) \\
& \leq \frac{1019\beta K}{\Delta_{\min}^2}\sqrt{\frac{\alpha}{\lambda}}\left(\ln\left(\frac{\beta K}{\Delta_{\min}}\right)\right)^{3/2} + 2K.
\end{aligned}
$$

where the equation (4.18) uses that for $d > 1$ we have $d\left(\ln d\right)^2 \leq d^2$, and equation (4.19) uses that $400 \leq 320^4 \leq \beta^4$.

□

4.8.4.2 Control over the Exponential Weights

Proposition 4.2 introduced in the proof sketch of Theorem 4.3 is based on the following result.

Proposition 4.7. *Under the conditions of Lemma 4.1 with $\gamma = 4$, $\beta = 320$ and $\lambda \in [1, K]$, the sum of exponential weights with sequence of learning rates $\eta_1, \eta_2, \ldots$ 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 4.6 (Bernstein's inequality for martingales)**.** *Let $X_1, \ldots, X_n$ be a martingale difference sequence with respect to filtration $\mathcal{F}_1, \ldots, \mathcal{F}_n$, 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:

$$\begin{aligned}
X_s &= \Delta_i - (\tilde{\ell}_{s,i} - \tilde{\ell}_{s,i^*}) \\
&\leq 1 + \tilde{\ell}_{s,i^*} \\
&\leq 1 + \frac{1}{P_{t,i^*}} \\
&\leq 1 + \max\left\{2K, 2\sqrt{\frac{sK^2}{\lambda \ln K}}, \frac{s\hat{\Delta}_{s,i^*}^2}{\beta \ln s}\right\} \qquad (4.20) \\
&\leq 1.25 \max\left\{2K, 2\sqrt{\frac{sK^2}{\lambda \ln K}}, \frac{s\hat{\Delta}_{s,i^*}^2}{\beta \ln s}\right\}
\end{aligned}$$

where equation (4.20) holds by definition of o_{t,i^*}. Using the same argument as in the proof of Lemma 4.1, 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:

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

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.

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

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:

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

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

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

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

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

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:

$$\begin{aligned}\sqrt{2\nu \ln\left(\frac{1}{\delta}\right)} + \frac{\kappa \ln\left(\frac{1}{\delta}\right)}{3} &= \sqrt{2\frac{2t^2\Delta_i^2}{\beta \ln t}\ln(Kt)} + \frac{\frac{1.25t\Delta_i^2}{\beta \ln t}\ln(Kt)}{3} \\ &\leq \sqrt{8\frac{t^2\Delta_i^2}{\beta \ln t}\ln(t)} + \frac{\frac{1.25t\Delta_i^2}{\beta \ln t}\ln t}{3} \qquad (4.21) \\ &\leq t\Delta_i\left(\frac{2\sqrt{2}}{\sqrt{\beta}} + \frac{2.5}{3\beta}\right) \\ &\leq \frac{1}{2}t\Delta_i,\end{aligned}$$

where equation (4.21) is due to $t \geq t_{\min}(i) \geq K$, so $\ln(Kt) \leq 2\ln t$. We can then use Theorem 4.6 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.

4.8.4.3 Control over the Exploration

We now provide a more general version of Proposition 4.3.

Proposition 4.8. *Let $S_1, S_2, \dots$ be a sequence of exploration sets generated by playing algorithm 3 the conditions of Lemma 4.1 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:

$$\sum_{t=t_{\min}}^{T} \sum_{i:\Delta_i>0} \Delta_i \frac{4}{t^2} \leq K \sum_{t=1}^{T} \frac{4}{t^2} \leq 8K. \tag{4.22}$$

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:

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

□

4.8.4.4 Proof of Theorem 4.3 and Corollary 4.1

The proof of Theorem 4.3 follows from the propositions in this section.

Proof of Theorem 4.3. We want to bound the pseudo-regret of algorithm 3 run with parameters defined in Lemma 4.1 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.

$$\begin{aligned}
R_T &= R_{t_{\min}} + \sum_{i\,:\,\Delta_i > 0} \sum_{t=t_{\min}}^{T} \Delta_i \mathbb{E}\left[p_{t,i}\right] \\
&\leq R_{t_{\min}} + \sum_{i\,:\,\Delta_i > 0} \Delta_i \sum_{t=t_{\min}}^{T} \left(\mathbb{E}\left[q_{t,i}\right] + \mathbb{E}\left[\varepsilon_{t,i}\right]\right), \qquad (4.23)
\end{aligned}$$

First, we apply the second part of Proposition 4.6 with $\lambda = \alpha$, and deduce that:

$$R_{t_{\min}} \leq \frac{1019\beta K}{\Delta_{\min}^2}\left(\ln\left(\frac{\beta K}{\Delta_{\min}}\right)\right)^{3/2} + 2K. \tag{4.24}$$

Then we bound the contribution of exponential weights by applying Proposition 4.7 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:

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

and then:

$$\begin{aligned}
\sum_{i:\Delta_i>0}\Delta_i\sum_{t=t_{\min}}^{T}\mathbb{E}\left[q_{t,i}\right] &\leq \sum_{i:\Delta_i>0}\Delta_i\left(\frac{6K}{\Delta_i^2} + \ln T\left(\frac{\alpha\ln K}{4K^2\beta^2} + \frac{1}{K}\right)\right) \\
&\leq \ln T\left(\frac{\alpha\ln K}{4K\beta^2} + 1\right) + \sum_{i:\Delta_i\geq 0}\frac{16K}{\Delta_i},
\end{aligned} \tag{4.25}$$

where the last step follows from Lemma 4.5. Furthermore, we bound the contribution of the extra exploration by applying Proposition 4.8 with $\lambda = \alpha$, which gives:

$$\begin{aligned}
\sum_{i\,:\,\Delta_i>0}\Delta_i\sum_{t=t_{\min}(i)}^{T}\mathbb{E}\left[\varepsilon_{t,i}\right] &\leq \sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\in S_t:\Delta_i>0}\frac{4\beta\ln t}{t\Delta_i}\right] + \frac{\alpha\ln K}{4K\beta^2}\ln T + 12K + 3 \\
&\leq \max_{Ind\in\mathcal{I}(G)}\left\{\sum_{i\in Ind:\Delta_i>0}\frac{4\beta\ln^2 T}{\Delta_i}\right\} + \frac{\alpha\ln K}{4K\beta^2}\ln T + 12K + 3,
\end{aligned} \tag{4.26}$$

where the last step follows from Proposition 4.1: 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 (4.24), (4.25) and (4.26) 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 4.1 follows the same structure.

Proof of Corollary 4.1. We decompose the regret following equation (4.23), where t_{min} is defined using $\lambda = \tilde{\alpha}$.

$R_{t_{min}}$ is bounded the first part of Proposition 4.5, which gives:

$$R_{t_{min}} \leq \frac{160\beta K}{\Delta_{\min}^2}\sqrt{\frac{\tilde{\alpha}}{\lambda}}\ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K. \tag{4.27}$$

Then $\sum_{i\,:\,\Delta_i>0}\Delta_i\sum_{t=t_{\min}}^{T}\mathbb{E}\left[q_{t,i}\right]$ is bounded by Proposition 4.7 with $\lambda = \tilde{\alpha}$, and using the derivation leading to (4.25), which gives:

$$\sum_{i\,:\,\Delta_i>0}\Delta_i\sum_{t=t_{\min}}^{T}\mathbb{E}\left[q_{t,i}\right] \leq \ln T\left(\frac{\tilde{\alpha}\ln K}{4K\beta^2}+1\right) + \sum_{i:\Delta_i\geq 0}\frac{16K}{\Delta_i}, \tag{4.28}$$

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 (4.26), which gives:

$$\sum_{i\,:\,\Delta_i>0}\Delta_i\sum_{t=t_{\min}(i)}^{T}\mathbb{E}\left[\varepsilon_{t,i}\right] \leq \max_{Ind\in\mathcal{I}(G)}\left\{\sum_{i\in Ind:\Delta_i>0}\frac{4\beta\ln^2 T}{\Delta_i}\right\} + \frac{\tilde{\alpha}\ln K}{4K\beta^2}\ln T + 12K + 3, \tag{4.29}$$

Finally, summing equations (4.27), (4.28) and (4.29) finishes the proof. □

4.8.5 Extension to Graphs that Change over Time

Proof of Theorem 4.4.

Adversarial Regime

In the adversarial regime, the proof follows the analysis with a fixed feedback graph up to equation (4.12), which gives:

$$\mathcal{R}_T \leq 2K + 2\sqrt{2\ln K}\;\mathbb{E}\left[\mathbb{E}_t\left[\sqrt{\sum_{t=K+1}^{T}\theta_t}\right]\right] + \sqrt{T\ln K}, \tag{4.30}$$

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 4.8 and Proposition 4.4 to deduce that for all $t \geq K+1$:

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

and using this bound on θ in equation (4.30) gives:

$$\mathcal{R}_T \leq 4\sqrt{\ln K \sum_{t=1}^{T} \tilde{\alpha}_t} + 2K. \tag{4.31}$$

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 4.7 at each round $t \geq K+1$, which gives:

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

using this bound on θ in equation (4.30) gives:

$$\mathcal{R}_T \leq 9\sqrt{\ln K}\sqrt{\ln(KT)}\sqrt{\sum_{t=1}^{T} \alpha_t} + 2K. \tag{4.32}$$

Taking the minimum over equations (4.31) and (4.32) finishes the proof.

Stochastic Regime
The structure of the proof follows from Theorem 4.3. We decompose the regret following equation (4.23), where t_{min} is chosen using $\lambda = 1$.

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

$R_{t_{min}}$ is bounded using the same approach as for Corollary 4.1, but using the time varying version of the bound given in equation (4.31). We bound $\tilde{\alpha}_t \leq K$ at each round, which gives:

$$R_{t_{min}} \leq \frac{160\beta K^{3/2}}{\Delta_{\min}^2} \ln\left(\frac{\sqrt{\beta}K}{\Delta_{\min}}\right) + 2K. \tag{4.33}$$

Then the second term is bounded by Proposition 4.7 with $\lambda = 1$, and using the derivation leading to equation (4.25), which gives:

$$\sum_{i\,:\,\Delta_i > 0} \Delta_i \sum_{t=t_{\min}}^{T} \mathbb{E}\left[q_{t,i}\right] \leq \ln T \left(\frac{\ln K}{4K\beta^2} + 1\right) + \sum_{i:\Delta_i \geq 0} \frac{16K}{\Delta_i}, \tag{4.34}$$

The last term follows Proposition 4.8 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 n^{th} 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 4.8 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

$$\begin{aligned}\sum_{t=t_{\min}}^{T} \mathbb{E}\left[\sum_{i\in S_t:\Delta_i\geq 0} \Delta_i \mathbb{E}\left[\varepsilon_{t,i}\right]\right] \leq \inf_{0\leq n\leq T}\left\{\max_{S\subset V:|S|=\tilde{\alpha}_n}\left\{\sum_{i\in S:\Delta_i>0} \frac{4\beta \ln^2 T}{\Delta_i}\right\} + n\right\} \\ + \frac{\ln T \ln K}{4K\beta^2} + 12K + 3. \qquad (4.35)\end{aligned}$$

We finish the proof by summing on equations (4.33), (4.34) and (4.35).

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

□

Chapter 5

Summary and Discussion

In Chapter 2, we studied the problem of decoupling exploration and exploitation in multi-armed bandits. We proposed an algorithm based on Tsallis-Inf that achieved a tight bound in the adversarial regime, as well as a time-independent bound in the stochastically constrained adversarial regime given a correct parametrization of the regularization function with $\alpha \in \left(\frac{1}{2}; \frac{2}{3}\right]$. Using two arms with their own distributions allows the learner to bypass the normal trade-off between exploration and exploitation, giving the algorithm the opportunity to simultaneously explore and exploit more. Whether this result can be improved further to eliminate the $\sqrt{K}$ factor in the stochastic regime without affecting the adversarial result remains an open question.

In Chapter 3, we considered the problem of multi-armed bandits with switching costs. We introduced an algorithm which enjoys an optimal rate in the adversarial regime as well as refined bounds in the stochastically constrained adversarial regime that scales as $O\big(\lambda^{2/3}K^{2/3}T^{1/3}\sum_{i\neq i^*}\Delta_i^{-1}\big)$. This result left a significant gap with existing $O(\log T\sum_{i\neq i^*}\Delta_i^{-1})$ results in the stochastic regime with a fixed switching cost $\lambda = 1$ (Gao et al., 2019; Esfandiari et al., 2021). Since then, Amir et al. (2022) have derived a lower bound showing that if an algorithm achieves an optimal rate of $O\big(K^{1/3}T^{2/3}\big)$ against adversarial sequences of losses, then this algorithm must suffer at least $\tilde{\Omega}\big(\min\big\{\frac{1}{\Delta_{\min}^2}, K^{1/3}T^{2/3}\big\}\big)$ pseudo-regret against a sequence of stochastically constrained adversarial losses with minimal sub-optimality gap $\Delta_{\min}$. This bound opens many questions regarding the difficulty of the problem in intermediate regimes, and of the possible trade-off between them.

In Chapter 4, we investigated the problem of online learning with feedback graphs.

Our approach combines the EXP3.G and EXP3++ algorithms with a novel arm-dependent extra exploration, which is tuned by taking advantage of the graph structure. Our algorithm enjoys near-optimal guarantees in both the adversarial and the stochastic regime, and generalizes to sequences of graphs that change over time. Improving upon the current results poses interesting challenges. One consideration is to improve upon our exploration set construction by solving a linear program instead, which could balance the cost and the benefit of playing each arm. Buccapatnam et al. (2014, 2017) study this approach in the stochastic regime. However, the biggest challenge comes from closing the suboptimality gaps in terms of T in the stochastic regime. Concurrently with our work, Ito et al. (2022) studied the same problem and proposed a different algorithm also based on EXP3.G. They derived more general results that have a weaker dependency on T but that depend on the independence number of the graph rather than its strong independence number. It is uncertain whether obtaining optimal best-of-both-worlds results for this problem can be achieved using the regularization function of the EXP3 algorithm, but changing the regularization function could come at the cost of bounds not scaling with the optimal graph dependent quantity.

List of Publications

The work presented in this thesis has lead to the following publications.

1. 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.

2. 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.

3. Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bibliography

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

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.

Idan Amir, Guy Azov, Tomer Koren, and Roi Livni. Better best of both worlds bounds for bandits with switching costs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peter Auer. Using confidence bounds for exploration-exploitation trade-offs. *Journal of Machine Learning Research*, 2002.

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, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002a.

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

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Avrim Blum and Yishay Mansour. Learning, regret minimization, and equilibria. In Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani, editors, *Algorithmic game theory*. Cambridge University Press, 2007.

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

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in finitely-armed and continuously-armed bandits. *Theoretical Computer Science*, 2011.

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.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 2007.

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.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 2014.

Ofer Dekel, Ambuj Tewari, and Raman Arora. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the Annual Symposium on the Theory of Computing (STOC)*, 2013.

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.

Hossein Esfandiari, Amin Karbasi, Abbas Mehrabian, and Vahab Mirrokni. Regret bounds for batched bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 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.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 2006.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1997.

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.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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 and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2015.

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.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Tor Lattimore and Csaba Svepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

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*, 1994.

Haipeng Luo and Robert E. Schapire. Achieving all with no parameters: Adanormalhedge. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2015.

Thodoris Lykouris, Éva Tardos, and Drishti Wali. Feedback graph regret bounds for Thompson Sampling and UCB. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 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.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 2004.

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.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.

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.

Chloé Rouyer, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Amir Sani, Gergely Neu, and Alessandro Lazaric. Exploiting easy data in online optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

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.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 2012.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning*, 2019.

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

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 1988.

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. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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.