UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Saeed Masoudian

Best-of-Both-Worlds Learning in Bandits with Delayed Feedback

Advisor: Yevgeny Seldin

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on August 31 2023.

Abstract

This thesis addresses two pivotal challenges in Multi-armed bandits: achieving *best-of-both-worlds* guarantees and effectively handling *delayed feedback*. In practical scenarios like recommender systems and clinical trials, environments may exhibit a blend of stochastic and adversarial characteristics. Concurrently, delays are prevalent in such applications.

The Tsallis-INF algorithm introduced by Zimmert and Seldin (2019) marked a breakthrough, demonstrating optimal performance in both adversarial and stochastic bandits. Building upon this foundation, our thesis refines the Tsallis-INF regret bound within intermediate scenarios that bridge stochastic and adversarial environments. We propose a comprehensive analysis that enhances understanding of the self-bounding technique used by Zimmert and Seldin (2019) and yields improved regret bounds for *stochastic with adversarial corruption* and *stochastically constrained adversarial regimes*.

Addressing the challenge of *delayed feedback* in bandits, we establish a best-of-both-worlds regret guarantee. Existing research within the Follow the Regularized Leader (FTRL) framework addresses the delayed problem only in the adversarial regime. We propose a minor adaptation to the algorithm of Zimmert and Seldin (2020), that relies on the knowledge of the maximal delay $d_{\max}$ and ensures control over the drift of the distribution over arms played by the algorithm, thereby realizing a best-of-both-worlds guarantee.

Furthermore, we complement our best-of-both-worlds algorithm for delayed bandits with the *skipping technique* (Zimmert and Seldin, 2020) and *implicit exploration* (Neu, 2015), eliminating the requirement for prior knowledge of $d_{\max}$. These techniques facilitate efficient distribution drift control, further enhancing our established best-of-both-worlds guarantees.

Lastly, we explore leveraging *intermediate observations* to mitigate delay impacts on the learning process. These observations, appearing as finite states S, provide the learner with real-time information. In each round, the corresponding state is revealed *immediately* upon the learner's action, followed by the actual loss after an adversarially set *delay*. We find that the complexity of the problem pivots on the state-loss mapping's nature, rather than the action-state relationship. For adversarial state-loss mappings, intermediate observations yield no advantages. However, in scenarios with stochastic state-loss mappings, we improve worst-case regret, replacing $\sqrt{(K+d)T}$ with $\sqrt{(K+\min\{S,d\})T}$, where d is the fixed delay, T is the time horizon, and K is the number of arms. This improvement extends to arbitrary delay settings, ensuring robust high probability guarantees.

Resumé

Denne afhandling beskæftiger sig med to afgørende udfordringer inden for Multi-armed bandits: opnåelse af garantier for *bedst-ud-af-begge-verdener* og effektiv håndtering af *forsinket feedback*. I praktiske scenarier som anbefalingssystemer og kliniske forsøg kan miljøer udvise en blanding af stokastiske og fjendtlige karakteristika. Samtidig er forsinkelser udbredte i sådanne anvendelser.

Algoritmen Tsallis-INF, introduceret af Zimmert and Seldin (2019), markerede et gennembrud ved at demonstrere optimal ydeevne både i fjendtlige og stokastiske bandits. Byggende på denne grundlæggende viden forfiner vores afhandling Tsallis-INF-regretsgrænsen inden for mellemliggende scenarier, der forbinder stokastiske og fjendtlige miljøer. Vi foreslår en omfattende analyse, der forbedrer forståelsen af den selvafgrænsende teknik, der anvendes af Zimmert and Seldin (2019), og giver forbedrede regretsgrænser for *stokastiske med fjendtlig korruption* og *stokastisk begrænsede fjendtlige regime*.

Ved at håndtere udfordringen med *forsinket feedback* i bandits etablerer vi en garant for det bedste fra begge verdener med hensyn til regrets. Eksisterende forskning inden for Follow the Regularized Leader (FTRL) rammer kun det forsinkede problem i fjendtlige regime. Vi foreslår en mindre tilpasning til algoritmen fra Zimmert and Seldin (2020), der bygger på kendskabet til den maksimale forsinkelse $d_{\max}$ og sikrer kontrol over fordelingens drift over arme, som algoritmen spiller, og realiserer derved en garant for det bedste fra begge verdener.

Desuden supplerer vi vores algoritme for det bedste fra begge verdener til forsinkede bandits med *spring-teknikken* (Zimmert and Seldin, 2020) og *implicit udforskning* (Neu, 2015), hvilket eliminerer kravet om forhåndskendskab til $d_{\max}$. Disse teknikker letter effektiv kontrol af fordelingsdrift, hvilket yderligere forbedrer vores etablerede garantier for det bedste fra begge verdener.

Endelig udforsker vi brugen af *mellemliggende observationer* for at mildne forsinkelsens indvirkning på læringsprocessen. Disse observationer, der fremtræder som begrænsede tilstande S, giver læseren realtidsinformation. I hvert trin afsløres den tilsvarende tilstand *øjeblikkeligt* efter deltagerens handling, efterfulgt af det faktiske tab efter en fjendtligt fastsat *forsinkelse*. Vi finder, at problemets kompleksitet drejer sig om karakteren af sammenhængen mellem tilstand og tab, snarere end mellem handling og tilstand. For fjendtlige sammenhænge mellem tilstand og tab giver mellemliggende observationer ingen fordele. Dog forbedrer vi i scenarier med stokastiske sammenhænge mellem tilstand og tab den værst tænkelige regret ved at erstatte $\sqrt{(K+d)T}$ med $\sqrt{(K+\min\{S,d\})T}$, hvor d er den faste forsinkelse, T er tidsrammen og K er antallet af arme. Denne forbedring strækker sig til vilkårlige

forsinkelsesindstillinger og sikrer robuste garantier med høj sandsynlighed.

Acknowledgements

My PhD was a remarkable journey, embracing new experiences in a foreign land, filled with both exciting "explorations" and, of course, "exploitations". I'm deeply grateful to my supervisor Yevgeny Seldin, whose guidance not only enriched my professional knowledge but also profoundly impacted me on a personal level. His network connected me to wonderful researchers globally, something I'm really proud of.

A special acknowledgment goes to my colleague and friend, Sadegh Talebi, with whom I shared enlightening discussions. Thanks to my colleagues from the Delta group, particularly my fellow PhD students Yi-Shan, Chloé, Hippolyte, Yijie, and Yunlian, who have been incredible companions on this journey. Your companionship made the path enjoyable and smooth and I will certainly miss our gatherings.

I would like to express my appreciation to my close collaborator, Julian Zimmert, who graciously hosted me in Berlin. Despite the brevity of our time together, our collaborations were remarkably fruitful. I am also very grateful to Nicolò Cesa-Bianchi for providing me a chance to visit his group in Milan, which was one of the most cherished aspects of my PhD journey. Thanks to the amazing people in his group, in particular Emmanuel, Dirk, Hao, Khaled, Pierre, Matilde, and Giulia for their friendship and the unforgettable memories we created.

Special gratitude to my parents for their blessings and support. Your encouragement has been always a driving force for me. Lastly, and most importantly, my heartfelt thanks to my wife Leila. She has been with me every step of the way, transforming challenges into opportunities and providing me constant support. As we continue to explore new horizons together, I am profoundly grateful for her presence in my life.

[I acknowledge funding from European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 801199.]

Table of Contents

Chapter 1

Introduction

The field of online learning includes a diverse array of problems connected to real life challenges. At its core lies the fundamental problem of the Multi-armed Bandit, a consequential challenge that has attracted extensive attention due to its applicability in a wide range of domains. Multi-armed bandits, with its roots in statistical decision theory and sequential analysis, encapsulates the essence of adaptive decision-making under uncertainty.

The simplicity inherent to the Multi-armed Bandit framework provides a fertile ground for analytical exploration. Nevertheless, the significance of the Multi-armed Bandit extends far beyond its theoretical confines, finding practical relevance in an array of real-world applications. One compelling domain where the Multi-armed Bandit paradigm shines is adaptive clinical trials. Here, the exploration-exploitation trade-off mirrors the challenge of experimenting with new treatments while capitalizing on the most promising options. Moreover, the flexibility of the Multi-armed Bandit framework makes it well-suited for enhancing recommendation systems, where it facilitates the balance between exploring uncharted options and exploiting well-performing choices. Although it might not fully captures the complexities of the real world, the Multi-armed Bandit framework still acts as an essential building block. Particularly, it significantly influences more structured domains in Online Learning such as Reinforcement Learning, Partial Monitoring, and Monte-Carlo Tree Search, leading to important progress that goes beyond its basic setup.

The Multi-armed Bandit problem can be visualized as a sequential decision-making game with an agent, often referred to as the *learner*, who interacts with an *environment* over a series of rounds. At each round the learner selects an action, so called *arm*, among a set of arms and incurs the associated loss with that arm. These losses are determined by the environment, and the learner *only* observes the

loss of the chosen action. The learner's objective is to minimize the cumulative regret - the gap between her accrued losses and the losses that would have been incurred by choosing the best action consistently.

Within the realm of Multi-armed Bandits, two conventional assumptions underpin the fabric of the problem. The first is related to the nature of the environment, mandating it to be either *fully stochastic*, where all losses are drawn i.i.d. from certain distributions, or *adversarial*, allowing arbitrary choices of losses within the $[0, 1]$ interval. The second assumption is *immediate feedback*, that dictates that the learner *promptly* observes the loss corresponding to her chosen action. However, these assumptions, while providing a foundation for analysis, can be violated in many real-world scenarios. This thesis focuses on the bandits works on both of these challenges.

In the bandit literature, there has been extensive research on both the stochastic regime (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a) and the adversarial setting (Auer et al., 2002b). However, these algorithms typically make strong assumptions about the type of environment, while real-world scenarios can be different. In practical applications, the environment might not be purely stochastic or completely adversarial. For instance, a recommender system, where user preferences usually follow a certain pattern, but occasional changes in behavior due to factors like mood can disrupt this pattern. In such cases, algorithms that are designed for fully stochastic regime might perform poorly, suffering linear regret. On the other hand, using adversarial algorithms isn't entirely fair either, as they only guarantee regret for worst-case scenarios, not accounting for slight disruptions in the stochastic environment. To address this challenge, there has been recent interest in developing algorithms that can *simultaneously* work well in both the stochastic and the adversarial regimes, without any prior knowledge about the type of the environment (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018). The ultimate goal of these works is to achieve the optimal bounds for both regimes, so called best-of-both-worlds guarantees. Unlike previous attempts that had limitations in one of the regimes, Zimmert and Seldin (2019) addressed this by introducing the Tsallis-INF algorithm, achieving optimal bounds in both regimes. Remarkably, this algorithm had been proposed previously for the adversarial regime by Audibert and Bubeck (2009, 2010) and Abernethy et al. (2015), but Zimmert and Seldin reanalyzed it for the stochastic regime, achieving logarithmic regret. Operating within the Follow the Regularized Leader (FTRL) framework, the algorithm uses a kind of regularization known as α-*Tsallis Entropy* (Tsallis, 1988), which has inspired others to seek "best-of-both-worlds" guarantees in various settings. The analysis of Tsallis-INF has been extended

further by Zimmert and Seldin (2021) to intermediate regimes between stochastic and adversarial environments, including *stochastically constrained adversarial regime* (Wei and Luo, 2018) and *stochastic bandits with adversarial corruptions* Lykouris et al. (2018). Yet, the analysis of the Tsallis-INF algorithm reveals a significant drawback in such intermediate scenarios. This limitation stems from the fact that the provided regret does not seamlessly bridge the gap between the two regimes. In this thesis, we will delve into this issue to propose a solution.

This thesis tackles another bandit challenge: *delayed feedback*, a natural occurrence in real-world applications. For instance, in clinical trials, there's a delay between giving patients a medication and seeing its effects. Decentralized recommender systems also face delays in communication. In the stochastic setting, Joulani et al. (2013) proved that the impact of fixed delay d is only an additive term $\mathcal{O}(d)$ on the regret, which does not grow with time. However, delays become more impactful in the adversarial setting. Cesa-Bianchi et al. (2019) demonstrated that fixed delay of d could lead to $\mathcal{O}(\sqrt{dT})$ regret. Later, Bistritz et al. (2019) and Thune et al. (2019) extended this to arbitrary delay scenarios but with a requirement of knowing the total delay beforehand. All these adversarial regime studies operate within the FTRL framework, using negative entropy as the regularizer. However, Zimmert and Seldin (2020) introduced a novel FTRL-based algorithm, employing a combination of Tsallis entropy and negative Shannon entropy for regularization. Furthermore, they introduced an effective *skipping technique* to skip rounds with significantly large delays. While, their algorithm achieves a minimax optimal bound in the adversarial regime with arbitrary delays, but it remained uncertain whether it also guarantees logarithmic performance in the stochastic setting. We show an adaptation to their algorithm that secures a best-of-both-worlds guarantee.

As we have observed, delay's impact on the adversarial regime grows at $\sqrt{dT}$ rate. Hence, if d is substantial, the delay cost could be significantly large. Yet, in many practical scenarios, there are *intermediate observations* available with no delay. For example, in medical trials, we can measure intermediate symptoms like blood pressure and heart rate when prescribing medication. This naturally leads to the question: Can the effect of delays be mitigated by utilizing these intermediate observations? While this problem has been addressed by Vernade et al. (2020) for non-stationary regime, in this thesis we tackle this problem in adversarial regime.

1.1 Outline of the Thesis

In the following we provide the structure of the thesis.

Chapter 2 introduces an improved analysis of Tsallis-INF algorithm provided by Zimmert and Seldin (2021). The analysis of Tsallis-INF algorithm in both stochastic with adversarial corruption, and stochastically constrained adversarial regimes is based on the self-bounding technique, however the analysis of adversarial setting requires a different approach. We provide a single analysis for all the regimes that not only provides a better understanding of self-bounding technique but also provides better results in both stochastic with adversarial corruption, and stochastically constrained adversarial regimes.

Chapter 3 and Chapter 3 both consider the problem of multi-armed bandits with arbitrary delays. While this issue has been addressed separately in adversarial and stochastic scenarios, no best-of-both-worlds solutions exist. In Chapter 3, we propose a modification to the state-of-the-arts algorithm for adversarial regime by Zimmert and Seldin (2020). Our modification ensures control over the drift of the distribution over arms played by the algorithm by utilizing information about the maximum delay $d_{\max}$. This enables us to achieve the first-ever best-of-both-worlds guarantee for this problem.

In Chapter 4, we empower the algorithm introduced in Chapter 3 with two techniques: the *skipping technique* (Zimmert and Seldin, 2020) and *implicit exploration* (Neu, 2015). These techniques allow distribution drift control without requiring any prior knowledge like $d_{\max}$. This advancement improves further the best-of-both-worlds guarantees established in Chapter 3.

In Chapter 5 we consider the problem of delayed bandits with intermediate observations. In this problem, the learner takes an action, observes an *intermediate state* from set of states, and suffers the loss of her action. However, the actual loss is observed after a certain delay. We address the fundamental question: *When do intermediate observations help in delayed bandits?*. We examine this question across different scenarios of the action-state mapping and state-loss mapping. While this problem has been studied by Vernade et al. (2020) within the non-stationary regime for action-state mapping and the stochastic regime for state-loss mapping, our analysis extends to all various scenarios, including both stochastic and adversarial mappings for each scenario.

Finally in Chapter 6, we comprehensively discuss the obtained results and potential future works.

1.2 Main Contributions

- We provide an alternative analysis for the Tsallis-INF algorithm by Zimmert and Seldin (2021). Let Δ_is be the suboptimal gaps, then the new analysis ensures
$$\mathcal{O}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log\left(T\frac{K-1}{(\sum_{i\neq i^*}1/\Delta_i)^2}\right)\right)$$
regret bound for both stochastically constrained adversarial regime and stochastic regime with adversarial corruption with small corruption level as $C \leq \sum_{i\neq i*}\frac{1}{\Delta_i}\big(\big(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\big)+1\big)$. Furthermore for the large amount of corruptions as $C \geq \sum_{i\neq i*}\frac{1}{\Delta_i}\big(\big(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\big)+1\big)$, we show Tsallis-INF achieves
$$\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log_+\left(T\frac{K-1}{C(\sum_{i\neq i^*}1/\Delta_i)}\right)}\right),$$
where $\log_+(x)=\max(1,\log x)$. Our bound also ensures a smooth transition from the optimal bounds of fully stochastic ($C=0$) and fully adversarial ($C=T$) regimes as we increase C from 0 to T.

- We improve the bound by Zimmert and Seldin (2021) in stochastic bandits with adversarial corruptions with the corruption budget $C\in[0,T]$, by a factor of $\sqrt{\frac{\log T}{\log T/C}}$. When $C=\mathcal{O}(T/\log T)$, this improvement can lead to a substantial improvement, reaching the order of $\sqrt{\frac{\log T}{\log\log T}}$.

- In the stochastically constrained adversarial regime, we improve over Zimmert and Seldin (2021) by replacing $\sum_{i\neq i^*}\frac{1}{\Delta_i}\log(T)$ with $\sum_{i\neq i^*}\frac{1}{\Delta_i}\log\big(T\frac{K-1}{(\sum_{i\neq i^*}1/\Delta_i)^2}\big)$.

- We provide a best-of-both-worlds analysis for Tsallis-INF algorithm that offers improved insights into the self-bounding analysis of this algorithm. This stems from our unified analysis, which covers fully adversarial, stochastic with adversarial corruption, and stochastically constrained adversarial regimes, whereas Zimmert and Seldin (2021) has a separate analysis for the fully adversarial case. Furthermore, our approach can be extended to improve the regret for the other variations of the Tsallis-INF algorithm in the bandit setting and beyond such as the work by Jin and Luo (2020).

- We provide a modification to the algorithm by Zimmert and Seldin (2020), that with an oracle-level knowledge of the maximum delay $d_{\max}$, simultaneously achieves near optimal regret bounds for both the adversarial and stochastic bandits with arbitrary delays.
- We show that the regret lower bound for adversarial bandits with arbitrary delays is $\Omega\left(\sqrt{KT} + \min_S(|S| + \sqrt{D_{\bar{S}} \log K})\right)$, where $D_{\bar{S}} = \sum_{t \in [T]/S} d_t$. Our lower bound shows optimality of the regret derived by Zimmert and Seldin (2020).
- We establish a best-of-both-worlds results for the *fixed delay* setting of bandits. This includes an *optimal* regret bound within the adversarial regime, alongside a *near-optimal* bound for the stochastic scenario.
- We lift the the assumption about prior knowledge of $d_{\max}$ in arbitrary delays regime by proposing an enhanced version of our best-of-both-worlds algorithm.
- Our new algorithm offers two improvements in best-of-both-worlds guarantee: firstly, it substitutes all occurrences of $d_{\max}$ with the maximum number of outstanding observations $\sigma_{\max}$; and secondly, it incorporates the possibility of skipping large delays, thus improving the dependence on the total delay. Notably, $\sigma_{\max}$ can be considerably smaller than $d_{\max}$ when dealing with unbalanced delays.
- We present an effective method for controlling distribution drift in delayed bandits. This technique operates within the FTRL framework, offering the possibility of its application in other delay-related works built upon FTRL principles.
- We demonstrate that the complexity of the problem involving delayed bandits with intermediate observations is primarily determined by the state-loss mapping, regardless of whether the action-state mapping is stochastic or adversarial.
- We prove that when the state-loss mapping is adversarial, incorporating intermediate observations yields no advantages to the learner.
- We show that when the state-loss mapping is stochastic, intermediate observations can be utilized to replace $\sqrt{(K+d)T}$ with $\sqrt{(K + \min\{S, d\})T}$, where the former is the regret of bandits with fixed delays. Here S represents the number of states, d is the fixed delay, and K is the number of actions.

- Finally, we extend our previous result to the arbitrary delays setting and attain high probability guarantees.

Chapter 2

Improved Analysis of the Tsallis-INF Algorithm in Stochastically Constrained Adversarial Bandits and Stochastic Bandits with Adversarial Corruptions

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

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 Conference on Learning Theory (COLT)*, 2021.

Abstract

We derive improved regret bounds for the Tsallis-INF algorithm of Zimmert and Seldin (2021). We show that in adversarial regimes with a (Δ, C, T) self-bounding constraint the algorithm achieves $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\Delta_i^{-1}\right)^2}\right) + \sqrt{C\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{C\sum_{i\neq i^*}\Delta_i^{-1}}\right)}\right)$ regret bound, where T is the time horizon, K is the number of arms, Δ_i are the suboptimality gaps, i^* is the best arm, C is the corruption magnitude, and $\log_+(x) = \max(1, \log x)$. InThe regime includes stochastic bandits, stochastically constrained adversarial bandits, and stochastic bandits with adversarial corruptions as special cases. Additionally, we provide a general analysis, which allows to achieve the same kind of improvement for generalizations of Tsallis-INF to other settings beyond multiarmed bandits.

2.1 Introduction

Most of the literature on multiarmed bandits is focused either on the stochastic setting (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a) or on the adversarial one (Auer et al., 2002b). However, in recent years there has been an increasing interest in algorithms that perform well in both regimes with no prior knowledge of the regime (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018), as well as algorithms that perform well in intermediate regimes between stochastic and adversarial (Seldin and Slivkins, 2014; Lykouris et al., 2018; Wei and Luo, 2018; Gupta et al., 2019). The quest for best-of-both-worlds algorithm culminated with the work of Zimmert and Seldin (2019), who proposed the Tsallis-INF algorithm and showed that its regret bound in both stochastic and adversarial environments matches the corresponding lower bounds within constants with no need of prior knowledge of the regime. Zimmert and Seldin (2020) further improved the analysis and introduced an *adversarial regime with a self-bounding constraint*, which is an intermediate regime between stochastic and adversarial environments, including *stochastically constrained adversaries* (Wei and Luo, 2018) and *stochastic bandits with adversarial corruptions* (Lykouris et al., 2018) as special cases. They have shown that the Tsallis-INF algorithm achieves the best known regret rate in this regime and its special cases.

The Tsallis-INF algorithm is based on regularization by Tsallis entropy with power $\frac{1}{2}$, which was also used in the earlier works by Audibert and Bubeck (2009,

2010) and Abernethy et al. (2015) for minimax optimal regret rates in the adversarial regime. The key novelty of the work of Zimmert and Seldin (2019, 2020) is an analysis of the algorithm in the stochastic setting based on a self-bounding property of the regret. The idea has been subsequently extended to derive best-of-both-worlds algorithms for combinatorial semi-bandits (Zimmert et al., 2019), decoupled exploration and exploitation (Rouyer and Seldin, 2020), bandits with switching costs (Rouyer et al., 2021), and ergodic MDPs (Jin and Luo, 2020).

We present a refined analysis based on the self-bounding property, which improves the regret bound in the adversarial regime with a self-bounding constraint and its special cases: stochastic bandits, stochastically constrained adversarial bandits, and stochastic bandits with adversarial corruption. The adversarial regime with a self-bounding constraint is defined in the following way. Let $\ell_1, \ell_2, \dots$ be a sequence of loss vectors with $\ell_t \in [0,1]^K$, let I_t be the action picked by the algorithm at round t, and let $\overline{Reg}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,I_t}\right] - \min_i \mathbb{E}\left[\sum_{t=1}^T \ell_{t,i}\right]$ be the pseudo-regret. For a triplet (Δ, C, T) with $\Delta \in [0,1]^K$ and $C \geq 0$, Zimmert and Seldin (2020) define an *adversarial regime with a* (Δ, C, T) *self-bounding constraint* as an adversarial regime, where the adversary picks losses, such that the pseudo-regret of any algorithm at time T satisfies

$$\overline{Reg}_T \geq \sum_{t=1}^{T} \sum_i \Delta_i \mathbb{P}(I_t = i) - C.$$

(The above condition is only assumed to be satisfied at time T, but there is no requirement that it is satisfied at time $t < T$.) A special case of this regime is the stochastically constrained adversarial regime, where $\overline{Reg}_T = \sum_{t=1}^T \sum_i \Delta_i \mathbb{P}(I_t = i)$ with Δ being the vector of suboptimality gaps. In particular, the stochastic regime is a special case of the stochastically constrained adversarial regime. (In the stochastic regime the expected loss of each arm is fixed over time. Stochastically constrained adversarial regime relaxes this requirement by only assuming that the expected gaps between the losses of pairs of arms are fixed, but the expected losses are allowed to fluctuate over time.) Another special case of an adversarial regime with a self-bounding constraint are stochastic bandits with adversarial corruptions. For two sequences of losses $\overline{\mathcal{L}}_T = (\bar{\ell}_1, \dots, \bar{\ell}_T)$ and $\mathcal{L}_T = (\ell_1, \dots, \ell_T)$ the amount of corruption is measured by $\sum_{t=1}^T \|\bar{\ell}_t - \ell_t\|_\infty$. In stochastic bandits with adversarial corruptions the adversary takes a stochastic sequence of losses and injects corruption with corruption magnitude bounded by C. Zimmert and Seldin (2020) show that a stochastic, as well as a stochastically constrained adversarial regime with a vector of suboptimality gaps Δ and injected corruption of magnitude bounded by C, satisfy $(\Delta, 2C, T)$ self-bounding constraint. As C grows from zero to T, the stochastic regime with

Setting	Zimmert and Seldin (2020)	Our paper
Small C	$\mathcal{O}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T\right)$	$\mathcal{O}\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\log_+\left(T\frac{K-1}{(\sum_{i\neq i^*}1/\Delta_i)^2}\right)\right)$
Large C	$\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log T}\right)$	$\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}\log_+\left(T\frac{K-1}{C(\sum_{i\neq i^*}1/\Delta_i)}\right)}\right)$

Table 2.1: Comparison of the leading terms in the regret bounds of Zimmert and Seldin (2020) and our paper, differences are highlighted in color. We define $\log_+(x) = \max(1, \log x)$. The "Small C" row compares the regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints with $C \leq \sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\right)+1\right)$. Here, C is a subdominant term and does not show up in the big-$\mathcal{O}$ notation. The "Large C" row compares the regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints with $C \geq \sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\right)+1\right)$. The regret bounds in the adversarial regime are identical, and hence omitted.

adversarial corruptions interpolates between stochastic and adversarial bandits.

Lykouris et al. (2018) were the first to introduce and study stochastic bandits with adversarial corruptions and their algorithm achieved $\mathcal{O}\left(\sum_{i:\Delta_i>0}\frac{KC+\log(T)}{\Delta_i}\log(T)\right)$ regret bound. Gupta et al. (2019) improved it to $\mathcal{O}\left(KC+\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\log^2(KT)\right)$. Zimmert and Seldin (2020) have shown that their best-of-both-worlds Tsallis-INF algorithm achieves $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{\log T}{\Delta_i}\right)+\sqrt{C\sum_{i\neq i^*}\frac{\log T}{\Delta_i}}\right)$ regret bound in the more general adversarial regime with (Δ, C, T) self-bounding constraint under the assumption that Δ has a unique zero entry (the assumption corresponds to uniqueness of the best arm *before* corruption). Neither of the algorithms requires prior knowledge of C.

Our contributions are summarized in the enumerated list below. The improvements relative to the work by Zimmert and Seldin (2020) are further highlighted in Table 2.1.

1. We present a refined analysis of the regret of Tsallis-INF in adversarial regimes with a (Δ, C, T) self-bounding constraint, achieving

$$\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)+\sqrt{C\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)}\right)$$

regret bound, where $\log_+(x) = \max(1, \log x)$.

2. In the stochastically constrained adversarial regime it improves the dominating term of the regret bound from $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log T\right)$ to $\mathcal{O}\left(\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)\right)$ relative to the work of Zimmert and Seldin (2020), see Table 2.1. A similar kind of improvement has been studied for UCB-type algorithms for stochastic bandits by Auer and Ortner (2010) and Lattimore (2018).

3. In the stochastic regime with adversarial corruptions the result yields an improvement by a multiplicative factor of $\mathcal{O}\left(\sqrt{\log T/\log\left(T/C\right)}\right)$ relative to the work of Zimmert and Seldin (2020), see Table 2.1 for a more refined statement. In particular, for $C=\Theta\left(\frac{TK}{(\log T)\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)$ it achieves an improvement by a multiplicative factor of $\sqrt{\frac{\log T}{\log\log T}}$.

4. While the analysis of Zimmert and Seldin (2020) used two different optimization problems to analyze the regret of Tsallis-INF in adversarial environments and in adversarial environments with a self-bounding constraint, we obtain both bounds from the same optimization problem. This provides continuity in the analysis in the sense that the $\mathcal{O}\left(\sqrt{KT}\right)$ adversarial regret bound is obtained as a natural limit case of the adversarial bound with a self-bounding constraint as C grows beyond $\mathcal{O}\left(\frac{KT}{\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)$. It also provides a better understanding of the self-bounding analysis technique.

5. We also provide a more general result, showing that any algorithm with adversarial pseudo-regret bound satisfying $\overline{Reg}_T \leq B\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{\mathbb{E}[w_{t,i}]}{t}}$, where $w_{t,i}$ are the probabilities of playing action i at round t and B is a constant, achieves

$$\mathcal{O}\left(B^2\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\Delta_i^{-1}\right)^2}\right)+B\sqrt{C\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)\log_+\left(\frac{KT}{C\sum_{i\neq i^*}\Delta_i^{-1}}\right)}\right)$$

regret in the adversarial regime with (Δ, C, T) self-bounding constraint. The result can be directly applied to achieve improved regret bounds for extensions of the Tsallis-INF algorithm, for example, the extension to episodic MDPs (Jin and Luo, 2020).

2.2 Problem Setting

We study multi-armed bandit problem in which at time $t = 1, 2, \ldots$ the learner chooses an arm I_t among a set of K arms $\{1, \ldots, K\}$. At the same time the environment selects a loss vector $\ell_t \in [0,1]^K$ and the learner only observes and suffers the loss ℓ_{t,I_t}. The performance of the learner is evaluated using pseudo-regret, which is defined as

$$\overline{Reg}_T = \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,I_t}\right] - \min_{i\in[K]} \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=t}^{T} \left(\ell_{t,I_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* \in \operatorname{argmin}_{i\in[K]} \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,i}\right]$ is a best arm in hindsight in expectation over the loss generation model and, in case of an adaptive adversary, the randomness of the learner.

Like Zimmert and Seldin (2020) we consider (*adaptive*) *adversarial regimes* and *adversarial regimes with a* (Δ, C, T) *self-bounding constraint*. In the former the losses at round t are generated arbitrarily, potentially depending on the preceding actions of the learner, $I_1 \ldots, I_{t-1}$. In the latter the adversary selects losses, such that for some $\Delta \in [0,1]^K$ and $C \geq 0$ the pseudo-regret of any algorithm at time T satisfies

$$\overline{Reg}_T \geq \left(\sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{P}(I_t = i)\Delta_i\right) - C. \tag{2.1}$$

The condition is only assumed to be satisfied at time T, but not necessarily at $t < T$. As we have already mentioned in the introduction, *stochastic* regime, *stochastically constrained adversarial* regime, and *stochastic bandits with adversarial corruptions* are all special cases of the adversarial regime with (Δ, C, T) self-bounding constraint.

Additional Notation: We use Δ^n to denote the probability simplex over $n+1$ points. The characteristic function of a closed convex set $\mathcal{A}$ is denoted by $\mathcal{I}_{\mathcal{A}}(x)$ and satisfies $\mathcal{I}_{\mathcal{A}}(x) = 0$ for $x \in \mathcal{A}$ and $\mathcal{I}_{\mathcal{A}}(x) = \infty$ otherwise. We denote the indicator function of an event $\mathcal{E}$ by $\mathbb{1}(\mathcal{E})$ and use $\mathbb{1}_t(i)$ as a shorthand for $\mathbb{1}(I_t = i)$. The probability distribution over arms that is played by the learner at round t is denoted by $w_t \in \Delta^{K-1}$. The convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by $f^*(y) = \sup_{x\in\mathbb{R}^n}\{\langle x, y\rangle - f(x)\}$.

2.3 Background: the Tsallis-INF algorithm

In this section we provide a brief background on the Tsallis-INF algorithm of Zimmert and Seldin (2020). The algorithm is based on Follow The Regularized Leader (FTRL) framework with Tsallis entropy regularization (Tsallis, 1988). The best-of-both-worlds version of Tsallis-INF uses Tsallis entropy regularizer with power $\frac{1}{2}$, defined by

$$\Psi(w) = 4\sum_{i=1}^{K}\left(\sqrt{w_i} - \frac{1}{2}w_i\right).$$

The regularization term at round t is given by

$$\Psi_t(w) = \eta_t^{-1}\Psi(w),$$

where η_t is the learning rate. The update rule for the distribution over actions is defined by

$$w_{t+1} = \nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-\sum_{\tau=1}^{t}\hat{\ell}_\tau) = \arg\max_{w\in\Delta^{K-1}}\left(\left\langle -\sum_{\tau=1}^{t}\hat{\ell}_\tau, w\right\rangle - \Psi_t(w)\right),$$

where $\hat{\ell}_\tau$ is an estimate of the loss vector ℓ_τ. It is possible to use the standard importance-weighed loss estimate $\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}(\mathcal{I}_t=i)}{w_{t,i}}$, but Zimmert and Seldin (2020) have shown that reduced-variance loss estimates defined by

$$\hat{\ell}_{t,i} = \frac{\mathbb{1}_t(i)(\ell_{t,i} - \mathbb{B}_t(i))}{w_{t,i}} + \mathbb{B}_t(i), \tag{2.2}$$

where $\mathbb{B}_t(i) = \frac{1}{2}\mathbb{1}(w_{t,i} \geq \eta_t^2)$, lead to better constants. The complete algorithm is provided in Algorithm 1 box. The regret bound derived by Zimmert and Seldin (2020) is provided in Theorem 2.1.

Algorithm 1: Tsallis-INF

Input: $(\Psi_t)_{t=1,2,\ldots}$
Initialize: Set $\hat{L}_0 = \mathbf{0}_K$ (where $\mathbf{0}_K$ is a zero vector in $\mathbb{R}^K$)
for t = 1, ... **do**
  choose $w_t = \nabla(\Psi_t + \mathcal{I}_{\Delta^{K-1}})^*(-\hat{L}_{t-1})$
  sample $I_t \sim w_t$
  observe ℓ_{t,I_t}
  construct a loss estimator $\hat{\ell}_t$ using (2.2)
  update $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$
end for

Theorem 2.1 (Zimmert and Seldin, 2020)**.** *The pseudo-regret of Tsallis-INF with* $\eta_t = \frac{4}{\sqrt{t}}$ *and reduced variance loss estimators defined in equation* (2.2)*, in any adversarial bandit problem satisfies*

$$\overline{Reg}_T \leq 2\sqrt{KT} + 10K\log(T) + 16.$$

Furthermore, if there exists a vector $\Delta \in [0,1]^K$ *with a unique zero entry* i^* *(i.e.,* $\Delta_{i^*} = 0$ *and* $\Delta_i > 0$ *for all* $i \neq i^*$*) and a constant* C*, such that the pseudo-regret at time* T *satisfies the* (Δ, C, T) *self-bounding constraint (equation* (2.1)*), then the pseudo-regret additionally satisfies:*

$$\overline{Reg}_T \leq \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + 28K\log(T) + \frac{1}{\Delta_{min}} + \frac{3}{2}\sqrt{K} + 32 + C, \tag{2.3}$$

where $\Delta_{min} = \min_{i \neq i^*}\{\Delta_i\}$*. Moreover, if* $C \geq \left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i}\right) + \frac{1}{\Delta_{min}}$*, then the pseudo-regret also satisfies:*

$$\overline{Reg}_T \leq 2\sqrt{\left(\sum_{i \neq i^*} \frac{\log(T)+3}{\Delta_i} + \frac{1}{\Delta_{min}}\right) C} + 28K\log(T) + \frac{3}{2}\sqrt{K} + 32. \tag{2.4}$$

Remark 2.2. While Theorem 2.1 requires uniqueness of the best arm for improved regret rates in the adversarial regime with a (Δ, C, T) self-bounding constraint, Zimmert and Seldin (2020) have shown experimentally that in the stochastic regime the presence of multiple best arms has no negative effect on the pseudo-regret of the algorithm. They conjecture that the requirement is an artifact of the analysis.

2.4 Main Results

In this section we provide our two main results. First, in Theorem 2.3 we provide a refined analysis of Tsallis-INF, which improves the pseudo-regret bounds in the adversarial regime with a (Δ, C, T) self-bounding constraint. Then, in Theorem 2.4 we provide a more general result, which allows to improve pseudo-regret bounds in adversarial regimes with (Δ, C, T) self-bounding constraints for extensions of Tsallis-INF to other problems. An advantage of both results is that the bounds for adversarial regimes and adversarial regimes with a self-bounding constraint are achieved from a single optimization problem, rather than from two different optimization problems, as in prior work. As a result, the regret bounds for the adversarial regime are achieved as a limit case of the regret bounds for adversarial regimes with a self-bounding constraint for large C.

2.4.1 Improved analysis of the Tsallis-INF algorithm

We start with an improved regret bound for Tsallis-INF.

Theorem 2.3. *The pseudo-regret of Tsallis-INF with $\eta_t = \frac{4}{\sqrt{t}}$ and reduced variance loss estimators defined in equation (2.2), in any adversarial bandit problem satisfies*

$$\overline{Reg}_T \leq 2\sqrt{(K-1)T} + \frac{1}{2}\sqrt{T} + 14K\log(T) + \frac{3}{4}\sqrt{K} + 15. \tag{2.5}$$

Furthermore, if there exists a vector $\Delta \in [0,1]^K$ with a unique zero entry i^* (i.e., $\Delta_{i^*} = 0$ and $\Delta_i > 0$ for all $i \neq i^*$) and a constant $C \geq 0$, such that the pseudo-regret at time T satisfies the (Δ, C, T) self-bounding constraint (equation (2.1)), then the pseudo-regret additionally satisfies:

$$\overline{Reg}_T \leq \sum_{i \neq i*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\right)^2} \right) + 6 \right) + 28K\log(T) + \frac{3}{2}\sqrt{K} + 30 + C. \tag{2.6}$$

Moreover, for $\sum_{i \neq i*} \frac{1}{\Delta_i} \left(\left(\log \frac{T(K-1)}{(\sum_{i \neq i*} \frac{1}{\Delta_i})^2} \right) + 1 \right) \leq C \leq \frac{T(K-1)}{\sum_{i \neq i*} \frac{1}{\Delta_i}}$ the regret also satisfies:

$$\overline{Reg}_T \leq \sqrt{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} \left(\sqrt{\log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}}} + 5 \right) + Q, \tag{2.7}$$

where $Q = \sum_{i \neq i^*} \frac{1}{\Delta_i} \left(\log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} + \sqrt{2 \log \frac{T(K-1)}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}}} + 2 \right) + \frac{3\sqrt{K}}{2} + 28K\log(T) + 30$ is a subdominant term.

A proof of the theorem is provided in Appendix 2.7.2. Theorem 2.3 improves on Theorem 2.1 in two ways. The bound in equation (2.6) improves the leading term of the regret bound under self-bounding constraint relative to equation (2.3) from $\sum_{i \neq i^*} \frac{1}{\Delta_i} \log T$ to $\sum_{i \neq i*} \frac{1}{\Delta_i} \left(\log \frac{T(K-1)}{\left(\sum_{i \neq i^*} \frac{1}{\Delta_i}\right)^2} \right)$. Related refinements of regret bounds for UCB strategies for ordinary stochastic bandits have been studied by Auer and Ortner (2010) and Lattimore (2018). More importantly, for large amount of corruption $C \in \left[\sum_{i \neq i*} \frac{1}{\Delta_i} \left(\log \left(\frac{T(K-1)}{(\sum_{i \neq i*} \frac{1}{\Delta_i})^2} \right) + 1 \right), \frac{T(K-1)}{\sum_{i \neq i*} \frac{1}{\Delta_i}} \right]$ the regret bound in equation (2.7) is of order $\mathcal{O}\left(\sqrt{C \left(\sum_{i \neq i^*} \frac{1}{\Delta_i} \right) \log_+ \left(\frac{KT}{C \sum_{i \neq i^*} \frac{1}{\Delta_i}} \right)} \right)$, whereas the regret bound

in equation (2.4) is of order $\mathcal{O}\left(\sqrt{C\sum_{i\neq i^*}\frac{\log T}{\Delta_i}}\right)$. For $C=\Theta\left(\frac{TK}{(\log T)\sum_{i\neq i^*}\frac{1}{\Delta_i}}\right)$ Theorem 2.3 improves the pseudo-regret bound by a multiplicative factor of $\sqrt{\frac{\log T}{\log\log T}}$. Another observation is that Theorem 2.3 successfully exploits the self-bounding property even when the amount of corruption is almost linear in T.

2.4.2 A general analysis based on the self-bounding property

Now we provide a general result, which can be used to analyze extensions of Tsallis-INF to other problem settings.

Theorem 2.4. *For any algorithm for an arbitrary problem domain with K possible actions that satisfies*

$$\overline{Reg}_T \leq B\sum_{t=1}^{T}\sum_{i\neq i^*}\sqrt{\frac{\mathbb{E}[w_{t,i}]}{t}}+D, \tag{2.8}$$

where $B, D \geq 0$ are some constants, the pseudo-regret of the algorithm in any adversarial environment satisfies

$$\overline{Reg}_T \leq 2B\sqrt{(K-1)T}+D. \tag{2.9}$$

Furthermore, if there exists a vector $\Delta\in[0,1]^K$ with a unique zero entry i^* (i.e., $\Delta_{i^*}=0$ and $\Delta_i>0$ for all $i\neq i^*$) and a constant $C\geq 0$, such that the pseudo-regret at time T satisfies the (Δ,C,T) self-bounding constraint (equation (2.1)), then the pseudo-regret additionally satisfies:

$$\overline{Reg}_T \leq B^2\sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)+3-2\log B\right)+C+2D. \tag{2.10}$$

Moreover, for $B^2\sum_{i\neq i*}\frac{1}{\Delta_i}\left(\left(\log\frac{T(K-1)}{B^2(\sum_{i\neq i*}\frac{1}{\Delta_i})^2}\right)+1\right)\leq C\leq\frac{T(K-1)}{\sum_{i\neq i*}\frac{1}{\Delta_i}}$ the pseudo-regret also satisfies:

$$\overline{Reg}_T \leq B\sqrt{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}\left(\sqrt{\log\frac{T(K-1)}{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}}+2\right)+M, \tag{2.11}$$

where $M=B^2\sum_{i\neq i^*}\frac{1}{\Delta_i}\left(\log\frac{T(K-1)}{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}+\sqrt{2\log\frac{T(K-1)}{C\sum_{i\neq i^*}\frac{1}{\Delta_i}}}+2\right)+2D$ is a subdominant term.

A proof is provided in Section 2.5. The Tsallis-INF algorithm satisfies the condition in equation (2.8) with $B = \frac{5}{4}$ (see equation (2.12) in Section 2.5, which follows from intermediate results by Zimmert and Seldin (2020)). Although the specialized analysis of Tsallis-INF in Theorem 2.3 is a bit tighter than the general result in Theorem 2.4, the latter can be applied to extensions of Tsallis-INF. One such example is the best-of-both-worlds algorithm of Jin and Luo (2020) for episodic MDPs. Jin and Luo (2020, Theorem 4) show that their algorithm satisfies the condition in (2.8) and use this result to achieve $\mathcal{O}\left((\log T) + \sqrt{C \log(T)}\right)$ pseudo-regret bound in the stochastic case with adversarial corruptions (Jin and Luo, 2020, Corollary 3). Application of our Theorem 2.4 improves the pseudo-regret bound to $\mathcal{O}\left((\log T) + \sqrt{C \log(T/C)}\right)$. In particular, for $C = \Theta(\frac{T}{\log T})$ the bound gets tighter by a multiplicative factor of $\frac{\log T}{\log \log T}$.

2.5 Proofs

In this section we provide a proof of Theorem 2.4. The proof of Theorem 2.3 is analogous, but more technical due to fine-tuning of the constants and is deferred to Appendix 2.7.2. Before showing the proof we revisit the key steps in the analysis of Tsallis-INF by Zimmert and Seldin (2020), which show that the pseudo-regret of Tsallis-INF satisfies the condition in equation (2.8) of Theorem 2.4.

Standard FTRL analysis (Lattimore and Szepesvári, 2020) uses a potential function $\Phi_t(-L) = \max_{w \in \Delta^{K-1}} \{\langle w, -L\rangle - \Psi_t(w)\}$ for breaking the pseudo-regret into *penalty* and *stability* terms, $\overline{Reg_T} = \mathit{stability} + \mathit{penalty}$, where

$$\mathit{stability} = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} + \Phi_t(-\hat{L}_t) - \Phi_t(-\hat{L}_{t-1})\right],$$
$$\mathit{penalty} = \mathbb{E}\left[\sum_{t=1}^{T} -\Phi_t(-\hat{L}_t) + \Phi_t(-\hat{L}_{t-1}) - \ell_{t,i_T^*}\right].$$

The two terms are then typically analyzed separately. Zimmert and Seldin (2020) proved the following bounds for the two terms for Tsallis-INF with reduced-variance

loss estimators:

$$\begin{aligned} \textit{stability} &\leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{2\sqrt{t}} + \frac{\mathbb{E}[w_{t,i}]}{2\sqrt{t}} \right) + 14K \log(T) + 15, \\ \textit{penalty} &\leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{2\sqrt{t}} - \frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} \right) + \frac{3}{4}\sqrt{K}. \end{aligned}$$

By summation of the two bounds the pseudo-regret satisfies

$$\overline{Reg}_T \leq \sum_{t=1}^{T} \left(\sum_{i \neq i^*} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} + \frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} \right) + 14K \log(T) + \frac{3}{4}\sqrt{K} + 15. \tag{2.12}$$

Since $\mathbb{E}[w_{t,i}] \leq \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$, the pseudo-regret of Tsallis-INF with reduced-variance loss estimators satisfies the condition in equation (2.8) with $B = \frac{5}{4}$ and $D = \frac{3}{4}\sqrt{K} + 14K \log(T) + 15$. (In the proof of Theorem 2.3 we keep the refined bound on the pseudo-regret from equation (2.12) to obtain better constants.) Now, after we have shown how the condition in equation (2.8) can be satisfied, we present a proof of Theorem 2.4. We start with a high-level overview of the key ideas and then present the technical details.

2.5.1 Overview of the Key Ideas Behind the Proof of Theorem 2.4

As observed by Zimmert and Seldin (2020), for any $\lambda \in [0, 1]$ we have

$$\overline{Reg}_T = (\lambda + 1)\overline{Reg}_T - \lambda \overline{Reg}_T. \tag{2.13}$$

The condition on $\overline{Reg}_T$ in equation (2.8) can be used to upper bound the first term and the self-bounding constraint (2.1) to lower bound the second, giving

$$\begin{aligned} \overline{Reg}_T &\leq (\lambda + 1) \left(B \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} + D \right) - \lambda \left(\sum_{t=1}^{T} \left(\sum_{i \neq i^*} \mathbb{E}[w_{t,i}] \Delta_i \right) - C \right) \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(B(\lambda + 1) \frac{\mathbb{E}[w_{t,i}]^{\frac{1}{2}}}{\sqrt{t}} - \lambda \mathbb{E}[w_{t,i}] \Delta_i \right) + \lambda C + (\lambda + 1) D. \end{aligned} \tag{2.14}$$

In the adversarial analysis, we take $\lambda = 0$ and maximize the right hand side of (2.14)

(which for $\lambda = 0$ is identical to the right hand side of (2.8)) under the constraint that $w_{t,i}$ is a probability distribution to obtain $\mathcal{O}(\sqrt{KT})$ regret bound. This is almost identical to the approach of Zimmert and Seldin (2020), except that in this case instead of the bound in equation (2.8) they use a bound involving summation over all arms, including i^*.

In the self-bounding analysis, Zimmert and Seldin (2020) relax the inequality in (2.14) to

$$\overline{Reg}_T \leq \sum_{t=1}^{T} \sum_{i \neq i^*} \left(2B\sqrt{\mathbb{E}\left[w_{t,i}\right]/t} - \lambda \Delta_i \mathbb{E}\left[w_{t,i}\right] \right) + \lambda C + 2D$$

and apply *individual* maximization of each $2B\sqrt{\mathbb{E}\left[w_{t,i}\right]/t} - \lambda\Delta_i\mathbb{E}\left[w_{t,i}\right]$ term, dropping the constraint that w_t is a probability distribution. We use (2.14) directly for bounding the regret and introduce two key novelties:

(a) we keep the constraint that w_t are probability distributions; and

(b) we jointly optimize with respect to all $w_{t,i}$ and λ, whereas Zimmert and Seldin (2020) first optimize w.r.t. $w_{t,i}$ and then w.r.t. λ.

Joint optimization over all $w_{t,i}$ and λ under the constraint that w_t are probability distributions is the major technical challenge that we resolve. Our analysis yields three advantages:

(A) The dependence on time is improved from $\log T$ to $\log(T(K-1)/(\sum_{i\neq i^*}\frac{1}{\Delta_i})^2)$ due to (a);

(B) We gain the $\sqrt{\log T/\log(T/C)}$ factor due to (b);

(C) Our adversarial and stochastic bounds come out of the same optimization problem, highlighting the relation and continuity between the two.

2.5.2 Proof of Theorem 2.4

Now we provide a detailed proof of Theorem 2.4.

Proof of the regret bound for an unconstrained adversarial regime (equation (2.9)**)**

In the unconstrained adversarial regime we take $\lambda = 0$ and plug the inequalities

$$\sum_{i \neq i^*} \mathbb{E}[w_{t,i}]^{\frac{1}{2}} \leq \sqrt{K-1}, \tag{2.15}$$

which holds since $\sum_{i \neq i^*} \mathbb{E}[w_{t,i}] \leq 1$, and $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$ into equation (2.14) and obtain the bound in equation (2.9).

Proof of the general regret bound for an adversarial regime with a self-bounding constraint (equation (2.10))

In the adversarial regime with a self-bounding constraint, we keep the constraint that w_t is a probability distribution, and thus $\sum_{i \neq i^*} \mathbb{E}[w_{t,i}]^{\frac{1}{2}} \leq \sqrt{K-1}$, and apply maximization directly to the sum over i under this constraint.

To simplify the notation, we use $a_{t,i} := \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$, $S := \sum_{i \neq i^*} \frac{1}{\Delta_i}$, and w.l.o.g. assume that $i^* = K$. We denote $R_t := \sum_{i \neq i^*} \left(B(\lambda+1)\frac{a_{t,i}}{\sqrt{t}} - \lambda \Delta_i a_{t,i}^2 \right)$ and $R := \sum_{t=1}^{T} R_t + \lambda C$. With this notation, by equation (2.14) we have

$$\overline{Reg}_T \leq R + (1+\lambda)D. \tag{2.16}$$

We bound R_t under the constraint that $\mathbb{E}[w_{t,i}]^{\frac{1}{2}}$ satisfy equation (2.15). We have

$$\begin{aligned} R_t \leq \max_{a_1,\ldots,a_{K-1}} & \sum_{i=1}^{K-1} B(\lambda+1)\frac{a_i}{\sqrt{t}} - \lambda \Delta_i a_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^{K-1} a_i \leq \sqrt{K-1}. \end{aligned}$$

By Lemma 2.2 provided in Appendix 2.7.1, the answer to this optimization problem is as follows:

1. If $\frac{B(\lambda+1)S}{2\lambda\sqrt{t}} \leq \sqrt{K-1}$, then $R_t \leq \frac{SB^2(\lambda+1)^2}{4\lambda t}$.
2. If $\frac{B(\lambda+1)S}{2\lambda\sqrt{t}} \geq \sqrt{K-1}$, then $R_t \leq \frac{\sqrt{K-1}B(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S}$.

This gives a threshold $T_0 = \frac{B^2(\lambda+1)^2S^2}{4\lambda^2(K-1)}$, so that for $t \leq T_0$ the second case applies to R_t, and otherwise the first case applies. We break the time steps into those before

T_0 and after T_0 and obtain:

$$\begin{aligned}
R &= \sum_{t=1}^{T_0} R_t + \sum_{T_0+1}^{T} R_t + \lambda C \\
&\leq \sum_{t=1}^{T_0} \left(\frac{\sqrt{K-1}B(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_0+1}^{T} \frac{SB^2(\lambda+1)^2}{4\lambda t} + \lambda C \\
&\leq 2\sqrt{T_0(K-1)}B(\lambda+1) - \frac{\lambda(K-1)T_0}{S} + \frac{SB^2(\lambda+1)^2}{4\lambda} \log \frac{T}{T_0} + \lambda C \\
&= \frac{B^2(\lambda+1)^2 S}{\lambda} - \frac{B^2(\lambda+1)^2 S}{4\lambda} + \frac{B^2(\lambda+1)^2 S}{4\lambda} \left(\log \frac{T(K-1)}{S^2} - 2\log \frac{B(\lambda+1)}{2\lambda} \right) + \lambda C \\
&= \frac{B^2(\lambda+1)^2 S}{4\lambda} \left[3 + \log \frac{T(K-1)}{S^2} \right] - \frac{B^2(\lambda+1)^2 S}{2\lambda} \log \frac{B(\lambda+1)}{2\lambda} + \lambda C. \qquad (2.17)
\end{aligned}$$

By taking $\lambda = 1$ we obtain

$$R \leq B^2 S \left(\log \frac{T(K-1)}{S^2} - 2\log(B) + 3 \right) + C,$$

which together with (2.16) gives the bound (2.10) in the theorem.

Proof of the refined regret bound for an adversarial regime with a self-bounding constraint (equation (2.11))

We continue from equation (2.17). We improve on the bound of Zimmert and Seldin (2020) in equation (2.4) by applying a smarter optimization over λ. We let $\alpha = \frac{2\lambda}{B(\lambda+1)}$ and rewrite the inequality in (2.17) as

$$R \leq \underbrace{\frac{B}{2 - B\alpha} \underbrace{\left[\frac{S}{\alpha} \left(3 + \log \left(\frac{T(K-1)}{S^2} \right) \right) + \frac{2S}{\alpha} \log(\alpha) + \alpha C \right]}_{f(\alpha)}}_{h(B,\alpha)}. \qquad (2.18)$$

We denote the right hand side of the expression by $h(B, \alpha)$. We restrict the range of α, so that $T \geq T_0 = \frac{S^2}{\alpha^2(K-1)}$, which gives $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$. Since $\lambda \in [0, 1]$, we also have $\alpha \leq \frac{1}{B}$. In order to bound $h(B, \alpha)$ we need to solve an optimization problem in α over the above interval. However, $h(B, \alpha)$ is not convex in α, but we show that the expression in the brackets, which we denote by $f(\alpha)$, is convex. We take the

point $\alpha^* = \text{argmin}_{\alpha \in [\frac{S}{\sqrt{T(K-1)}}, \frac{1}{B}]} f(\alpha)$, which achieves the minimum of $f(\alpha)$, and use $h(B, \alpha^*) = \frac{B}{2 - B\alpha^*} f(\alpha^*)$ as an upper bound for R. Since $R \leq h(B, \alpha)$ for any α, in particular we have $R \leq h(B, \alpha^*)$.

In order to show that $f(\alpha)$ is convex and find its minimum we take the first and second derivatives.

$$f'(\alpha) = \frac{-1}{\alpha^2}\left[2S\log(\alpha) - C\alpha^2 + S\log\frac{(K-1)T}{S^2} + S\right] = 0,$$
$$f''(\alpha) = \frac{2S}{\alpha^3}\left(2\log\alpha + \log\frac{T(K-1)}{S^2}\right).$$

For $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$ the second derivative is positive and, therefore, $f(\alpha)$ is convex and the minimum is achieved when $f'(\alpha) = 0$. This happens when

$$-\log\frac{\alpha^2(K-1)T}{S^2} + \frac{C}{S}\alpha^2 - 1 = 0.$$

We define $\beta = \frac{\alpha^2(K-1)T}{S^2}$, then

$$g(\beta) = \frac{CS}{(K-1)T}\beta - \log(\beta) - 1 = 0.$$

Since $\alpha \in [\frac{S}{\sqrt{T(K-1)}}, \frac{1}{B}]$, we have $\beta \in [1, \frac{(K-1)T}{B^2S^2}]$. We recall that equation (2.11) holds under the assumption that $B^2S\left(\log\frac{(K-1)T}{B^2S^2} + 1\right) \leq C \leq \frac{(K-1)T}{S}$. We note that for $C \leq \frac{(K-1)T}{S}$ we have $g(1) = \frac{CS}{(K-1)T} - 1 \leq 0$. We also note that for $C \geq B^2S\left(\log\frac{(K-1)T}{B^2S^2} + 1\right)$ we have $g\left(\frac{(K-1)T}{B^2S^2}\right) \geq 0$. Since $g(\beta)$ is continuous, the root of $g(\beta) = 0$ for C in the above range is thus achieved by $\beta \in [1, \frac{(K-1)T}{B^2S^2}]$ and since $g(\beta)$ is convex the solution is unique.

We find the root of $g(\beta) = 0$ by using the -1-branch of the *Lambert W function*, called $W_{-1}(x)$, which is defined as the solution of equation $we^w = x$. If $g(\beta) = 0$, then β satisfies

$$\frac{-CS\beta}{(K-1)T}e^{\frac{-CS\beta}{(K-1)T}} = \frac{-CS}{e(K-1)T},$$

and thus

$$\beta = \frac{-T(K-1)}{CS}W_{-1}\left(\frac{-CS}{e(K-1)T}\right).$$

We conclude that the minimum of $f(\alpha)$ is attained at

$$\alpha^* = \sqrt{\frac{-S}{C} W_{-1}\left(\frac{-CS}{e(K-1)T}\right)} \tag{2.19}$$

and, consequently, $\log\left(\frac{T(K-1)(\alpha^*)^2}{S^2}\right) = \frac{C}{S}(\alpha^*)^2 - 1$. By substituting this identity into $h(B, \alpha^*)$, we obtain:

$$\begin{aligned} h(B,\alpha^*) &= \frac{B}{2 - B\alpha^*}\left(2\frac{S}{\alpha^*} + 2C\alpha^*\right) \leq B(1 + B\alpha^*)\left(\frac{S}{\alpha^*} + C\alpha^*\right) \\ &= B\left(\frac{S}{\alpha^*} + C\alpha^* + BS + BC(\alpha^*)^2\right) = B\left(\sqrt{\frac{CS}{w}} + \sqrt{CSw} + BS + BSw\right), \end{aligned} \tag{2.20}$$

where $w := -W_{-1}\left[\frac{-CS}{e(K-1)T}\right]$ and the inequality follows by the fact that $\forall x \in [0,1]$: $\frac{2}{2-x} \leq 1 + x$. This provides a closed form upper bound for the pseudo-regret, but we still need an estimate of w to obtain an explicit bound. We use the result of Chatzigeorgiou (2013), who provides the following bounds for $W_{-1}(x)$.

Lemma 2.1 (Chatzigeorgiou 2013)**.** *For any $x \leq 1$*

$$1 + \sqrt{2\log(1/x)} + \frac{2}{3}\log(1/x) \leq -W_{-1}(-x/e) \leq 1 + \sqrt{2\log(1/x)} + \log(1/x).$$

To complete the proof it suffices to use Lemma 2.1 with $x = \frac{CS}{(K-1)T}$, which gives

$$1 \leq w \leq 1 + \sqrt{2\log\frac{T(K-1)}{CS}} + \log\frac{T(K-1)}{CS} \leq \left(1 + \sqrt{\log\frac{T(K-1)}{CS}}\right)^2.$$

By substituting this into (2.20) we obtain:

$$\begin{aligned} h(B,\alpha^*) &\leq B\sqrt{CS} + B\sqrt{CS}\left(1 + \sqrt{\log\frac{T(K-1)}{CS}}\right) \\ &\quad + 2B^2S + B^2S\log\frac{T(K-1)}{CS} + B^2S\sqrt{2\log\frac{T(K-1)}{CS}} \\ &= B\sqrt{CS}\left(\sqrt{\log\frac{T(K-1)}{CS}} + 2\right) + B^2S\left(\log\frac{T(K-1)}{CS} + \sqrt{2\log\frac{T(K-1)}{CS}} + 2\right). \end{aligned} \tag{2.21}$$

Finally, by (2.18) we have $R \leq h(B, \alpha^*)$, which together with (2.16) and the fact that $\lambda \leq 1$ completes the proof. ■

2.6 Discussion

We have presented a refined analysis of the Tsallis-INF algorithm in adversarial regimes with a self-bounding constraint. The result improves on prior work in two ways. First, it improves the dependence of the regret bound on time horizon from $\log T$ to $\log \frac{(K-1)T}{(\sum_{i \neq i^*} \frac{1}{\Delta_i})^2}$. Second, it improves the dependence of the regret bound on corruption amount C. In particular, for $C = \Theta\left(\frac{TK}{(\log T) \sum_{i \neq i^*} \sum \frac{1}{\Delta_i}}\right)$ it improves the pseudo-regret bound by a multiplicative factor of $\sqrt{\frac{\log T}{\log \log T}}$. Moreover, we have provided a generalized result that can be used to improve regret bounds for extensions of Tsallis-INF to other problem settings, where the regret satisfies a self-bounding constraint. Due to versatility and rapidly growing popularity of regret analysis based on the self-bounding property, the result provides a powerful tool for tightening regret bounds in a broad range of corrupted settings.

2.7 Appendix

2.7.1 Technical Lemmas

Lemma 2.2. *Let b and $c_1, \ldots, c_n$ be non-negative real numbers and let*

$$Z = \max_{x \in \mathbb{R}^n} \sum_{i=1}^{n} (b x_i - c_i x_i^2)$$
$$s.t. \sum_{i=1}^{n} x_i \leq M.$$

Then

$$Z = \begin{cases} bM - \frac{M^2}{\sum_{i=1}^n \frac{1}{c_i}}, & \text{if } \sum_{i=1}^n \frac{b}{2c_i} > M, \\ \frac{b^2}{4} \sum_{i=1}^n \frac{1}{c_i}, & \text{otherwise.} \end{cases}$$

Moreover, we always have $bM - \frac{M^2}{\sum_{i=1}^n \frac{1}{c_i}} \leq \frac{b^2}{4} \sum_{i=1}^n \frac{1}{c_i}$ and, therefore, we always have $Z \leq \frac{b^2}{4} \sum_{i=1}^n \frac{1}{c_i}$.

Proof. Since $c_i \geq 0$, the objective function is a sum of downward-pointing parabolas and, therefore, concave. Thus, the maximum is attained when the first derivative of the Lagrangian with Lagrange variable $v \geq 0$ for the inequality constraint satisfies

$$b - 2c_i x_i - v = 0,$$

where $v(\sum_{i=1}^n x_i - M) = 0$. Thus, $x_i = \frac{b-v}{2c_i}$. The KKT conditions provide two cases:

i) If $\sum_{i=1}^n \frac{b}{2c_i} > M$, then $v > 0$ and $\sum_{i=1}^n x_i = M$. As a consequence, $v = b - \frac{M}{\sum_{i=1}^n \frac{1}{2c_i}}$. So $x_i = \frac{M}{c_i \sum_{i=1}^n \frac{1}{c_i}}$ and $Z = bM - \frac{M^2}{\sum_{i=1}^n \frac{1}{c_i}}$.

ii) If $\sum_{i=1}^n \frac{b}{2c_i} \leq M$, then $v = 0$ and, as a consequence, $x_i = \frac{b}{2c_i}$ and $Z = \frac{b^2}{4} \sum_{i=1}^n \frac{1}{c_i}$.

Finally, by the AM-GM inequality we have

$$\frac{M^2}{\sum_{i=1}^n \frac{1}{c_i}} + \frac{b^2}{4} \sum_{i=1}^n \frac{1}{c_i} \geq bM,$$

which gives the final statement of the lemma. □

We also use the following result by Zimmert and Seldin (2020, Lemma 15).

Lemma 2.3 (Zimmert and Seldin, 2020)**.** *For any $b > 0$ and $c > 0$ and $T_0, T \in \mathbb{N}$, such that $T_0 < T$ and $b\sqrt{T_0} > c$, it holds that*

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{b\sqrt{T_0} - c}\,.$$

By doubling the lower threshold on $b\sqrt{T_0}$ we obtain the following corollary.

Corollary 2.1. *For any $b > 0$ and $c > 0$ and $T_0, T \in \mathbb{N}$, such that $T_0 < T$ and $b\sqrt{T_0} \geq 2c$, it holds that*

$$\sum_{t=T_0+1}^{T} \frac{1}{bt^{\frac{3}{2}} - ct} \leq \frac{2}{c}.$$

2.7.2 Proof of Theorem 2.3

Proof. Similar to the proof of Theorem 2.4, for any $\lambda \in [0,1]$ we use the self-bounding constraint and the regret bound of Zimmert and Seldin (2020) given in equation (2.12) to provide the following bound for the pseudo-regret:

$$\begin{aligned}\overline{Reg}_T &= (\lambda+1)\overline{Reg}_T - \lambda\overline{Reg}_T \\ &\leq (\lambda+1)\left(\sum_{i\neq i^*}\left[\sum_{t=1}^{T}\frac{\mathbb{E}[w_{t,i}]}{4\sqrt{t}} + \sum_{t=1}^{T}\frac{\sqrt{\mathbb{E}[w_{t,i}]}}{\sqrt{t}}\right] + \frac{3}{4}\sqrt{K} + 14K\log(T) + 15\right) \\ &\quad - \lambda\left(\left[\sum_{t=1}^{T}\sum_{i\neq i^*}\mathbb{E}[w_{t,i}]\Delta_i\right] - C\right).\end{aligned}$$

As before, to simplify the notation, let $a_{t,i} = \mathbb{E}[w_{t,i}]^{\frac{1}{2}}$ and $S = \sum_{i\neq i^*}\frac{1}{\Delta_i}$ and w.l.o.g. assume that $i^* = K$ and define

$$R_t = \sum_{i\neq i^*}\left(\frac{\lambda+1}{\sqrt{t}}a_{t,i} - \left(\lambda\Delta_i - \frac{\lambda+1}{4\sqrt{t}}\right)a_{t,i}^2\right), \tag{2.22}$$

$$R = \sum_{t=1}^{T} R_t + \lambda C.$$

Then

$$\overline{Reg}_T \leq R + (1+\lambda)\left(\frac{3}{4}\sqrt{K} + 14K\log(T) + 15\right). \tag{2.23}$$

Hence, in order to obtain a bound for the pseudo-regret, it suffices to derive a bound for R. We start with the bound for a general adversarial environment and then prove the refinements.

Proof of the regret bound for an unconstrained adversarial regime (equation (2.5))

We take $\lambda = 0$. By plugging it into the definition of R_t in equation (2.22) we obtain

$$R_t \leq \frac{\sqrt{K-1}}{\sqrt{t}} + \frac{1}{4\sqrt{t}}$$

and

$$R = \sum_{t=1}^{T} R_t \leq 2\sqrt{(K-1)T} + \frac{1}{2}\sqrt{T}.$$

Plugging this into (2.23) completes the proof of (2.5).

Proof of the regret bounds for an adversarial regime with a self-bounding constraint (equations (2.6) and (2.7))

Now we prove the refined bounds for adversarial environments satisfying the self-bounding constraint with unique best arm. Similarly to the proof of Theorem 2.4, we bound R_t for each $t \geq 1$ by solving a constrained maximization problem over $\{a_{t,i}\}_{i=1}^n$, where the constraint is $\sum_{i=1}^{K-1} a_{t,i} \leq \sqrt{K-1}$. But the challenge here is that the coefficients $\lambda\Delta_i - \frac{\lambda+1}{4\sqrt{t}}$ in front of $a_{t,i}^2$ in the definition of R_t are not necessarily positive, and if they are not, then Lemma 2.2 cannot be applied. More precisely, if

$$\forall i \neq i^* : \lambda\Delta_i \geq \frac{\lambda+1}{4\sqrt{t}} \Rightarrow t \geq \left(\frac{\lambda+1}{4\lambda\Delta_{min}}\right)^2, \tag{2.24}$$

where $\Delta_{min} = \min_{i\neq i^*}\{\Delta_i\}$, then all the coefficients are positive. We denote $\alpha = \frac{2\lambda}{\lambda+1}$ and define a threshold $T_1 = \left(\frac{\lambda+1}{2\lambda\Delta_{min}}\right)^2 = \left(\frac{1}{\alpha\Delta_{min}}\right)^2$. We note that T_1 is four times larger than what is required for satisfaction of the condition in equation (2.24). The reason is that at a later point in the proof we apply Corollary 2.1 for $t \geq T_1$ and we need to satisfy the condition of the corollary. For $t \geq T_1$ we can use Lemma 2.2 to bound R_t. By the lemma we obtain:

$$R_t \leq \frac{(\lambda+1)^2}{4t}\sum_{i=1}^{K-1}\frac{1}{\lambda\Delta_i - \frac{\lambda+1}{\sqrt{t}}} = \sum_{i=1}^{K-1}\frac{\lambda+1}{\frac{4\lambda}{\lambda+1}\Delta_i t - \sqrt{t}} = \sum_{i=1}^{K-1}\frac{\lambda+1}{2\alpha\Delta_i t - \sqrt{t}}.$$

We rewrite each term in the summation in the following way

$$\frac{\lambda+1}{2\alpha\Delta_i t - \sqrt{t}} = \frac{\lambda+1}{2\alpha\Delta_i t} + \frac{\lambda+1}{4\alpha^2\Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}$$

and obtain

$$\text{for } t \geq T_1\text{:} \qquad R_t \leq \frac{S(\lambda+1)}{2\alpha t} + \sum_{i=1}^{K-1}\frac{\lambda+1}{4\alpha^2\Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t}. \tag{2.25}$$

In order to bound R_t for $t < T_1$, we break it into two parts as follows:

$$\begin{aligned} R_t =& \sum_{i\neq i^*}\left(\frac{\lambda+1}{\sqrt{t}}a_{t,i} - \lambda\Delta_i a_{t,i}^2\right) + \sum_{i\neq i^*}\left(\frac{\lambda+1}{4\sqrt{t}}a_{t,i}^2\right) \\ \leq& \sum_{i\neq i^*}\left(\frac{\lambda+1}{\sqrt{t}}a_{t,i} - \lambda\Delta_i a_{t,i}^2\right) + \frac{1}{2\sqrt{t}}, \end{aligned}$$

where the inequality holds because $\lambda \leq 1$ and $\sum_{i \neq i^*} a_{t,i}^2 \leq 1$. We use Lemma 2.2 to bound the summation in the latter expression. The solution depends on a threshold $T_2 = \frac{(\lambda+1)^2 S^2}{4\lambda^2(K-1)} = \frac{S^2}{(K-1)\alpha^2}$:

$$\text{for } t \leq T_2: \qquad R_t \leq \frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} + \frac{1}{2\sqrt{t}}, \tag{2.26}$$

$$\text{for } t \geq T_2: \qquad R_t \leq \frac{S(\lambda+1)^2}{4\lambda t} + \frac{1}{2\sqrt{t}} = \frac{S(\lambda+1)}{2\alpha t} + \frac{1}{2\sqrt{t}}. \tag{2.27}$$

Note that for $t \geq T_1$ we have a choice between using the bound in equation (2.25) or one of the bounds in (2.26) or (2.27), depending on whether $t \leq T_2$ or $t \geq T_2$. The relation between the thresholds, $T_1 \leq T_2$ or $T_2 \leq T_1$, depends on the relation between $\left(\frac{1}{\Delta_{min}}\right)^2$ and $\frac{S^2}{K-1}$. Also note that the choice of α (which determines λ) affects the thresholds T_1 and T_2, but not their relation. Similar to the proof of Theorem 2.4, we restrict the range of α, so that $T \geq T_2 = \frac{S^2}{\alpha^2(K-1)}$, which gives $\alpha \geq \frac{S}{\sqrt{T(K-1)}}$.

We now derive a bound on R. We consider three cases: $T_2 \leq T \leq T_1$, $T_2 \leq T_1 \leq T$, and $T_1 \leq T_2 \leq T$.

First case: $T_2 \leq T \leq T_1$**.** By (2.26) and (2.27) we have:

$$\begin{aligned}
&\sum_{t=1}^{T} R_t \leq \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T} R_t \\
&\leq \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) + \sqrt{T} \\
&\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \sqrt{T_1},
\end{aligned} \tag{2.28}$$

where in the second line we used $\sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \leq \sqrt{T}$ and in the third line $\sum_{t=T_2+1}^{T} \frac{1}{t} \leq \log(T/T_2)$ and $\lambda \leq 1$ and $T \leq T_1$.

Second case: $T_2 \le T_1 \le T$**.** By (2.26), (2.27), and (2.25) we have:

$$\begin{aligned}
\sum_{t=1}^{T} R_t &\le \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T_1} R_t + \sum_{t=T_1+1}^{T} R_t \\
&\le \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) + \sqrt{T_1} \\
&\qquad + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{\lambda+1}{4\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t} \\
&\le 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) + \sqrt{T_1} \\
&\qquad + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - \alpha\Delta_i t}, \qquad (2.29)
\end{aligned}$$

where in the second line we used $\sum_{t=1}^{T_1} \frac{1}{2\sqrt{t}} \le \sqrt{T_1}$ and in the third line $\sum_{t=T_2+1}^{T} \frac{1}{t} \le \log(T/T_2)$ and $\lambda \le 1$.

Third case: $T_1 \le T_2 \le T$**.** By (2.26) and (2.25) we have:

$$\begin{aligned}
\sum_{t=1}^{T} R_t &\le \sum_{t=1}^{T_2} R_t + \sum_{t=T_2+1}^{T} R_t \\
&\le \sum_{t=1}^{T_2} \left(\frac{\sqrt{K-1}(\lambda+1)}{\sqrt{t}} - \frac{\lambda(K-1)}{S} \right) + \sqrt{T_2} + \sum_{t=T_2+1}^{T} \left(\frac{S(\lambda+1)}{2\alpha t} \right) \\
&\qquad + \sum_{i=1}^{K-1} \sum_{t=T_2+1}^{T} \frac{\lambda+1}{4\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - 2\alpha\Delta_i t} \\
&\le 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \sqrt{T_2} + \frac{S(\lambda+1)}{2\alpha} \log(\frac{T}{T_2}) \\
&\qquad + \sum_{i=1}^{K-1} \sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2 \Delta_i^2 t^{\frac{3}{2}} - \alpha\Delta_i t}. \qquad (2.30)
\end{aligned}$$

Merging the cases: Corollary 2.1 provides an upper bound for the last terms of (2.29) and (2.30):

$$\sum_{t=T_1+1}^{T} \frac{1}{2\alpha^2\Delta_i^2 t^{\frac{3}{2}} - \alpha\Delta_i t} \leq \frac{2}{\alpha\Delta_i}.$$

Now we combine (2.28), (2.29), and (2.30), and obtain following bound for R:

$$\begin{aligned} R &= \sum_{t=1}^{T} R_t + \lambda C \\ &\leq 2\sqrt{T_2(K-1)}(\lambda+1) - \frac{\lambda(K-1)T_2}{S} + \frac{S(\lambda+1)}{2\alpha}\log(\frac{T}{T_2}) + \lambda C \\ &\quad + \sqrt{\max\{T_1,T_2\}} + \sum_{i=1}^{K-1}\frac{2}{\alpha\Delta_i}. \end{aligned} \tag{2.31}$$

We note that $\max\{T_1,T_2\} = \max\left\{\frac{S^2}{(K-1)\alpha^2}, \frac{1}{\Delta_{min}^2\alpha^2}\right\} \leq \frac{S^2}{\alpha^2}$. Moreover, by substituting $T_2 = \frac{S^2}{\alpha^2(K-1)}$ into (2.31) we obtain:

$$\begin{aligned} R &\leq 2(\lambda+1)\frac{S}{\alpha} - \frac{\lambda S}{\alpha^2} + \frac{S(\lambda+1)}{2\alpha}\log\left(\frac{\alpha^2(K-1)T}{S^2}\right) + \lambda C + \frac{S}{\alpha} + \sum_{i=1}^{K-1}\frac{2}{\alpha\Delta_i} \\ &= \frac{\lambda+1}{2}\left[4\frac{S}{\alpha} - \frac{S}{\alpha} + \frac{S}{\alpha}\log\left(\frac{(K-1)T}{S^2}\right) + \frac{2S}{\alpha}\log(\alpha) + \alpha C\right] + \frac{3S}{\alpha} \\ &= \underbrace{\frac{1}{2-\alpha}\left[\frac{S}{\alpha}\left(3 + \log\left(\frac{T(K-1)}{S^2}\right)\right) + \frac{2S}{\alpha}\log(\alpha) + \alpha C\right]}_{h(1,\alpha)} + \frac{3S}{\alpha}. \end{aligned} \tag{2.32}$$

We recognize that the first term in equation (2.32) is $h(1,\alpha)$, which was defined earlier in equation (2.18).

Proof of the general bound in equation (2.6)**:** By taking $\lambda = 1$, which corresponds to $\alpha = 1$, we obtain

$$\begin{aligned} R &\leq S\left(\log\left(\frac{T(K-1)}{S^2}\right) + 3\right) + C + 3S \\ &= S\left(\log\left(\frac{T(K-1)}{S^2}\right) + 6\right) + C. \end{aligned}$$

Plugging this and the value of λ into (2.16) completes the proof of (2.6).

Proof of the refined bound in equation (2.7)**:** We note that the range of C in the refined bound in equation (2.7) is the same as in the refined bound in (2.11) in Theorem 2.4 for $B = 1$. We take α^* as in equation (2.19), i.e., $\alpha^* = \sqrt{\frac{-S}{C} W_{-1}\left(\frac{-CS}{e(K-1)T}\right)}$. By Lemma 2.1 we have $-W_{-1}\left(\frac{-CS}{e(K-1)T}\right) \geq 1$, and thus $\alpha^* \geq \sqrt{\frac{S}{C}}$. By plugging this bound and the bound on $h(1, \alpha^*)$ from equation (2.21) into equation (2.32), we obtain:

$$\begin{aligned} R &\leq \sqrt{CS}\left(\sqrt{\log\frac{T(K-1)}{CS}} + 2\right) + S\left(\log\frac{T(K-1)}{CS} + \sqrt{2\log\frac{T(K-1)}{CS}} + 2\right) \\ &\quad + 3\sqrt{CS} \\ &= \sqrt{CS}\left(\sqrt{\log\frac{T(K-1)}{CS}} + 5\right) + S\left(\log\frac{T(K-1)}{CS} + \sqrt{2\log\frac{T(K-1)}{CS}} + 2\right). \end{aligned}$$

Plugging this bound into (2.16) and using the fact that $\lambda \leq 1$ completes the proof of (2.7). □

Chapter 3

A Best-of-Both-Worlds Algorithm for Bandits with Delayed Feedback

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

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.

Abstract

We present a modified tuning of the algorithm of Zimmert and Seldin (2020) for adversarial multiarmed bandits with delayed feedback, which in addition to the minimax optimal adversarial regret guarantee shown by Zimmert and Seldin simultaneously achieves a near-optimal regret guarantee in the stochastic setting with fixed delays. Specifically, the adversarial regret guarantee is $\mathcal{O}(\sqrt{TK} + \sqrt{dT \log K})$, where T is the time horizon, K is the number of arms, and d is the fixed delay, whereas the stochastic regret guarantee is $\mathcal{O}\left(\sum_{i \neq i^*}(\frac{1}{\Delta_i}\log(T) + \frac{d}{\Delta_i \log K}) + dK^{1/3}\log K\right)$, where Δ_i are the suboptimality gaps. We also present an extension of the algorithm to the case of arbitrary delays, which is based on an oracle knowledge of the maximal delay d_{max} and achieves $\mathcal{O}(\sqrt{TK} + \sqrt{D \log K} + d_{max}K^{1/3}\log K)$ regret in the adversarial regime, where D is the total delay, and $\mathcal{O}\left(\sum_{i \neq i^*}(\frac{1}{\Delta_i}\log(T) + \frac{\sigma_{max}}{\Delta_i \log K}) + d_{max}K^{1/3}\log K\right)$ regret in the stochastic regime, where σ_{max} is the maximal number of outstanding observations. Finally, we present a lower bound that matches the refined adversarial regret upper bound achieved by the skipping technique of Zimmert and Seldin (2020) in the adversarial setting.

3.1 Introduction

Delayed feedback is a common challenge in many online learning problems, including multi-armed bandits. The literature studying multi-armed bandit games with delayed feedback builds on prior work on bandit problems with no delays. The researchers have traditionally separated the study of bandit games in stochastic environments (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985; Auer et al., 2002a) and in adversarial environments(Auer et al., 2002b). However, in practice the environments are rarely purely stochastic, whereas they may not be fully adversarial either. Furthermore, the exact nature of an environment is not always known in practice. Therefore, in recent years there has been an increasing interest in algorithms that perform well in both regimes with no prior knowledge of the regime (Bubeck and Slivkins, 2012; Seldin and Slivkins, 2014; Auer and Chiang, 2016; Seldin and Lugosi, 2017; Wei and Luo, 2018). The quest for best-of-both-worlds algorithms for no-delay setting culminated with the Tsallis-INF algorithm proposed by Zimmert and Seldin (2019), which achieves the optimal regret bounds in both stochastic and adversarial environments. The algorithm and analysis were further improved by Zim-

mert and Seldin (2021) and Masoudian and Seldin (2021), who, in particular, derived improved regret bounds for intermediate regimes between stochastic and adversarial, while Ito (2021) removed an assumption on uniqueness of the best arm, which was used in the early works.

Our goal is to extend best-of-both-worlds results to multi-armed bandits with delayed feedback. So far the literature on multi-armed bandits with delayed feedback has followed the traditional separation into stochastic and adversarial. In the stochastic regime Joulani et al. (2013) showed that if the delays are random (generated i.i.d), then compared to the non-delayed stochastic multi-armed bandit setting, the regret only increases additively by a factor that is proportional to the expected delay. In the adversarial setting Cesa-Bianchi et al. (2019) have studied the case of uniform delays d. They derived a lower bound $\Omega(\max(\sqrt{KT}, \sqrt{dT \log K}))$ and an almost matching upper bound $\mathcal{O}(\sqrt{KT \log K} + \sqrt{dT \log K})$. Thune et al. (2019) and Bistritz et al. (2019) extended the results to arbitrary delays, achieving $\mathcal{O}(\sqrt{KT \log K} + \sqrt{D \log K})$ regret bounds based on oracle knowledge of the total delay D and time horizon T. Thune et al. (2019) also proposed a skipping technique based on advance knowledge of the delays "at action time", which allowed to exclude excessively large delays from D. Finally, Zimmert and Seldin (2020) introduced an FTRL algorithm with a hybrid regularizer that achieved $\mathcal{O}(\sqrt{KT} + \sqrt{D \log K})$ regret bound, matching the lower bound in the case of uniform delays and requiring no prior knowledge of D or T. The regularizer used by Zimmert and Seldin was a mix of the negative Tsallis entropy regularizer used in the Tsallis-INF algorithm for bandits and the negative entropy regularizer used in the Hedge algorithm for full information games, mixed with separate learning rates:

$$F_t(x) = -2\eta_t^{-1} \left(\sum_{i=1}^{K} \sqrt{x_i} \right) + \gamma_t^{-1} \left(\sum_{i=1}^{K} x_i (\log x_i - 1) \right). \tag{3.1}$$

Zimmert and Seldin (2020) also improved the skipping technique and achieved a refined regret bound $\mathcal{O}(\sqrt{KT} + \min_S(|S| + \sqrt{D_{\bar{S}} \log K}))$, where S is a set of skipped rounds and $D_{\bar{S}}$ is the total delay in non-skipped rounds. The refined skipping technique requires no advance knowledge of the delays. Their key step toward elimination of the need of advance knowledge of delays was to base the analysis on the count of the number of outstanding observations rather than the delays. The great advantage of skipping is that a few rounds with excessively large or potentially even infinite delays have a very limited impact on the regret bound. One of our contributions in this paper is a lower bound for the case of non-uniform delays, which matches the refined regret upper bound achieved by skipping.

Even though the hybrid regularizer used by Zimmert and Seldin (2020) was sharing the Tsallis entropy part with their best-of-both-worlds Tsallis-INF algorithm from Zimmert and Seldin (2021), and even though the adversarial analysis was partly similar to the analysis of the Tsallis-INF algorithm, Zimmert and Seldin (2020) did not manage to derive a regret bound for their algorithm in the stochastic setting with delayed feedback and left it as an open problem. The stochastic analysis of the Tsallis-INF algorithm is based on the self-bounding technique (Zimmert and Seldin, 2021). Application of this technique in the no delay setting is relatively straightforward, but in presence of delays it requires control of the drift of the playing distribution from the moment an action is played to the moment the feedback arrives. Cesa-Bianchi et al. (2019) have bounded the drift of the playing distribution of the EXP3 algorithm in the uniform delays setting with a fixed learning rate. But best-of-both-worlds algorithms require decreasing learning rates (Mourtada and Gaïffas, 2019), which makes the drift control much more challenging. The problem gets even more challenging in the case of arbitrary delays, because it requires drift control over arbitrary long periods of time.

We apply an FTRL algorithm with the same hybrid regularizer as the one used by Zimmert and Seldin (2020), but with a different tuning of the learning rates. The new tuning has a minor effect on the adversarial regret bound, but allows us to make progress with the stochastic analysis. For the stochastic analysis we use the self-bounding technique. One of our key contributions is a general lemma that bounds the drift of the playing distribution derived from the time-varying hybrid regularizer over arbitrary delays. Using this lemma we derive near-optimal best-of-both-worlds regret guarantees for the case of fixed delays. But even with the lemma at hand, application of the self-bounding technique in presence of arbitrary delays is still much more challenging than in the no delays or fixed delay setting. Therefore, we resort to introducing an assumption of oracle knowledge of the maximal delay, which limits the maximal period of time over which we need to keep control over the drift. Our contributions are summarized below. To keep the presentation simple we assume uniqueness of the best arm throughout the paper. Tools for eliminating the uniqueness of the best arm assumption were proposed by Ito (2021).

1. We show that in the arbitrary delays setting with an oracle knowledge of the maximal delay d_{max}, our algorithm achieves $\mathcal{O}(\sqrt{KT} + \sqrt{D \log K} + d_{max} K^{1/3} \log K)$ regret bound in the adversarial regime simultaneously with $\mathcal{O}\left(\sum_{i \neq i^*} (\frac{\log T}{\Delta_i} + \frac{\sigma_{max}}{\Delta_i \log K}) + d_{max} K^{1/3} \log K\right)$ regret bound in the stochastic regime, where σ_{max} is the maximal number of outstanding observations. We note that $\sigma_{max} \leq d_{max}$, but it may potentially be much smaller. For example,

if the first observation has a delay of T and all the remaining observations have zero delay, then $d_{max} = T$, but $\sigma_{max} = 1$.

2. In the case of uniform delays the above bounds simplify to $\mathcal{O}(\sqrt{KT} + \sqrt{dT \log K} + dK^{1/3} \log K)$ in the adversarial case and $\mathcal{O}\left(\sum_{i \neq i^*} (\frac{\log T}{\Delta_i} + \frac{d}{\Delta_i \log K}) + dK^{1/3} \log K\right)$ in the stochastic case. For $T \geq dK^{2/3} \log K$ the last term in the adversarial regret bound is dominated by the middle term, which leads to the minimax optimal $\mathcal{O}(\sqrt{KT} + \sqrt{dT \log K})$ adversarial regret. The stochastic regret lower bound is trivially $\Omega(\min\{d\frac{\sum_{i \neq i*} \Delta_i}{K}, \sum_{i \neq i^*} \frac{\log T}{\Delta_i}\}) = \Omega(d\frac{\sum_{i \neq i*} \Delta_i}{K} + \sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ and, therefore, our stochastic regret upper bound is near-optimal.

3. We present an $\Omega\left(\sqrt{KT} + \min_S(|S| + \sqrt{D_{\bar{S}} \log K})\right)$ regret lower bound for adversarial multi-armed bandits with non-uniformly delayed feedback, which matches the refined regret upper bound achieved by the skipping technique of Zimmert and Seldin (2020).

3.2 Problem setting

We study the multi-armed bandit with delays problem, in which at time $t = 1, 2, \ldots$ the learner chooses an arm I_t among a set of K arms and instantaneously suffers a loss ℓ_{t,I_t} from a loss vector $\ell_t \in [0,1]^K$ generated by the environment, but ℓ_{t,I_t} is not observed by the learner immediately. After a delay of d_t, at the end of round $t + d_t$, the learner observes the pair (t, ℓ_{t,I_t}), namely, the loss and the index of the game round the loss is coming from. The sequence of delays $d_1, d_2, \ldots$ is selected arbitrarily by the environment. Without loss of generality we can assume that all the outstanding observations are revealed at the end of the game, i.e., $t + d_t \leq T$ for all t, where T is the time horizon, unknown to the learner. We consider two regimes, oblivious adversarial and stochastic.

The performance of the learner is evaluated using pseudo-regret, which is defined as

$$\overline{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_{i \in [K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_{t,I_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* \in \operatorname{argmin}_{i \in [K]} \mathbb{E}\left[\sum_{t=t}^{T} \ell_{t,i}\right]$ is a best arm in hindsight in expectation over the loss generation model and the randomness of the learner. In the oblivious adver-

sarial setting the losses are independent of the actions taken by the algorithm and considered to be deterministic, and the pseudo-regret is equal to the expected regret.

Additional Notation: We use Δ^n to denote the probability simplex over $n+1$ points. The characteristic function of a closed convex set $\mathcal{A}$ is denoted by $\mathcal{I}_{\mathcal{A}}(x)$ and satisfies $\mathcal{I}_{\mathcal{A}}(x) = 0$ for $x \in \mathcal{A}$ and $\mathcal{I}_{\mathcal{A}}(x) = \infty$ otherwise. The convex conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by $f^*(y) = \sup_{x \in \mathbb{R}^n}\{\langle x, y\rangle - f(x)\}$. We also use bar to denote that the function domain is restricted to Δ^n, e.g., $\bar{f}(x) = \begin{cases} f(x), & \text{if } x \in \Delta^n \\ \infty, & \text{otherwise} \end{cases}$. We denote the indicator function of an event $\mathcal{E}$ by $\mathbb{1}(\mathcal{E})$ and use $\mathbb{1}_t(i)$ as a shorthand for $\mathbb{1}(I_t = i)$. The probability distribution over arms that is played by the learner at round t is denoted by $x_t \in \Delta^{K-1}$.

3.3 Algorithm

The algorithm is based on Follow The Regularized Leader (FTRL) algorithm with the hybrid regularizer used by Zimmert and Seldin (2020), stated in equation (3.1). At each time step t let $\sigma_t = \sum_{s=1}^{t-1} \mathbb{1}(s + d_s \geq t)$ be the number of outstanding observations and $\mathcal{D}_t = \sum_{s=1}^{t} \sigma_t$ be the cumulative number of outstanding observations, then the learning rates are defined as

$$\eta_t^{-1} = \sqrt{t + \eta_0}, \qquad \gamma_t^{-1} = \sqrt{\frac{\sum_{s=1}^{t} \sigma_s + \gamma_0}{\log K}}, \tag{3.2}$$

where $\eta_0 = 10 d_{max} + d_{max}^2 / \left(K^{1/3} \log(K)\right)^2$ and $\gamma_0 = 24^2 d_{max}^2 K^{2/3} \log(K)$. The update rule for the distribution over actions played by the learner is

$$x_t = \nabla \bar{F}_t^*(-\hat{L}_t^{obs}) = \arg\min_{x \in \Delta^{K-1}} \langle \hat{L}_t^{obs}, x\rangle + F_t(x), \tag{3.3}$$

where $\hat{L}_t^{obs} = \sum_{s=1}^{t-1} \hat{\ell}_s \mathbb{1}(s + d_s < t)$ is the cumulative importance-weighted observed loss and $\hat{\ell}_s$ is an importance-weighted estimate of the loss vector ℓ_s defined by

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i} \mathbb{1}(I_t = i)}{x_{t,i}}.$$

At the beginning of round t the algorithm calculates the cumulative number of outstanding observations $\mathcal{D}_t$ and uses it to define the learning rate γ_t. Next, it

uses the FTRL update rule defined in (3.3) to define a distribution over actions x_t from which to draw action I_t. Finally, at the end of round t it receives the delayed observations and updates the cumulative loss estimation vector accordingly, so that $\hat{L}_{t+1}^{obs} = \hat{L}_t^{obs} + \sum_{s=1}^{t} \hat{\ell}_s \mathbb{1}(s + d_s = t)$. The complete algorithm is provided in Algorithm 2.

Algorithm 2: FTRL with advance tuning for delayed bandit

Initialize $\mathcal{D}_0 = 0$ *and* $\hat{L}_1^{obs} = \mathbf{0}_K$ *(where* $\mathbf{0}_K$ *is a zero vector in* $\mathbb{R}^K$*)*
for $t = 1, \ldots, T$ **do**
  Set $\sigma_t = \sum_{s=1}^{t-1} \mathbb{1}(s + d_s > t)$
  Update $\mathcal{D}_t = \mathcal{D}_{t-1} + \sigma_t$
  Set $x_t = \arg\min_{x \in \Delta^{K-1}} \langle \hat{L}_t^{obs}, x \rangle + F_t(x)$ *// F_t is defined in (3.1) and η_t and γ_t in (3.2)*
  Sample $I_t \sim x_t$
  Observe (s, ℓ_{s,I_s}) for all s that satisfy $s + d_s = t$
  $\hat{L}_{t+1}^{obs} = \hat{L}_t^{obs} + \sum_{s=1}^{t} \hat{\ell}_s \mathbb{1}(s + d_s = t)$

3.4 Best-of-both-worlds regret bounds for Algorithm 2

In this section we provide best-of-both-worlds regret bounds for Algorithm 2. First, in Theorem 3.1 we provide regret bounds for an arbitrary delay setting, where we assume an oracle access to d_{max}. Then, in Corollary 3.1 we specialize the result to a fixed delay setting.

Theorem 3.1. *Assume that Algorithm 2 is given an oracle knowledge of d_{max}. Then its pseudo-regret for any sequence of delays and losses satisfies*

$$\overline{Reg}_T = \mathcal{O}(\sqrt{TK} + \sqrt{D \log K} + d_{max} K^{1/3} \log K).$$

Furthermore, in the stochastic regime the pseudo-regret additionally satisfies

$$\overline{Reg}_T = \mathcal{O}\left(\sum_{i \neq i^*} \left(\frac{1}{\Delta_i} \log(T) + \frac{\sigma_{max}}{\Delta_i \log K} \right) + d_{max} K^{1/3} \log K \right).$$

A sketch of the proof is provided in Section 3.5 and detailed constants are worked out in Appendix 3.8.3. For fixed delays Theorem 3.1 gives the following corollary.

Corollary 3.1. *If the delays are fixed and equal to d, and $T \geq dK^{2/3} \log K$, then the pseudo-regret of Algorithm 2 always satisfies*

$$\overline{Reg}_T = \mathcal{O}(\sqrt{TK} + \sqrt{dT \log K})$$

and in the stochastic setting it additionally satisfies

$$\overline{Reg}_T = \mathcal{O}\left(\sum_{i \neq i^*} (\frac{1}{\Delta_i} \log(T) + \frac{d}{\Delta_i \log K}) + dK^{1/3} \log K\right).$$

In the adversarial regime with fixed delays d, regret lower bound is $\Omega\left(\sqrt{KT} + \sqrt{dT \log K}\right)$, whereas in the stochastic regime with fixed delays the regret lower bound is trivially $\Omega(d\frac{\sum_{i \neq i*} \Delta_i}{K} + \sum_{i \neq i^*} \frac{\log T}{\Delta_i})$. Thus, in the adversarial regime the corollary yields the minimax optimal regret bound and in the stochastic regime it is near-optimal. More explicitly, it is optimal within a multiplicative factor of $\sum_{i \neq i^*} \frac{1}{\Delta_i \log K} + \frac{K^{4/3} \log K}{\sum_{i \neq i*} \Delta_i}$ in front of d.

If we fix a total delay budget D, then uniform delays $d = D/T$ is a special case, and in this sense Theorem 3.1 is also optimal in the adversarial regime and near-optimal in the stochastic regime, although for non-uniform delays improved regret bounds can potentially be achieved by skipping. We also note that having the dependence on σ_{max} in the middle term of the stochastic regret bound in Theorem 3.1 is better than having a dependence on d_{max}, since $\sigma_{max} \leq d_{max}$, and in some cases it can be significantly smaller, as shown in the example in the Introduction and quantified by the following lemma.

Lemma 3.1. *Let $d_{max}(S) = \max_{s \in S} d_s$, where $S \subseteq \{1, \ldots, T\}$ is a subset of rounds. Let $\bar{S} = \{1, \ldots, T\} \setminus S$ be the remaining rounds. Then*

$$\sigma_{max} \leq \min_{S \subseteq \{1, \ldots, T\}} \left\{ |S| + d_{max}(\bar{S}) \right\}.$$

A proof of Lemma 3.1 is provided in Appendix 3.8.1.

Finally, we note that the result in Theorem 3.1 is easily extendable to the corrupted regime, because the proof relies on the same self-bounding technique as the one used by Zimmert and Seldin (2021). If we denote by B_T^{stoch} the regret upper bound in the stochastic regime in Theorem 3.1 and by C the total corruption budget, then in the corrupted regime the regret would be $\mathcal{O}(B_T^{stoch} + \sqrt{B_T^{stoch} C})$. The proof is straightforward, following the lines of Zimmert and Seldin (2021), and, therefore, left out.

3.5 A proof sketch of Theorem 3.1

In this section we provide a sketch of a proof of Theorem 3.1. We provide a proof sketch for the stochastic bound in Section 3.5.1. Afterwards, in Section 3.5.2, we show how the analysis of Zimmert and Seldin (2020) gives the adversarial bound stated in Theorem 3.1.

3.5.1 Stochastic Bound

We start by providing a key lemma (Lemma 3.2) that controls the drift of the playing distribution derived from the time-varying hybrid regularizer over arbitrary delays. We then introduce a drifted version of the pseudo-regret defined in (3.4), for which we use the key lemma to show that the drifted version of the pseudo-regret is close to the actual one. As a result, it is sufficient to bound the drifted version. The analysis of the drifted pseudo-regret follows by the standard analysis of the FTRL algorithm (Lattimore and Szepesvári, 2020) that decomposes the pseudo-regret (drifted pseudo-regret in our case) into stability and penalty terms. Thereafter, we proceed by using Lemma 3.2 again, this time to bound the stability term in order to apply the self-bounding technique (Zimmert and Seldin, 2019), which yields logarithmic regret in the stochastic setting. Our key lemma is the following.

Lemma 3.2 (The Key Lemma)**.** *For any $i \in [K]$ and $s, t \in [T]$, where $s \leq t$ and $t - s \leq d_{max}$, we have*

$$x_{t,i} \leq 2x_{s,i}.$$

A detailed proof of the lemma is provided in Appendix 3.8.2. Below we explain the high level idea behind the proof.

Proof sketch. We know that $x_t = \nabla \bar{F}_t^*(-\hat{L}_t^{obs})$ and $x_s = \nabla \bar{F}_s^*(-\hat{L}_s^{obs})$, so we introduce $\tilde{x} = \nabla \bar{F}_s^*(-\hat{L}_t^{obs})$ as an auxiliary variable to bridge between x_t and x_s. The analysis consists of two key steps and is based on induction on (t, s).
Deviation Induced by the Loss Shift: This step controls the drift when we fix the learning rates and shift the cumulative loss. We prove the following inequality:

$$\tilde{x}_i \leq \frac{3}{2} x_{s,i}.$$

Note that this step uses the induction assumption for $(s, s - d_r)$ for all $r < s : r + d_r = s$.
Deviation Induced by the Change of Regularizer: In this step we bound the

drift when the cumulative loss vector is fixed and we change the regularizer. We show that

$$x_{t,i} \leq \frac{4}{3}\tilde{x}_i.$$

Combining these two steps gives us the desired bound. A proof of these steps is provided in Appendix 3.8.2. □

We use Lemma 3.2 to relate the drifted pseudo-regret to the actual pseudo-regret. Let $A_t = \{s : s \leq t \text{ and } s + d_s = t\}$ be the set of rounds for which feedback arrives at round t. We define the observed loss vector at time t as $\hat{\ell}_t^{obs} = \sum_{s \in A_t} \hat{\ell}_s$ and the drifted pseudo-regret as

$$\overline{Reg}_T^{drift} = \mathbb{E}\left[\sum_{t=1}^{T}\left(\langle x_t, \hat{\ell}_t^{obs}\rangle - \hat{\ell}_{t,i_T^*}^{obs}\right)\right]. \tag{3.4}$$

We rewrite the drifted regret as

$$\begin{aligned}
\overline{Reg}_T^{drift} &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s\in A_t}\left(\langle x_t, \hat{\ell}_s\rangle - \hat{\ell}_{s,i_T^*}\right)\right] \\
&= \sum_{t=1}^{T}\sum_{s\in A_t}\sum_{i=1}^{K}\mathbb{E}[x_{t,i}(\hat{\ell}_{s,i} - \hat{\ell}_{s,i_T^*})] \\
&= \sum_{t=1}^{T}\sum_{s\in A_t}\sum_{i=1}^{K}\mathbb{E}[x_{t,i}]\Delta_i = \sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}[x_{t+d_t,i}]\Delta_i,
\end{aligned}$$

where when taking the expectation we use the facts that $\hat{\ell}_s$ has no impact on the determination of x_t and that the loss estimators are unbiased. Using Lemma 3.2 we make a connection between pseudo-regret and the drifted version:

$$\begin{aligned}
\overline{Reg}_T^{drift} = \sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}[x_{t+d_t,i}]\Delta_i &\geq \sum_{t=1}^{T-d_{max}}\sum_{i=1}^{K}\frac{1}{2}\mathbb{E}[x_{t+d_{max},i}]\Delta_i \\
&= \frac{1}{2}\sum_{t=d_{max}+1}^{T}\sum_{i=1}^{K}\mathbb{E}[x_{t,i}]\Delta_i \\
&\geq \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{K}\mathbb{E}[x_{t,i}]\Delta_i - \frac{d_{max}}{2} = \frac{1}{2}\overline{Reg}_T - \frac{d_{max}}{2},
\end{aligned}$$

where the first inequality follows by Lemma 3.2, and the second inequality uses $\sum_{t=1}^{d_{max}} \mathbb{E}[x_{t,i}]\Delta_i \leq d_{max}$. As a result, we have $\overline{Reg}_T \leq 2\overline{Reg}_T^{drift} + d_{max}$ and it suffices to upper bound $\overline{Reg}_T^{drift}$. We follow the standard analysis of FTRL, which decomposes the drifted pseudo-regret into *stabiltiy* and *penalty* terms as

$$\begin{aligned}\overline{Reg}_T^{drift} =& \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \langle x_t, \hat{\ell}_t^{obs}\rangle + \bar{F}_t^*(-\hat{L}_{t+1}^{obs}) - \bar{F}_t^*(-\hat{L}_t^{obs})}_{stability}\right] \\ &+ \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \bar{F}_t^*(-\hat{L}_t^{obs}) - \bar{F}_t^*(-\hat{L}_{t+1}^{obs}) - \ell_{t,i_T^*}}_{penalty}\right].\end{aligned}$$

For the penalty term we have the following bound by Abernethy et al. (2015)

$$penalty \leq \sum_{t=2}^{T} (F_{t-1}(x_t) - F_t(x_t)) + F_T(\mathrm{e}_{i_T^*}) - F_1(x_1),$$

where $\mathrm{e}_{i_T^*}$ denotes a the unit vector in $\mathbb{R}^K$ with the i_T^*-th element being one and zero elsewhere. By replacing the closed form of the regularizer in this bound and using the facts that $\eta_t^{-1} - \eta_{t-1}^{-1} = \mathcal{O}(\eta_t)$, $\gamma_t^{-1} - \gamma_{t-1}^{-1} = \mathcal{O}(\sigma_t\gamma_t/\log K)$, and $x_{t,i_T^*}^{\frac{1}{2}} - 1 \leq 0$, we obtain

$$penalty \leq \mathcal{O}\left(\sum_{t=2}^{T}\sum_{i\neq i^*} \eta_t x_{t,i}^{\frac{1}{2}} + \sum_{t=2}^{T}\sum_{i=1}^{K} \frac{\sigma_t\gamma_t x_{t,i}\log(1/x_{t,i})}{\log K}\right) + 2\sqrt{\eta_0(K-1)} + \sqrt{\gamma_0 \log K}. \tag{3.5}$$

In order to control the stability term we derive Lemma 3.3.

Lemma 3.3 (Stability)**.** *Let $\upsilon_t = |A_t|$. For any $\alpha_t \leq \gamma_t^{-1}$ we have*

$$stability \leq \sum_{t=1}^{T}\sum_{i=1}^{K} 2f_t''(x_{t,i})^{-1}(\hat{\ell}_{t,i}^{obs} - \alpha_t)^2.$$

Furthermore, $\alpha_t = \frac{\sum_{j=1}^K f''(x_{t,j})^{-1}\hat{\ell}_{t,j}^{obs}}{\sum_{j=1}^K f''(x_{t,j})^{-1}}$ *satisfies* $\alpha_t \le \gamma_t^{-1}$ *and yields*

$$\mathbb{E}[\textit{stability}] \le \sum_{t=1}^{T}\sum_{i\neq i^*} 2\gamma_t(\upsilon_t - 1)\upsilon_t \mathbb{E}[x_{t,i}]\Delta_i + \sum_{t=1}^{T}\sum_{s\in A_t}\sum_{i=1}^{K} 2\eta_t \mathbb{E}[x_{t,i}^{3/2}x_{s,i}^{-1}(1-x_{s,i})]. \tag{3.6}$$

A proof of the stability lemma is provided in Appendix 3.8.1.6. We apply Lemma 3.2 to (3.6) to give bounds $\upsilon_t x_{t,i} = \sum_{s\in A_t} x_{t,i} \le 2\sum_{s\in A_t} x_{s,i}$ and $x_{t,i}^{3/2}x_{s,i}^{-1}(1-x_{s,i}) \le 2^{3/2}x_{s,i}^{1/2}(1-x_{s,i})$. Moreover, in order to remove the best arm i^* from the summation in the later bound we use $x_{s,i^*}^{1/2}(1-x_{s,i^*}) \le \sum_{i\neq i^*} x_{s,i} \le \sum_{i\neq i^*} x_{s,i}^{1/2}$. These bounds together with the facts that we can change the order of the summations and that each t belongs to exactly one A_s, gives us the following stability bound

$$\mathbb{E}[\textit{stability}] = \mathcal{O}\left(\sum_{t=1}^{T}\sum_{i\neq i^*}\eta_t\mathbb{E}[x_{t,i}^{1/2}] + \sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t}(\upsilon_{t+d_t}-1)\mathbb{E}[x_{t,i}]\Delta_i\right). \tag{3.7}$$

By combining (3.7), (3.5), and the fact that $\overline{Reg}_T \le 2\overline{Reg}_T^{drift} + d_{max}$, we show that there exist constants $a, b, c \ge 0$, such that

$$\overline{Reg}_T \le \mathbb{E}\left[\underbrace{a\sum_{t=1}^{T}\sum_{i\neq i^*}\eta_t x_{t,i}^{1/2}}_{A} + \underbrace{b\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t}(\upsilon_{t+d_t}-1)x_{t,i}\Delta_i}_{B} + \underbrace{c\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\sigma_t\gamma_t x_{t,i}\log(1/x_{t,i})}{\log K}}_{C}\right] + \underbrace{4\sqrt{\eta_0(K-1)} + 2\sqrt{\gamma_0\log K} + d_{max}}_{D}. \tag{3.8}$$

Self bounding analysis: We use the self-bounding technique to write $\overline{Reg}_T = 4\overline{Reg}_T - 3\overline{Reg}_T$, and then based on (3.8) we have

$$\overline{Reg}_T \le \mathbb{E}\left[4aA - \overline{Reg}_T\right] + \mathbb{E}\left[4bB - \overline{Reg}_T\right] + \mathbb{E}\left[4cC - \overline{Reg}_T\right] + 4D. \tag{3.9}$$

For D we can substitute the values of γ_0 and η_0 and get

$$D = \mathcal{O}(d_{max}(K-1)^{1/3}\log K). \tag{3.10}$$

Upper bounding A, B, and C requires separate and elaborate analysis, which we do in Lemmas 3.4, 3.5 and 3.6, respectively. Proofs of these lemmas are provided in Appendix 3.8.1.2.

Lemma 3.4 (A bound for $4aA - \overline{Reg}_T$)**.** *We have the following bound for any $a \geq 0$:*

$$4aA - \overline{Reg}_T \leq \sum_{i \neq i^*} \frac{4a^2}{\Delta_i} \log(T/\eta_0 + 1) + 1. \tag{3.11}$$

Lemma 3.4 contributes the logarithmic (in T) term to the regret bound.

Lemma 3.5 (A bound for $4bB - \overline{Reg}_T$)**.** *Let $\upsilon_{max} = \max_{t \in [T]} \upsilon_t$, then for any $b \geq 0$:*

$$4bB - \overline{Reg}_T \leq 64b^2 \upsilon_{max} \log K. \tag{3.12}$$

It is evident that $\upsilon_{max} \leq \sigma_{max} \leq d_{max}$, so the bound in Lemma 3.5 contributes an $\mathcal{O}(d_{max} \log K)$ term to the regret bound.

Lemma 3.6 (A bound for $4cC - \overline{Reg}_T$)**.** *For any $c \geq 0$:*

$$4cC - \overline{Reg}_T \leq \sum_{i \neq i^*} \frac{128c^2 \sigma_{max}}{\Delta_i \log K}. \tag{3.13}$$

Part of the pseudo-regret bound that corresponds to Lemma 3.6 comes from the penalty term related to the negative entropy part of the regularizer. In this part, despite the fact that σ_{max} can be much smaller than d_{max} (Lemma 3.1), the $\sum_{i \neq i^*} \frac{\sigma_{max}}{\Delta_i \log K}$ term could be very large when the suboptimality gaps are small. In Appendix 3.8.4 we show how an asymmetric oracle learning rate $\gamma_{t,i} \simeq \gamma_t / \sqrt{\Delta_i}$ for the negative entropy regularizer can be used to remove the $\sum_{i \neq i^*} 1/\Delta_i$ factor in front of σ_{max}. The possibility of removing this factor without the oracle knowledge is left as an open question.

Finally, by plugging (3.10),(3.11),(3.12),(3.13) into (3.9) we obtain the desired regret bound.

3.5.2 Adversarial bound

For the adversarial regime we use the final bound of Zimmert and Seldin (2021), which holds for any non-increasing learning rates:

$$\overline{Reg}_T \leq \sum_{t=1}^{T} \eta_t \sqrt{K} + \sum_{t=1}^{T} \gamma_t \sigma_t + 2\eta_T^{-1} \sqrt{K} + \gamma_T^{-1} \log K.$$

It suffices to substitute the values of the learning rates and use Lemma 3.7 for function $\frac{1}{\sqrt{x}}$:

$$\begin{aligned}\overline{Reg}_T &\leq \sum_{t=1}^{T} \frac{\sqrt{K}}{\sqrt{t+\eta_0}} + \sum_{t=1}^{T} \frac{\sigma_t \sqrt{\log K}}{\sqrt{D_t + \gamma_0}} + 2\sqrt{KT + K\eta_0} + \sqrt{\log(K) D_T + \gamma_0 \log(K)} \\ &= \mathcal{O}\left(\sqrt{KT} + \sqrt{\log(K) D_T} + d_{max} K^{1/3} \log K\right).\end{aligned}$$

3.6 Refined lower bound

In this section, we prove a tight lower bound for adversarial regret with arbitrary delays. Thune et al. (2019) have proposed a skipping technique to achieve refined regret upper bounds in the adversarial regime with non-uniform delays. The technique was improved by Zimmert and Seldin (2020), but it remained unknown whether the refined regret bounds for regimes with non-uniform delays are tight. We answer this question positively by showing that the regret bound of Zimmert and Seldin (2020) is not improvable without additional assumptions. We first derive a refined lower bound for full-information games with variable loss ranges, which might be of independent interest. A proof is provided in Appendix 3.8.5.

Theorem 3.2. *Let $L_1 \geq L_2 \geq \cdots \geq L_T \geq 0$ be a non-increasing sequence of positive reals and assume that there exists a permutation $\rho : [T] \to [T]$, such that the losses at time t are bounded in $[0, L_{\rho(t)}]^K$. The minimax regret Reg^* in the corresponding adversarial full-information game satisfies*

$$Reg^* \geq \max\left\{\frac{1}{2}\sum_{t=1}^{\lfloor \log_2(K) \rfloor} L_t, \frac{1}{32}\sqrt{\sum_{t=\lfloor \log_2(K) \rfloor}^{T} L_t^2 \log(K)}\right\}.$$

From here we can directly obtain a lower bound for the full-information game with variable delays. This implies the same lower bound for bandits, since we have strictly less information available.

Corollary 3.2. *Let $(d_t)_{t=1}^T$ be a sequence of non-increasing delays, such that $d_t \leq T+1-t$ and let an oblivious adversary select all loss vectors $(\ell_t)_{t=1}^T$ in $[0,1]^K$ before the start of the game. The minimax regret of the full-information game is bounded from below by*

$$Reg^* = \Omega\left(\min_{S \subset [T]} |S| + \sqrt{D_{\bar{S}} \log(K)}\right), \text{ where } D_{\bar{S}} = \sum_{t \in [T] \setminus S} d_t.$$

Proof. We divide the time horizon greedily into M buckets, such that the actions for all timesteps inside a bucket have to be chosen before the first feedback from any timestep inside the bucket is received. In other words, let bucket $B_m = \{b_m, \dots, b_{m+1} - 1\}$, then $\forall t \in B_m : t + d_t > b_{m+1} - 1$, while $\exists t \in B_m : t + d_t = b_{m+1}$. This division of buckets has the following properties:

(i) monotonically decreasing sizes: $|B_1| \geq |B_2| \geq \cdots \geq |B_M|$.

(ii) upper bound on the sum of delays: $\forall m \in [M-1] : |B_m|^2 \geq \sum_{t \in B_{m+1}} d_t$.

Both properties follow directly from the non-decreasing nature of the delays.

$$
\begin{aligned}
|B_m| &= b_{m+1} - b_m \leq b_m + d_{b_m} - b_m = d_{b_m} \\
|B_m| &= \min_{t \in B_m} \{d_t + t - b_m\} \geq d_{b_{m+1}-1} + \min_{t \in B_m} \{t - b_m\} \geq d_{b_{m+1}-1} \,.
\end{aligned}
$$

Hence

$$
\begin{aligned}
&|B_m| \geq d_{b_{m+1}-1} \geq d_{b_{m+1}} \geq |B_{m+1}| \,, \\
&\textstyle\sum_{t \in B_{m+1}} d_t \leq |B_{m+1}| \cdot d_{b_{m+1}} \leq |B_{m+1}| \cdot |B_m| \leq |B_m|^2 \,.
\end{aligned}
$$

Set $S' = \bigcup_{m=1}^{\lfloor \log_2(K) \rfloor} B_m$ and let the adversary set all losses within a bucket to the same value, then the game reduces to a full information game over M rounds with loss ranges $|B_1|, |B_2|, \dots, |B_M|$. Applying Theorem 3.2 yields

$$
\begin{aligned}
Reg^* &\geq \max \left\{ \frac{1}{2} \sum_{m=1}^{\lfloor \log_2(K) \rfloor} |B_m|, \frac{1}{32} \sqrt{\sum_{m=\lfloor \log_2(K) \rfloor}^{M} |B_m|^2 \log(K)} \right\} \\
&\geq \max \left\{ \frac{1}{2} |S'|, \frac{1}{32} \sqrt{\sum_{t \in \bar{S}'} d_t \log(K)} \right\} = \Omega \left(\min_{S \subset [T]} |S| + \sqrt{\sum_{t \in \bar{S}} d_t \log(K)} \right) \,.
\end{aligned}
$$

□

3.7 Discussion

We have presented a best-of-both-worlds analysis of a slightly modified version of the algorithm of Zimmert and Seldin (2020) for bandits with delayed feedback. The key novelty of our analysis is the control of the drift of the playing distribution over arbitrary, but bounded, time intervals when the learning rate is changing over time.

This control is necessary for best-of-both-worlds guarantees, but it is much more challenging than the drift control over fixed time intervals with fixed learning rate that appeared in prior work.

We also presented an adversarial regret lower bound matching the skipping-based refined regret upper bound of Zimmert and Seldin (2020) within constants.

Our work leads to several exciting open questions. The main one is whether skipping can be used to eliminate the need in oracle knowledge of d_{max}. If possible, this would remedy the deterioration of the adversarial bound by the additive factor of d_{max}, because the skipping threshold would be dominated by $\sqrt{D_{\bar{S}} \log K}$. Another open question is whether the $\frac{\sigma_{max}}{\Delta_i}$ term can be eliminated from the stochastic bound. Yet another open question is whether the d_{max} factor in the stochastic bound can be reduced to σ_{max} and whether the multiplicative terms dependent on K can be eliminated. An extension of the results to first order bounds, that depend on the cumulative loss of the best action rather than T, and extension to arm dependent delays are also open questions. For now it was only done in the adversarial setting (Gyorgy and Joulani, 2021; Van der Hoeven and Cesa-Bianchi, 2022).

3.8 Appendix

3.8.1 Proofs of the lemmas for the analysis of Algorithm 2

3.8.1.1 A proof of Lemma 3.1

Proof. Let $S \subseteq \{1, \dots, T\}$ and $\bar{S} = \{1, \dots, T\} \setminus S$ be an arbitrary split of the game rounds. Consider the number of outstanding observations σ_t at an arbitrary round t. The number σ_t is bounded by the sum of the number of outstanding observations from actions taken in the rounds in S and the number of outstanding observations from actions taken in the rounds in $\bar{S}$. The former is bounded by $|S|$, and the latter is bounded by $d_{max}(\bar{S})$, since by definition of $d_{max}(\bar{S})$ any observation from an action taken in a round in $\bar{S}$ can be outstanding for at most $d_{max}(\bar{S})$ rounds. Since this holds for any split of the rounds $\{1, \dots, T\}$ into S and $\bar{S}$, we have $\sigma_{max} = \max_t \sigma_t \leq \min_{S \subseteq \{1,\dots,T\}} \big(|S| + d_{max}(\bar{S})\big)$. □

3.8.1.2 Proofs of the lemmas supporting the proof of Theorem 3.1

We start with providing some auxiliary lemmas.

Lemma 3.7 (Integral inequality: Lemma 4.13 of Orabona (2019))**.** *Let $g(x)$ be a positive nonincreasing function, then for any non-negative sequence $\{z_n\}_{n \in \{0,\dots,N\}}$ we*

have

$$\sum_{n=1}^{N} z_n g\left(\sum_{i=0}^{n} z_i\right) \leq \int_{z_0}^{\sum_{i=0}^{N} z_i} g(x)dx.$$

Lemma 3.8. *Let σ_t and υ_t be the number of outstanding observations and arriving observations at time t, respectively, then the following inequality holds for all t*

$$\sum_{s=1}^{t} \sigma_s \geq \sum_{s=1}^{t} \frac{\upsilon_s^2 - \upsilon_s}{2}.$$

Proof. Note that $A_s = \{r : r + d_r = t\}$. We define $D_s = \{d_r : r \in A_s\}$ be the set of delays corresponding to observations that arrive at round s, then D_s must have $\upsilon_s = |A_s|$ different number of elements, because $\forall r \in A_s : r + d_r = s$. As a result, we have

$$\sum_{r \in A_s} d_r \geq 0 + 1 + \ldots + (\upsilon_s - 1) = \frac{\upsilon_s(\upsilon_s - 1)}{2}.$$

This gives us the following inequality

$$\begin{aligned}\sum_{s=1}^{t} \frac{\upsilon_s^2 - \upsilon_s}{2} &\leq \sum_{s=1}^{t} \sum_{r \in A_s} d_r \\ &= \sum_{r: r + d_r \leq t} d_r.\end{aligned}$$

On the other hand, $\sum_{s=1}^{t} \sigma_s \geq \sum_{r: r+d_r \leq t} d_r$, since every observation from an action taken at round r with delay d_r counts as outstanding over d_r rounds, i.e., contributes 1 to $\sigma_{r+1}, \ldots, \sigma_{r+d_r}$, and observations that have not arrived by round t contribute only to the left hand side of the inequality. Together with the preceding inequality this completes the proof. □

3.8.1.3 A proof of Lemma 3.4

Proof. We bound $4aA - \overline{Reg}_T$.

$$\begin{aligned}4aA - \overline{Reg}_T &= \sum_{t=1}^{T} \sum_{i \neq i^*} \left(\frac{4a x_{t,i}^{\frac{1}{2}}}{\sqrt{t + \eta_0}} - x_{t,i}\Delta_i\right) \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \frac{4a^2}{(t + \eta_0)\Delta_i} \leq \sum_{i \neq i^*} \frac{4a^2}{\Delta_i} \log(T/\eta_0 + 1) + 1, \qquad (3.14)\end{aligned}$$

where the first inequality uses the AM-GM inequality, by which for any z and y we have $z + y \geq 2\sqrt{zy} \Rightarrow 2\sqrt{zy} - y \leq z$. The second inequality follows by the integral bound on the harmonic series, by which $\sum_{t=1}^{T} 1/(t+\eta_0) \leq \log(T+\eta_0) - \log(\eta_0) + 1$. □

3.8.1.4 Proof of Lemma 3.5

Proof. We have

$$4bB - \overline{Reg}_T = \sum_{t=1}^{T} \sum_{i \neq i^*} x_{t,i} \Delta_i \left(4b(\upsilon_{t+d_t} - 1)\gamma_{t+d_t} - 1\right).$$

We define T_0 to be the first round t with $\gamma_t^{-1} \geq 4b(\upsilon_{max} - 1)$, where $\upsilon_{max} = \max_{s \in [T]}\{\upsilon_s\}$. Then in the summation over time, the rounds with $t + d_t \geq T_0$ provide a negative contribution, since $4b(\upsilon_{t+d_t} - 1)\gamma_{t+d_t} - 1 \leq \frac{4b(\upsilon_{t+d_t}-1)}{4b(\upsilon_{max}-1)} - 1 \leq 0$. Therefore,

$$\begin{aligned} 4bB - \overline{Reg}_T &\leq \sum_{t+d_t < T_0} \sum_{i \neq i^*} x_{t,i} \Delta_i \left(4b(\upsilon_{t+d_t} - 1)\gamma_{t+d_t} - 1\right) \\ &\leq \sum_{t+d_t < T_0} 4b(\upsilon_{t+d_t} - 1)\gamma_{t+d_t} = \sum_{t=1}^{T_0-1} \sum_{s+d_s = t} 4b(\upsilon_t - 1)\gamma_t = \sum_{t=1}^{T_0-1} 4b\upsilon_t(\upsilon_t - 1)\gamma_t, \end{aligned} \tag{3.15}$$

where the second inequality holds because $\sum_{i \neq i^*} x_{t,i}\Delta_i \leq 1$ and $\upsilon_{t+d_t} \geq 1$. For simplicity of notation, we denote $\tilde{\upsilon}_t = \upsilon_t(\upsilon_t - 1)/2$, for which Lemma 3.8 gives us $\sum_{s=1}^{t} \tilde{\upsilon}_t \leq \sum_{s=1}^{t} \sigma_s$. Therefore, we have

$$\begin{aligned} \sum_{t=1}^{T_0-1} 4b\upsilon_t(\upsilon_t - 1)\gamma_t &\leq \sum_{t=1}^{T_0-1} \frac{8b\sqrt{\log K}\,\tilde{\upsilon}_t}{\sqrt{\sum_{s=1}^{t} \tilde{\upsilon}_t}} \\ &\leq 16b\sqrt{(\log K) \sum_{t=1}^{T_0-1} \tilde{\upsilon}_t} \leq 16b\sqrt{(\log K) \sum_{t=1}^{T_0-1} \sigma_t} \leq 16b(\log K)\gamma_{T_0-1}^{-1}, \end{aligned} \tag{3.16}$$

where the second inequality uses integral inequality Lemma 3.7 for $g(x) = \frac{1}{\sqrt{x}}$. Moreover, by the choice of T_0 we have $\gamma_{T_0-1}^{-1} \leq 4b(\upsilon_{max} - 1)$. Combining this with (3.15) and (3.16) gives us $4bB - \overline{Reg}_T \leq 64b^2\upsilon_{max} \log K$. □

3.8.1.5 Proof of Lemma 3.6

Proof. First, we remove i^* from the summation in C by using the following inequality

$$-x_{t,i^*}\log(x_{t,i^*}) \leq (1-x_{t,i^*}) = \sum_{i\neq i^*} x_{t,i},$$

which follows by the fact that $z\log(z)+1-z$ is a decreasing function for $z\in[0,1]$, and the minimum value is zero, therefore, it is non-negative for $z\in[0,1]$. By using this inequality we have

$$\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{-4c\sigma_t x_{t,i}\log(x_{t,i})}{\sqrt{(S_t+\gamma_0)\log K}} \leq 4c\underbrace{\sum_{t=1}^{T}\sum_{i\neq i^*}\frac{-\sigma_t x_{t,i}\log(x_{t,i})}{\sqrt{(S_t+\gamma_0)\log K}}}_{C_1} + 4c\underbrace{\sum_{t=1}^{T}\sum_{i\neq i^*}\frac{\sigma_t x_{t,i}}{\sqrt{(S_t+\gamma_0)\log K}}}_{C_2},$$

where $S_t=\sum_{s=1}^{t}\sigma_s$. We break the expression $4cC-\overline{Reg}_T$, into $4\left(cC_1-\alpha\overline{Reg}_T\right)+4\left(cC_2-\beta\overline{Reg}_T\right)$, where $\alpha+\beta=1/4$.

Controlling $cC_2-\beta\overline{Reg}_T$

Let $\sigma_{max}=\max_{t\in[T]}\{\sigma_t\}$ and let T_i be the first round t when $S_t+\gamma_0 \geq \frac{c^2\sigma_{max}^2}{\beta^2\Delta_i^2\log K}$. Then for all $t\geq T_i$ we have

$$\frac{c\sigma_t x_{t,i}}{\sqrt{(S_t+\gamma_0)\log K}} - \beta x_{t,i}\Delta_i \leq 0.$$

Therefore, rounds after T_i provide negative contribution to the summation, and we have

$$\begin{aligned}
cC_2-\beta\overline{Reg}_T &\leq \beta\sum_{i\neq i^*}\sum_{t=1}^{T_i-1}x_{t,i}\left(\frac{c\sigma_t}{\beta\sqrt{(S_t+\gamma_0)\log K}}-\Delta_i\right)\\
&\leq \sum_{i\neq i^*}\sum_{t=1}^{T_i-1}\frac{c\sigma_t}{\sqrt{(S_t+\gamma_0)\log K}}\\
&\leq \sum_{i\neq i^*}\frac{2c(\sqrt{S_{T_i-1}+\gamma_0}-\sqrt{\gamma_0})}{\sqrt{\log K}}\\
&\leq \sum_{i\neq i^*}\frac{2c^2\sigma_{max}}{\beta\Delta_i\log K}, \qquad (3.17)
\end{aligned}$$

where the third inequality uses Lemma 3.7 for $g(x)=\frac{1}{\sqrt{x}}$ and the last inequality follows by the choice of T_i, which gives $S_{T_i-1}+\gamma_0 \leq \frac{c^2\sigma_{max}^2}{\beta^2\Delta_i^2\log K}$.

Controlling $cC_1 - \alpha\overline{Reg}_T$
For $cC_1 - \alpha\overline{Reg}_T$, let $b_t = \frac{c\sigma_t}{\alpha\sqrt{(S_t+\gamma_0)\log K}}$, then

$$\begin{aligned} cC_1 - \alpha\overline{Reg}_T &= \alpha \sum_{t=1}^{T} \sum_{i \neq i^*} \left(-b_t x_{t,i} \log(x_{t,i}) - \Delta_i x_{t,i}\right) \\ &\leq \alpha \sum_{t=1}^{T} \sum_{i \neq i^*} \max_{z \in [0,1]} \left\{-b_t z \log(z) - \Delta_i z\right\}. \end{aligned}$$

The function $g(z) = -b_t z \log(z) - \Delta_i z$ is a concave function for $z \in [0,1]$ and the maximum occurs when the derivative is zero. So we must have $-b_t \log(z) - b_t - \Delta_i = 0 \Rightarrow z = e^{-\frac{\Delta_i}{b_t}-1}$, and by substitution $\max_{z\in[0,1]} g(z) = b_t e^{-\frac{\Delta_i}{b_t}-1}$. Therefore,

$$\begin{aligned} cC_1 - \alpha\overline{Reg}_T &\leq \alpha \sum_{t=1}^{T} \sum_{i \neq i^*} b_t e^{-\frac{\Delta_i}{b_t}-1} \\ &= \sum_{i \neq i^*} \sum_{t=1}^{T} \frac{c\sigma_t}{\sqrt{(S_t+\gamma_0)\log K}} \exp\left(-\frac{\alpha\Delta_i\sqrt{(S_t+\gamma_0)\log K}}{c\sigma_t} - 1\right) \\ &\leq \sum_{i \neq i^*} \sum_{t=1}^{T} \sigma_t \times \frac{c}{\sqrt{(S_t+\gamma_0)\log K}} \exp\left(-\frac{\alpha\Delta_i\sqrt{(S_t+\gamma_0)\log K}}{c\sigma_{max}} - 1\right), \end{aligned}$$

where $\sigma_{max} = \max_{t\in[T]}\{\sigma_t\}$. Let $g_i(x) = \frac{c}{\sqrt{x\log K}} \exp\left(-\frac{\alpha\Delta_i\sqrt{x\log K}}{c\sigma_{max}} - 1\right)$, then for each i we need to upper bound $\sum_{t=1}^{T} \sigma_t g_i(S_t+\gamma_0)$, which by Lemma 3.7 can be upper bounded by $\int_{\gamma_0}^{S_T+\gamma_0} g_i(x)dx$, because g is nonincreasing. On the other hand, for any $\delta, a \geq 0$, we have $\int \frac{a}{\sqrt{x}} \exp(-\frac{\delta\sqrt{x}}{a} - 1)dx = -\frac{2a^2}{\delta}\exp(-\frac{\delta\sqrt{x}}{a} - 1)$. So, using

the closed form of $\int g_i(x)dx$ with $\delta = \frac{\alpha\Delta_i}{\sigma_{max}}, a = \frac{c}{\sqrt{\log K}}$, we have

$$\begin{aligned} cC_1 - \alpha\overline{Reg}_T &\leq \sum_{i\neq i^*} \int_{\gamma_0}^{S_T+\gamma_0} g_i(x)dx \\ &= \sum_{i\neq i^*} \frac{-2c^2\sigma_{max}}{\alpha\Delta_i \log K} \exp\left(-\frac{\alpha\Delta_i\sqrt{x\log K}}{c\sigma_{max}} - 1\right)\Bigg|_{x=\gamma_0}^{x=S_T+\gamma_0} \\ &= \frac{2c^2\sigma_{max}\left(\exp\left(-\frac{\alpha\Delta_i\sqrt{\gamma_0\log K}}{c\sigma_{max}} - 1\right) - \exp\left(-\frac{\alpha\Delta_i\sqrt{(S_T+\gamma_0)\log K}}{c\sigma_{max}} - 1\right)\right)}{\alpha\Delta_i \log K} \\ &\leq \sum_{i\neq i^*} \frac{2c^2\sigma_{max}}{\alpha\Delta_i \log K}. \end{aligned} \tag{3.18}$$

Taking together (3.17) and (3.18) gives us

$$\begin{aligned} 4cC - \overline{Reg}_T \leq \sum_{i\neq i^*} \frac{8c^2\sigma_{max}}{\Delta_i \log K}\left(\frac{1}{\beta} + \frac{1}{\alpha}\right) &= \sum_{i\neq i^*} \frac{8c^2\sigma_{max}}{\Delta_i \log K}\left(\frac{1}{1/4-\alpha} + \frac{1}{\alpha}\right) \\ &\leq \sum_{i\neq i^*} \frac{128c^2\sigma_{max}}{\Delta_i \log K}, \end{aligned} \tag{3.19}$$

where the second inequality uses $\alpha = \frac{1}{8}$.

□

3.8.1.6 Proof of the stability lemma

The lemma has two parts, the first part is the general bound for the stability term and the second is a special case of that bound where we set α to a specific value to get the desirable bound.

Before starting the proof we provide one fact and one lemma that help us in the proof of the stability lemma. We recall that our regularization function is $F_t(x) = \sum_{i=1}^K f_t(x)$, where $f_t(x) = -2\eta_t^{-1}\sqrt{x} + \gamma_t^{-1}x(\log x - 1)$.

Fact 3.3 ((Zimmert and Seldin, 2020)). *$f_t^{*'}(x)$ is a convex monotonically increasing function.*

Proof. The proof is available in Section 7.3 of the supplementary material of Zimmert and Seldin (2020). □

Lemma 3.9. *Let $D_F(x,y) = F(x) - F(y) - \langle x-y, \nabla F(y)\rangle$ be the Bergman divergence of a function F. Then for any $x \in \textbf{dom}(f_t)$, and any ℓ such that $\ell \geq -\gamma_t^{-1}$:*

$$D_{f_t^*}(f_t^{'}(x) - \ell, f_t^{'}(x)) \leq \frac{\ell^2}{2f_t^{''}(ex)}.$$

Moreover, it is easy to see $(f_t^{''}(ex))^{-1} \leq 4(f_t^{''}(x))^{-1}$, which implies $D_{f_t^*}(f_t^{'}(x) - \ell, f_t^{'}(x)) \leq \frac{2\ell^2}{f_t^{''}(x)}$.

Proof. By Taylor's theorem there exists $\tilde{x} \in \left[f_t^{*'}(f_t^{'}(x) - \ell), f_t^{*'}(f_t^{'}(x))\right]$, such that

$$D_{f_t^*}(f_t^{'}(x) - \ell, f_t^{'}(x)) = \frac{1}{2}\ell^2 f_t^{*''}(f_t^{'}(\tilde{x})) = \frac{1}{2}\ell^2 f_t^{''}(\tilde{x})^{-1},$$

where the second equality is a property of the convex conjugate operation. We have two cases for ℓ:

1. If $\ell \geq 0$, then based on Fact 3.3 we know that $f_t^{*'}$ is increasing, so $\tilde{x} \leq x$. On the other hand, $f^{''}(x)^{-1}$ is increasing, so $f_t^{''}(\tilde{x})^{-1} \leq f_t^{''}(x)^{-1} \leq f_t^{''}(ex)^{-1}$.

2. If $\ell < 0$, then $\tilde{x} \in \left[f_t^{*'}(f_t^{'}(x)), f_t^{*'}(f_t^{'}(x) - \ell)\right]$. We show that $f_t^{*'}(f_t^{'}(x) - \ell) \leq ex$, which by the choice of $\tilde{x}$ implies $\tilde{x} \leq ex$, and consequently, like in the other case, we end up having $f_t^{''}(\tilde{x})^{-1} \leq f_t^{''}(ex)^{-1}$.

   Since $f^{*'}$ is increasing and $ex = f^{*'}(f^{'}(ex))$, it suffices to prove that $f^{'}(ex) \geq f^{'}(x) - \ell$, or, equivalently, $f^{'}(ex) - f^{'}(x) \geq -\ell$. So

$$\begin{aligned} f^{'}(ex) - f^{'}(x) &= \left(-\eta_t^{-1}(ex)^{-1/2} + \gamma_t^{-1}\log(ex)\right) - \left(-\eta_t^{-1}x^{-1/2} + \gamma_t^{-1}\log(x)\right) \\ &= \eta_t^{-1}x^{-1/2}\left(1 - \frac{1}{\sqrt{2}}\right) + \gamma_t^{-1} \geq \gamma_t^{-1} \geq -\ell \end{aligned}$$

□

Proof of the First Part of the Stability Lemma. We have $x_t = \arg\min_{x\in\Delta^{K-1}}\langle \hat{L}_t^{obs}, x\rangle + F_t(x)$, so by the KKT conditions there exists $c_0 \in \mathbb{R}$, such that $-\hat{L}_t^{obs} = \nabla F_t(x_t) - c_0\mathbf{1}_K$. On the other hand, $\bar{F}_t(-L + c\mathbf{1}_K) = \bar{F}_t(-L) + c$ for any $c \in \mathbb{R}$ and $L \in \mathbb{R}^K$ and the equality holds iff $c = 0$. Therefore, using these

two facts we can rewrite the stability term as

$$\sum_{t=1}^{T}\langle x_t, \hat{\ell}_t^{obs}\rangle + \bar{F}_t^*(-\hat{L}_{t+1}^{obs}) - \bar{F}_t^*(-\hat{L}_t^{obs}) \tag{3.20}$$

$$= \sum_{t=1}^{T}\langle x_t, \hat{\ell}_t^{obs} - \alpha_t \mathbf{1}_K\rangle + \bar{F}_t^*(-\hat{L}_{t+1}^{obs} + (\alpha_t + c_0)\mathbf{1}_K) - \bar{F}_t^*(-\hat{L}_t^{obs} + c_0\mathbf{1}_K)$$

$$= \sum_{t=1}^{T}\langle x_t, \hat{\ell}_t^{obs} - \alpha_t \mathbf{1}_K\rangle + \bar{F}_t^*(\nabla F_t(x_t) - (\hat{\ell}_t^{obs} - \alpha_t\mathbf{1}_K)) - \bar{F}_t^*(\nabla F_t(x_t))$$

$$\leq \sum_{t=1}^{T}\langle x_t, \hat{\ell}_t^{obs} - \alpha_t \mathbf{1}_K\rangle + F_t^*(\nabla F_t(x_t) - (\hat{\ell}_t^{obs} - \alpha_t\mathbf{1}_K)) - F_t^*(\nabla F_t(x_t))$$

$$= \sum_{i=1}^{K} D_{f_t^*}\left(f_t^{'}(x_{t,i}) - (\hat{\ell}_{t,i}^{obs} - \alpha_t), f_t^{'}(x_{t,i})\right), \tag{3.21}$$

where the inequality holds because $\bar{F}_t^*(L) \leq F_t^*(L)$ for all $L \in \mathbb{R}^K$ and $\bar{F}_t^*(\nabla F_t(x)) = F_t^*(\nabla F_t(x))$ for all $x \in \mathbb{R}^K$. Hence, since $\alpha_t \leq \gamma_t^{-1}$, we have $\hat{\ell}_{t,i}^{obs} - \alpha_t \geq -\alpha_t \geq -\gamma_t^{-1}$. This implies that we can apply Lemma 3.9 to get the following bound for (3.21)

$$stability \leq \sum_{i=1}^{K} 2f_t^{''}(x_{t,i})^{-1}(\hat{\ell}_{t,i}^{obs} - \alpha_t)^2.$$

□

Proof of the Second Part of the Stability Lemma. First, we must check whether $\alpha_t = \frac{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}\tilde{\ell}_{t,j}}{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}}$ satisfies $\alpha_t \leq \gamma_t^{-1}$ or not:

$$\begin{aligned}
\alpha_t &= \frac{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}\tilde{\ell}_{t,j}}{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}} \\
&= \frac{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}\sum_{s\in A_t}\hat{\ell}_{s,j}}{\sum_{j=1}^{K} f^{''}(x_{t,j})^{-1}} \\
&\leq 8|A_t|(K-1)^{\frac{1}{3}} \leq 8d_{max}(K-1)^{\frac{1}{3}} \leq \gamma_t^{-1},
\end{aligned}$$

where the first inequality uses Lemma 3.10. To simplify the analysis, for all i let $z_i = f_t^{''}(x_{t,i})^{-1}$, then by substitution of the value of α_t in the stability expression we

have

$$\begin{aligned}
\sum_{i=1}^{K} z_i(\tilde{\ell}_{t,i} - \alpha_t)^2 &= \sum_{i=1}^{K} z_i \tilde{\ell}_{t,i}^2 - 2\sum_{i=1}^{K} z_i \tilde{\ell}_{t,i}\alpha_t + \sum_{i=1}^{K} z_i \alpha_t^2 \\
&= \sum_{i=1}^{K} z_i \tilde{\ell}_{t,i}^2 - \frac{(\sum_{i=1}^{K} z_i \tilde{\ell}_{t,i})^2}{\sum_{i=1}^{K} z_i} \\
&= \sum_{i=1}^{K} z_i \tilde{\ell}_{t,i}^2 - \frac{\sum_{i=1}^{K} z_i^2 \tilde{\ell}_{t,i}^2}{\sum_{i=1}^{K} z_i} - \frac{\sum_{i,j,i\neq j} z_i z_j \tilde{\ell}_{t,i}\tilde{\ell}_{t,j}}{\sum_{i=1}^{K} z_i} \\
&= \sum_{i=1}^{K} \left(z_i - \frac{z_i^2}{\sum_{j=1}^{K} z_j} \right) \left(\sum_{s\in A_t} \hat{\ell}_{s,i} \right)^2 - \frac{\sum_{i,j,i\neq j} z_i z_j \left(\sum_{r,s\in A_t} \hat{\ell}_{r,i}\hat{\ell}_{s,j} \right)}{\sum_{i=1}^{K} z_i} \\
&= \sum_{i=1}^{K} \left(z_i - \frac{z_i^2}{\sum_{j=1}^{K} z_j} \right) \left(\sum_{s\in A_t} \hat{\ell}_{s,i}^2 \right) \qquad (3.22) \\
&\quad + \sum_{i=1}^{K} \left(z_i - \frac{z_i^2}{\sum_{j=1}^{K} z_j} \right) \left(\sum_{r,s\in A_t, r\neq s} \hat{\ell}_{r,i}\hat{\ell}_{s,i} \right) - \frac{\sum_{i,j,i\neq j} z_i z_j \left(\sum_{r,s\in A_t} \hat{\ell}_{s,i}\hat{\ell}_{r,j} \right)}{\sum_{i=1}^{K} z_i}. \qquad (3.23)
\end{aligned}$$

We call the term in line (3.22) Stab1 and the two terms in line (3.23) Stab2. We

first bound the expectation of Stab1.

$$\begin{aligned}
\mathbb{E}[\text{Stab1}] &= \mathbb{E}\left[\sum_{i=1}^{K}\left(z_i - \frac{z_i^2}{\sum_{i=1}^{K} z_i}\right)\left(\sum_{s\in A_t}\hat{\ell}_{s,i}^2\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\left(z_i - \frac{z_i^2}{\sum_{i=1}^{K} z_i}\right)\left(\sum_{s\in A_t}\mathbb{E}_s[\hat{\ell}_{s,i}^2]\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{K}\left(z_i - \frac{z_i^2}{\sum_{i=1}^{K} z_i}\right)\left(\sum_{s\in A_t}\ell_{s,i}^2 x_{s,i}^{-1}\right)\right] \\
&\le \sum_{s\in A_t}\mathbb{E}\left[\sum_{i=1}^{K} z_i x_{s,i}^{-1} - \frac{\sum_{i=1}^{K} z_i^2 x_{s,i}^{-1}}{\sum_{i=1}^{K} z_i}\right] \\
&\le \sum_{s\in A_t}\mathbb{E}\left[\sum_{i=1}^{K} z_i x_{s,i}^{-1}(1-x_{s,i})\right] \\
&\le \sum_{s\in A_t}\mathbb{E}\left[\sum_{i=1}^{K} 2\eta_t x_{t,i}^{3/2} x_{s,i}^{-1}(1-x_{s,i})\right], \qquad (3.24)
\end{aligned}$$

where the first inequality bounds losses by one and changes the order of summations, the second inequality uses Cauchy-Schwarz inequality $\sum_{i=1}^{K} z_i^2 x_{s,i}^{-1} = \left(\sum_{i=1}^{K} z_i^2 x_{s,i}^{-1}\right)\underbrace{\left(\sum_{i=1}^{K} x_{s,i}\right)}_{=1} \ge \left(\sum_{i=1}^{K} z_i\right)^2$, and the last inequality uses the fact that $z_i = f_t''(x_{t,i})^{-1} \le 2\eta_t x_{t,i}^{3/2}$.

For Stab2 we have

$$
\begin{aligned}
\mathbb{E}[\text{Stab2}] &= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{K} z_i}\left(\sum_{i=1}^{K}\sum_{r,s\in A_t, r\neq s}\sum_{j\neq i} z_i z_j \hat{\ell}_{r,i}\hat{\ell}_{s,i} - \sum_{i,j,i\neq j}\sum_{r,s\in A_t} z_i z_j \hat{\ell}_{s,i}\hat{\ell}_{r,j}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{K} z_i}\left(\sum_{i=1}^{K}\sum_{r,s\in A_t, r\neq s}\sum_{j\neq i} z_i z_j \mu_i^2 - \sum_{i,j,i\neq j}\sum_{r,s\in A_t} z_i z_j \mu_i \mu_j\right)\right] \\
&= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{K} z_i}\left(\upsilon_t(\upsilon_t-1)\sum_{i=1}^{K}\sum_{j\neq i} z_i z_j \mu_i^2 - \upsilon_t^2 \sum_{i,j,i\neq j} z_i z_j \mu_i \mu_j\right)\right] \qquad (3.25) \\
&\leq \mathbb{E}\left[\frac{\upsilon_t(\upsilon_t-1)}{\sum_{i=1}^{K} z_i}\left(\sum_{i=1}^{K} z_i(\sum_{j=1}^{K} z_j)\mu_i^2 - \sum_{i=1}^{K} z_i^2\mu_i^2 - \sum_{i,j,i\neq j} z_i z_j \mu_i \mu_j\right)\right] \\
&= \mathbb{E}\left[\frac{\upsilon_t(\upsilon_t-1)}{\sum_{i=1}^{K} z_i}\left((\sum_{i=1}^{K} z_i\mu_i^2)(\sum_{i=1}^{K} z_i) - (\sum_{i=1}^{K} z_i\mu_i)^2\right)\right] \\
&\leq \mathbb{E}\left[\frac{\upsilon_t(\upsilon_t-1)}{\sum_{i=1}^{K} z_i}\left((\sum_{i=1}^{K} z_i\mu_i^2)(\sum_{i=1}^{K} z_i) - (\sum_{i=1}^{K} z_i)^2\mu_{i^*}^2\right)\right] \\
&= \mathbb{E}\left[\upsilon_t(\upsilon_t-1)(\sum_{i=1}^{K} z_i\mu_i^2 - \sum_{i=1}^{K} z_i\mu_{i^*}^2)\right] \\
&\leq \mathbb{E}\left[\upsilon_t(\upsilon_t-1)(\sum_{i\neq i^*} 2z_i\Delta_i)\right] \\
&\leq \mathbb{E}\left[\sum_{i\neq i^*} 2\upsilon_t(\upsilon_t-1)\gamma_t x_{t,i}\Delta_i\right], \qquad (3.26)
\end{aligned}
$$

where the second equality follows by the fact that for all $s \in A_t$, x_s has no impact on x_t, and for all different elements of A_t, such as $r, s \in A_t$ and $r < s$, x_r has no impact on x_s. Regarding the inequalities, the first one follows by $\upsilon_t^2 \geq \upsilon_t(\upsilon_t - 1)$, the second one holds because for all i we have $\mu_i^* \leq \mu_i$, the third inequality follows by $\mu_i + \mu_{i^*} \leq 2$ and $\mu_i - \mu_{i^*} = \Delta_i$, and the last one substitutes $z_i = f''(x_{t,i})^{-1} \leq \gamma_t x_{t,i}$.

Combining (3.24) and (3.26) completes the proof. □

3.8.2 Proof of the Key Lemma

3.8.2.1 Auxiliary results for the proof of the key lemma

First, we provide two facts and a lemma, which are needed for the proof of the key lemma. We recall that $f_t(x) = -2\eta_t^{-1}\sqrt{x} + \gamma_t^{-1}x(\log x - 1)$.

Fact 3.4. *$f_t^{'}(x)$ is a concave function.*

Proof. $f_t^{'}(x) = -\eta^{-1}x^{-1/2} + \gamma_t^{-1}\log x$, so the second derivative is $-\frac{3}{4}\eta^{-1}x^{-5/2} - \gamma_t^{-1}x^{-2} \leq 0$. □

Fact 3.5. *$f_t^{''}(x)^{-1}$ is a convex function.*

Proof. Let $g(x) = f_t^{''}(x)^{-1} = (\frac{\eta_t^{-1}x^{-3/2}}{2} + \gamma_t^{-1}x^{-1})^{-1}$, then the second derivative of $g(x)$ is

$$g^{''}(x) = \frac{\eta_t\gamma_t^2 \cdot \left(2\eta_t x^{\frac{7}{2}} + 3\gamma_t x^3\right)}{2\sqrt{x}\left(2\eta_t x^{\frac{3}{2}} + \gamma_t x\right)^3},$$

which is positive. □

Lemma 3.10. *Fix t and s where $t \geq s$, and assume that there exists α, such that $x_{t,i} \leq \alpha x_{s,i}$ for all $i \in [K]$, and let $f(x) = \left(-2\eta_t^{-1}\sqrt{x} + \gamma_t^{-1}x(\log x - 1)\right)$, then we have the following inequality*

$$\frac{\sum_{j=1}^K f''(x_{t,j})^{-1}\hat{\ell}_{s,j}}{\sum_{j=1}^K f''(x_{t,j})^{-1}} \leq 2\alpha(K-1)^{\frac{1}{3}}.$$

Proof for Lemma 3.10. We begin the proof as the following

$$\begin{aligned}
\frac{\sum_{i=1}^K f''(x_{t,i})^{-1}\hat{\ell}_{s,i}}{\sum_{i=1}^K f''(x_{t,i})^{-1}} &= \frac{f''(x_{t,i_s})^{-1}x_{s,i_s}^{-1}\ell_{s,i_s}}{\sum_{i=1}^K f''(x_{t,i})^{-1}} \\
&\leq \frac{f''(x_{t,i_s})^{-1}x_{t,i_s}^{-1}\left(x_{t,i_s}/x_{s,i_s}\right)}{\sum_{i=1}^K f''(x_{t,i})^{-1}} \\
&\leq \frac{f''(x_{t,i_s})^{-1}\alpha x_{t,i_s}^{-1}}{\sum_{i=1}^K f''(x_{t,i})^{-1}} \\
&\leq \frac{\alpha f''(x_{t,i_s})^{-1}x_{t,i_s}^{-1}}{(K-1)f''\left(\frac{1-x_{t,i_s}}{K-1}\right)^{-1} + f''(x_{t,i_s})^{-1}} \quad \text{Define } z := x_{t,i_s} \\
&= \frac{\alpha\left(\eta_t^{-1}z^{-3/2} + 2\gamma_t^{-1}z^{-1}\right)^{-1}z^{-1}}{(K-1)\left(\eta_t^{-1}(\frac{1-z}{K-1})^{-3/2} + 2\gamma_t^{-1}(\frac{1-z}{K-1})^{-1}\right)^{-1} + \left(\eta_t^{-1}z^{-3/2} + 2\gamma_t^{-1}z^{-1}\right)^{-1}} \\
&= \alpha\left((1-z)\frac{\eta_t^{-1}z^{-1/2} + 2\gamma_t^{-1}}{\eta_t^{-1}\sqrt{K-1}(1-z)^{-1/2} + 2\gamma_t^{-1}} + z\right)^{-1}, \qquad (3.27)
\end{aligned}$$

where the first inequality follows by $\ell_{s,i_s} \leq 1$, the second one uses the assumption of the lemma that $x_{t,i} \leq \alpha x_{s,i}$, and the third inequality is due to convexity of $f''(x)^{-1}$ from Fact 3.5. We consider two cases for z: $z < \frac{1}{K}$ and $z \geq \frac{1}{K}$.

a) $z \leq \frac{1}{K}$: This case implies

$$\begin{aligned}
\frac{1-z}{z} = \frac{1}{z} - 1 \geq K-1 \Rightarrow (1-z)^{-1/2}\sqrt{K-1} &\leq z^{-1/2} \\
\Rightarrow 1 &\leq \frac{\eta_t^{-1}z^{-1/2} + 2\gamma_t^{-1}}{\eta_t^{-1}\sqrt{K-1}(1-z)^{-1/2} + 2\gamma_t^{-1}}. \qquad (3.28)
\end{aligned}$$

Plugging (3.28) into (3.27) gives

$$\frac{\sum_{i=1}^K f''(x_{t,i})^{-1}\hat{\ell}_{s,i}}{\sum_{i=1}^K f''(x_{t,i})^{-1}} \leq \alpha\left(1 - z + z\right)^{-1} = \alpha.$$

b) $z \geq \frac{1}{K}$: Similar to the previous case, $z \geq \frac{1}{K}$ implies $\eta_t^{-1}z^{-1/2} \leq \eta_t^{-1}\sqrt{K-1}(1-z)^{-1/2}$, so the minimum of $\frac{\eta_t^{-1}z^{-1/2}+2\gamma_t^{-1}}{\eta_t^{-1}\sqrt{K-1}(1-z)^{-1/2}+2\gamma_t^{-1}}$ occurs when $2\gamma_t^{-1} = 0$. Substitution of $2\gamma_t^{-1} = 0$ in (3.27) gives

$$\frac{\sum_{i=1}^K f''(x_{t,i})^{-1}\hat{\ell}_{s,i}}{\sum_{i=1}^K f''(x_{t,i})^{-1}} \leq \alpha\left((1-z)^{3/2}z^{-1/2}(K-1)^{-1/2} + z\right)^{-1}. \qquad (3.29)$$

Here we have the following two subcases

b1) $z \geq \frac{1}{(K-1)^{1/3}+1}$: This gives

$$\alpha\left((1-z)^{3/2}z^{-1/2}(K-1)^{-1/2}+z\right)^{-1} \leq \alpha z^{-1} \leq \alpha\left((K-1)^{1/3}+1\right) \leq 2\alpha(K-1)^{1/3}.$$

b2) $z \leq \frac{1}{(K-1)^{1/3}+1}$: This implies $(1-z) \geq \frac{(K-1)^{1/3}}{(K-1)^{1/3}+1} \geq \frac{1}{2}$ and we can use it in (3.29) in the following way

$$\begin{aligned}
&\alpha\left((1-z)^{3/2}z^{-1/2}(K-1)^{-1/2}+z\right)^{-1} \\
&\leq \alpha\left(\frac{z^{-1/2}(K-1)^{-1/2}}{\sqrt{8}}+z\right)^{-1} \\
&= \alpha\left(\frac{z^{-1/2}(K-1)^{-1/2}}{2\sqrt{8}}+\frac{z^{-1/2}(K-1)^{-1/2}}{2\sqrt{8}}+z\right)^{-1} \\
&\leq \frac{\alpha}{3}\left(\frac{(K-1)^{-1}}{32}\right)^{-1/3} \leq 2\alpha(K-1)^{1/3},
\end{aligned}$$

where the second inequality is by the AM-GM inequality.

Combining the results for all cases and setting $\alpha = 4$ we obtain the upper bound $8(K-1)^{1/3}$. □

3.8.2.2 Proof of the key lemma

Proof of Lemma 3.2. To show $x_{t,i} \leq 2x_{s,i}$ for all i we do induction on *valid* pairs (t,s), where we call a pair (t,s) valid if $s \leq t$ and $t-s \leq d_{max}$. The induction step for (t,s) uses the induction assumption for all valid pairs (t',s'), such that $s',t' < t$, and all valid pairs (t',s'), such that $t' = t$ and $s < s' \leq t$. Thus, the induction base would be all the pairs of (t',t') for all $t' \in [T]$, for which the statement $x_{t',i} \leq 2x_{t',i}$ trivially holds. Hence, it suffices to prove the induction step for the valid pair (t,s).

As we mentioned in the proof sketch, we have $x_t = \bar{F}_t^*(-\hat{L}_t^{obs})$ and $x_s = \bar{F}_s^*(-\hat{L}_s^{obs})$, and we introduce $\tilde{x} = \bar{F}_s^*(-\hat{L}_t^{obs})$ as an auxiliary variable to bridge from x_t and x_s. We bridge from x_t to x_s via $\tilde{x}$ in the following way.

Deviation Induced by the Loss Shift: This step controls the drift when we fix the regularization (more precisely, the learning rates) and shift the cumulative loss. We prove the following inequality:

$$\tilde{x}_i \leq \frac{3}{2}x_{s,i}.$$

Note that this step uses the induction assumption for $(s, s-d_r)$ for all $r < s : r+d_r = s$.

Deviation Induced by the Change of Regularizer: In this step we bound the drift when the cumulative loss vector is fixed and we change the regularizer. We show that

$$x_{t,i} \leq \frac{4}{3}\tilde{x}_i.$$

Deviation induced by the change of regularizer

The regularizer at any round r is $F_r(x) = \sum_{i=1}^K f_r(x_i) = \sum_{i=1}^K \left(-2\eta_r^{-1}\sqrt{x_i} + \gamma_r^{-1}x_i(\log x_i - 1)\right)$. Since $x_t = \nabla \bar{F}_t^*(-\hat{L}_t^{obs})$ and $\tilde{x} = \nabla \bar{F}_s^*(-\hat{L}_t^{obs})$, by the KKT conditions $\exists \mu, \tilde{\mu}$ s.t. $\forall i$:

$$f_s^{'}(\tilde{x}_i) = -L_{t,i}^{obs} + \tilde{\mu},$$
$$f_t^{'}(x_{t,i}) = -L_{t,i}^{obs} + \mu.$$

We also know that $\exists j : \tilde{x}_j \geq x_{t,j}$ which leads to

$$-L_{t,j}^{obs} + \mu = f_t^{'}(x_{t,j}) \leq f_s^{'}(x_{t,j}) \leq f_s^{'}(\tilde{x}_j) = -L_{t,j}^{obs} + \tilde{\mu},$$

where the first inequality holds because the learning rates are decreasing, and the second inequality is due to the fact that $f_s^{'}(x)$ is increasing. This implies that $\mu \leq \tilde{\mu}$, which gives us the following inequality for all i:

$$f_t^{'}(x_{t,i}) = -\frac{1}{\eta_t\sqrt{x_{t,i}}} + \frac{\log(x_{t,i})}{\gamma_t} \leq -\frac{1}{\eta_s\sqrt{\tilde{x}_i}} + \frac{\log(\tilde{x}_i)}{\gamma_s} = f_s^{'}(\tilde{x}_i).$$

Define $\alpha = x_{t,i}/\tilde{x}_i$. Using the above inequality we have

$$\begin{aligned}
\frac{1}{\eta_s\sqrt{\tilde{x}_i}} - \frac{\log(\tilde{x}_i)}{\gamma_s} &\leq \frac{1}{\eta_t\sqrt{\alpha\tilde{x}_i}} - \frac{\log(\tilde{x}_i)}{\gamma_t} - \frac{\log(\alpha)}{\gamma_t} \quad \text{(multiply by } \eta_t\sqrt{\tilde{x}_i} \text{ and rearrange)} \\
\Rightarrow \frac{1}{\sqrt{\alpha}} &\geq \frac{\eta_t}{\eta_s} + 2\sqrt{\tilde{x}_i}\log(\sqrt{\tilde{x}_i})\left(\frac{\eta_t}{\gamma_t} - \frac{\eta_t}{\gamma_s}\right) + \log(\alpha)\frac{\eta_t}{\gamma_t}\sqrt{\tilde{x}_i} \\
&\geq \frac{\eta_t}{\eta_s} + \min_{0\leq z\leq 1}\left\{2z\log(z)\left(\frac{\eta_t}{\gamma_t} - \frac{\eta_t}{\gamma_s}\right) + \log(\alpha)\frac{\eta_t}{\gamma_t}z\right\} \\
&\overset{(a)}{=} \frac{\eta_t}{\eta_s} - \frac{2}{e}\left(\frac{\eta_t}{\gamma_t} - \frac{\eta_t}{\gamma_s}\right)\left(\frac{1}{\sqrt{\alpha}}\right)^{\frac{\gamma_t^{-1}}{\gamma_t^{-1}-\gamma_s^{-1}}} \\
&\overset{(b)}{\geq} \frac{\eta_t}{\eta_s} - \left(\frac{\eta_t}{\gamma_t} - \frac{\eta_t}{\gamma_s}\right)\frac{1}{\sqrt{\alpha}},
\end{aligned}$$

where (a) holds because the minimized function is convex and equating the first derivative to zero gives $z = \left(\frac{1}{\sqrt{\alpha}}\right)^{\frac{\gamma_t^{-1}}{\gamma_t^{-1}-\gamma_s^{-1}}}$, and (b) follows by $\frac{\gamma_t^{-1}}{\gamma_t^{-1}-\gamma_s^{-1}} \geq 1$ and $e \geq 2$. Rearranging the above result gives

$$\alpha \leq \left(\frac{\eta_s}{\gamma_t} - \frac{\eta_s}{\gamma_s} + \frac{\eta_s}{\eta_t}\right)^2 = \left(\eta_s(\gamma_t^{-1} - \gamma_s^{-1}) + \frac{\eta_s}{\eta_t}\right)^2. \tag{3.30}$$

Now we need to substitute the closed form of learning rates to obtain an upper bound for α. As a reminder, the learning rates are

$$\gamma_s^{-1} = \frac{1}{\sqrt{\log K}}\sqrt{\sum_{r=1}^{s}\sigma_r + \gamma_0},\ \eta_s^{-1} = \sqrt{s+\eta_0},$$
$$\gamma_t^{-1} = \frac{1}{\sqrt{\log K}}\sqrt{\sum_{r=1}^{s+d}\sigma_r + \gamma_0},\ \eta_t^{-1} = \sqrt{s+d+\eta_0},$$

where $d = t - s$, $\eta_0 = 10d_{max} + d_{max}^2/\left(K^{1/3}\log(K)\right)^2$, and $\gamma_0 = 24^2 d_{max}^2 K^{2/3}\log(K)$. Therefore, in (3.30) we have

$$\begin{aligned}
\eta_s\left(\gamma_t^{-1} - \gamma_s^{-1}\right) &\leq \eta_s \frac{\sum_{r=s+1}^{s+d}\sigma_r}{\sqrt{\log(K)\left(\sum_{r=1}^{s+d}\sigma_r + \gamma_0\right)}} \\
&\leq \eta_s \frac{\sum_{r=s+1}^{s+d}\sigma_r}{\sqrt{\log(K)\gamma_0}} \\
&\leq \frac{d_{max}^2}{\sqrt{\log(K)\gamma_0\eta_0}} \leq \frac{d_{max}^2}{\sqrt{24^2 d_{max}^4}} = \frac{1}{24},
\end{aligned} \tag{3.31}$$

where the third inequality follows by $d, \sigma_r \leq d_{max}$ for all r and $\eta_s \leq \frac{1}{\sqrt{\eta_0}}$, and the last inequality holds because $\eta_0 \geq 16 d_{max}^2/K^{2/3}$. On the other hand, for $\frac{\eta_s}{\eta_t}$ in (3.30) we have

$$\begin{aligned}
\frac{\eta_s}{\eta_t} = \sqrt{\frac{s+d+\eta_0}{s+\eta_0}} &= \sqrt{1 + \frac{d}{s+\eta_0}} \\
&\leq \sqrt{1 + \frac{d}{10 d_{max}}} \\
&\leq \sqrt{1 + \frac{d_{max}}{10 d_{max}}} = \sqrt{\frac{11}{10}},
\end{aligned} \tag{3.32}$$

where the first and the second inequalities hold because $\eta_0 \geq 10d_{max}$ and $d \leq d_{max}$, respectively.
Plugging (3.31) and (3.32) into (3.30) gives us the following bound for α:

$$\alpha \leq \left(\sqrt{\frac{11}{10}} + \frac{1}{24}\right)^2 \leq \frac{4}{3}. \tag{3.33}$$

Deviation Induced by the Loss Shift

We have $x_s = \nabla \bar{F}_s^*(-L_s^{obs})$ and $\tilde{x} = \nabla \bar{F}_s^*(-L_t^{obs})$. Since they both share the same regularizer $F_s(x) = \sum_{i=1}^K f_s(x_i)$, to simplify the notation we drop s and use $f(x)$ to refer to $f_s(x)$. By the KKT conditions $\exists \mu, \tilde{\mu}$ s.t. $\forall i$:

$$\begin{aligned} f^{'}(x_{s,i}) &= -L_{s,i}^{obs} + \mu, \\ f^{'}(\tilde{x}_i) &= -L_{t,i}^{obs} + \tilde{\mu}. \end{aligned}$$

Let $\tilde{\ell} = L_t^{obs} - L_s^{obs}$, then by the concavity of $f^{'}(x)$ from Fact 3.4, we have

$$(x_{s,i} - \tilde{x}_i) f^{''}(x_{s,i}) \leq \underbrace{f^{'}(x_{s,i}) - f^{'}(\tilde{x}_i)}_{\mu - \tilde{\mu} + \tilde{\ell}_i} \leq (x_{s,i} - \tilde{x}_i) f^{''}(\tilde{x}_i). \tag{3.34}$$

Since $f^{''}(x_{s,i}) \geq 0$, from the left side of (3.34) we get $x_{s,i} - \tilde{x}_i \leq f^{''}(x_{s,i})^{-1}\left(\mu - \tilde{\mu} + \tilde{\ell}_i\right)$. Taking summation over all i and using the fact that both vectors x_s and $\tilde{x}$ are probability vectors, we have

$$\begin{aligned} 0 = \sum_{i=1}^K x_{s,i} - \tilde{x}_i &\leq \sum_{i=1}^K f^{''}(x_{s,i})^{-1}\left(\mu - \tilde{\mu} + \tilde{\ell}_i\right) \\ &\Rightarrow \tilde{\mu} - \mu \leq \frac{\sum_{i=1}^K f^{''}(x_{s,i})^{-1}\tilde{\ell}_i}{\sum_{i=1}^K f^{''}(x_{s,i})^{-1}}. \end{aligned} \tag{3.35}$$

Combining the right hand sides of (3.34) and (3.35) gives

$$(\tilde{x}_i - x_{s,i}) f^{''}(\tilde{x}_i) \leq \tilde{\mu} - \mu - \tilde{\ell}_i \leq \frac{\sum_{j=1}^K f^{''}(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^K, f^{''}(x_{s,j})^{-1}}$$

and by rearrangement

$$\begin{aligned}\tilde{x}_i &\leq x_{s,i} + f''(\tilde{x}_i)^{-1} \times \frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \\ &\leq x_{s,i} + \gamma_s \tilde{x}_i \times \frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}}, \end{aligned} \tag{3.36}$$

where the last inequality holds because $f''(\tilde{x}_i)^{-1} = \left(\eta_s^{-1}\frac{1}{2}\tilde{x}_i^{-3/2} + \gamma_s^{-1}\tilde{x}_i^{-1}\right)^{-1}$. The next step for bounding $\tilde{x}_i$ is to bound $\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}}$ in (3.36), where $\tilde{\ell}_j = \sum_{r\in A}\hat{\ell}_{r,j}$ and $A = \{r : s \leq r + d_r < t\}$.

If there exists $r \in A$, such that $r > s$ and $2x_{r,i} \leq x_{s,i}$, then combining it with the induction assumption for (t, r), i.e., $x_{t,i} \leq 2x_{r,i}$, leads to $x_{t,i} \leq 2x_{r,i} \leq x_{s,i}$, which completes the proof. Otherwise, that for all $r \in A$ we have either $r \leq s$ or $x_{s,i} \leq 2x_{r,i}$. If $r \leq s$, we can use the induction assumption for (s, r), which gives $x_{s,i} \leq 2x_{r,i}$. Consequently, in either case, the inequality $x_{s,i} \leq 2x_{r,i}$ holds for all $r \in A$, and we can plug it into Lemma 3.10 to get the following bound for all $r \in A$:

$$\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\hat{\ell}_{r,j}}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \leq 4(K-1)^{\frac{1}{3}}. \tag{3.37}$$

We then proceed by doing a summation over all $r \in A$ on both sides of the above inequality and get $\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \leq 4|A|(K-1)^{\frac{1}{3}}$. Now it suffices to plug this result into (3.36):

$$\begin{aligned} &\tilde{x}_i \leq x_{s,i} + 4|A|\gamma_s\tilde{x}_i(K-1)^{\frac{1}{3}} \Rightarrow \\ &\tilde{x}_i \leq x_{s,i} \times \left(\frac{1}{1 - 4|A|\gamma_s(K-1)^{1/3}}\right) \qquad (3.38) \\ &\quad \leq x_{s,i} \times \left(\frac{1}{1 - 8\gamma_s d_{max}(K-1)^{1/3}}\right) \\ &\quad \leq x_{s,i} \times \left(\frac{1}{1 - 8\sqrt{\log K/\gamma_0}\, d_{max}(K-1)^{1/3}}\right) = \frac{x_{s,i}}{1-1/3} = \frac{3}{2}x_{s,i}, \qquad (3.39) \end{aligned}$$

where the third inequality uses $|A| \leq d_{max} + t - s \leq 2d_{max}$, and the last one uses the facts that $\gamma_s \leq \sqrt{\log(K)/\gamma_0}$ and $\gamma_0 = 24^2 d_{max}^2 (K-1)^{2/3}\log(K)$.

Combining (3.39) and (3.33) completes the proof. □

3.8.3 Detailed constant factors in the regret bound for Algorithm 2

In this section we provide a detailed regret bound for Algorithm 2.
As we proved in Section 3.5 we have the following inequality for the drifted regret:

$$\overline{Reg}_T \leq 2\overline{Reg}_T^{drift} + d_{max} \tag{3.40}$$

We first derive a bound for the drifted regret by splitting the drifted regret into stability and penalty terms, as mentioned in Section 3.5. Following the general analysis of the penalty term for FTRL (Abernethy et al., 2015), we have

$$penalty \leq \sum_{t=2}^{T} (F_{t-1}(x_t) - F_t(x_t)) + F_T(x^*) - F_1(x_1),$$

which gives us

$$\begin{aligned}
penalty =& \sum_{t=2}^{T} \left(2(\sum_{i=1}^{K} x_{t,i}^{\frac{1}{2}} - 1)(\eta_t^{-1} - \eta_{t-1}^{-1}) - \sum_{i=1}^{K} x_{t,i} \log(x_{t,i})(\gamma_t^{-1} - \gamma_{t-1}^{-1}) \right) \\
& - 2\eta_1^{-1} + 2\sqrt{K}\eta_1^{-1} + \gamma_1^{-1} \log K \\
\leq & \sum_{t=2}^{T} \left(2 \sum_{i \neq i^*} x_{t,i}^{\frac{1}{2}} (\eta_t^{-1} - \eta_{t-1}^{-1}) - \sum_{i=1}^{K} x_{t,i} \log(x_{t,i})(\gamma_t^{-1} - \gamma_{t-1}^{-1}) \right) \\
& + 2\sqrt{\eta_0(K-1)} + \sqrt{\gamma_0 \log K} \\
\leq & \sum_{t=2}^{T} \left(2 \sum_{i \neq i^*} \eta_t x_{t,i}^{\frac{1}{2}} - \sum_{i=1}^{K} \frac{\sigma_t \gamma_t x_{t,i} \log(x_{t,i})}{\sqrt{\log K}} \right) + 2\sqrt{\eta_0(K-1)} + \sqrt{\gamma_0 \log K},
\end{aligned} \tag{3.41}$$

where the first inequality holds because $x_{t,i^*}^{\frac{1}{2}} \leq 1$ and the second inequality follows by $\eta_t^{-1} - \eta_{t-1}^{-1} = \sqrt{t+\eta_0} - \sqrt{t-1+\eta_0} \leq \frac{1}{\sqrt{t+\eta_0}} = \eta_t$ and $\gamma_t^{-1} - \gamma_{t-1}^{-1} = \frac{\gamma_t^{-2} - \gamma_{t-1}^{-2}}{\gamma_t^{-1} + \gamma_{t-1}^{-1}} \leq \frac{\gamma_t^{-2} - \gamma_{t-1}^{-2}}{\gamma_t^{-1}}$.

For the stability term, we start from the bound given by Lemma 3.3:

$$\mathbb{E}[stability] \leq \sum_{t=1}^{T} \sum_{i \neq i^*} 2\gamma_t(\upsilon_t - 1)\upsilon_t \mathbb{E}[x_{t,i}]\Delta_i + \sum_{t=1}^{T} \sum_{s \in A_t} \sum_{i=1}^{K} \eta_t \mathbb{E}[x_{t,i}^{3/2} x_{s,i}^{-1}(1 - x_{s,i})]. \tag{3.42}$$

In above inequality, we know that $\upsilon_t x_{t,i} = \sum_{s \in A_t} x_{t,i}$, and by Lemma 3.2 we have $x_{t,i} \leq 2x_{s,i}$ for $s \in A_t$. Then for the first term in (3.42):

$$\begin{aligned}\sum_{t=1}^{T} \sum_{i \neq i^*} 2\gamma_t(\upsilon_t - 1)\upsilon_t x_{t,i}\Delta_i &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \sum_{s \in A_t} 4\gamma_t(\upsilon_t - 1)\upsilon_t x_{s,i}\Delta_i \\ &= \sum_{t=1}^{T} \sum_{i \neq i^*} 4\gamma_{t+d_t}(\upsilon_{t+d_t} - 1)x_{t,i}\Delta_i. \end{aligned} \tag{3.43}$$

Furthermore, we can bound $x_{t,i}^{3/2} x_{s,i}^{-1}(1 - x_{s,i}) \leq 2^{3/2} x_{s,i}^{1/2}(1 - x_{s,i})$. Moreover, in order to remove the best arm i^* from the summation in the later bound we use $x_{t,i^*}^{3/2} x_{s,i^*}^{-1}(1 - x_{s,i^*}) \leq 2\sum_{i \neq i^*} x_{s,i} \leq \sum_{i \neq i^*} 2x_{s,i}^{1/2}$.

For the second term in (3.42) we have

$$\begin{aligned}\sum_{t=1}^{T} \sum_{s \in A_t} \sum_{i=1}^{K} \eta_t x_{t,i}^{3/2} x_{s,i}^{-1}(1 - x_{s,i}) &\leq \sum_{t=1}^{T} \sum_{s \in A_t} \sum_{i=1}^{K} \eta_t 2^{3/2} x_{s,i}^{1/2}(1 - x_{s,i}) \\ &\leq \sum_{t=1}^{T} \sum_{s \in A_t} \sum_{i \neq i^*} \sqrt{8}\eta_t x_{s,i}^{1/2} + \sum_{t=1}^{T} \sum_{s \in A_t} \sum_{i \neq i^*} 2\eta_t x_{s,i}^{1/2} \\ &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} 5\eta_t x_{t,i}^{1/2}, \end{aligned} \tag{3.44}$$

where the last inequality follows by the facts that we can change the order of the summations and that each t belongs to exactly one A_s. Plugging (3.43) and (3.44) into (3.42) we have

$$\mathbb{E}[\mathit{stability}] \leq \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \neq i^*} 4\gamma_{t+d_t}(\upsilon_{t+d_t} - 1)x_{t,i}\Delta_i + \sum_{t=1}^{T} \sum_{i \neq i^*} 5\eta_t x_{t,i}^{1/2}\right]. \tag{3.45}$$

Now it suffices to combine (3.45), (3.41), and (3.40) to get

$$\overline{Reg}_T \leq \mathbb{E}\left[\underbrace{14\sum_{t=1}^{T}\sum_{i\neq i^*}\eta_t x_{t,i}^{1/2}}_{A} + \underbrace{8\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t}(\upsilon_{t+d_t}-1)x_{t,i}\Delta_i}_{B}\right] + \mathbb{E}\left[\underbrace{2\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\sigma_t\gamma_t x_{t,i}\log(1/x_{t,i})}{\log K}}_{C}\right] + \underbrace{4\sqrt{\eta_0(K-1)}+2\sqrt{\gamma_0\log K}+d_{max}}_{D}. \tag{3.46}$$

We rewrite the regret as

$$\overline{Reg}_T = 4\overline{Reg}_T - 3\overline{Reg}_T \leq 4\times 14A - \overline{Reg}_T + 4\times 8B - \overline{Reg}_T + 4\times 2C - \overline{Reg}_T + 4D,$$

where by applying Lemmas 3.4, 3.5, and 3.6 we achieve

$$\begin{aligned} 4\times 14A - \overline{Reg}_T &\leq \sum_{i\neq i^*}\frac{28^2}{\Delta_i}\log(T/\eta_0+1)\\ 4\times 8B - \overline{Reg}_T &\leq 64^2\upsilon_{max}\log K\\ 4\times 2C - \overline{Reg}_T &\leq \sum_{i\neq i^*}\frac{512\sigma_{max}}{\Delta_i\log K}. \end{aligned}$$

Therefore, the final regret bound is

$$\begin{aligned} \overline{Reg}_T \leq \sum_{i\neq i^*}\frac{28^2}{\Delta_i}\log(T/\eta_0+1) + 64^2\upsilon_{max}\log K + \sum_{i\neq i^*}\frac{512\sigma_{max}}{\Delta_i\log K}\\ + 16\sqrt{\eta_0(K-1)} + 8\sqrt{\gamma_0\log K} + 4d_{max}. \end{aligned}$$

3.8.4 Removing the multiplicative factor $1/\Delta_i$ from σ_{max}/Δ_i in the regret bound

In this section we discuss how an asymmetric *oracle* learning rate $\gamma_{t,i} \simeq \gamma_t/\sqrt{\Delta_i}$ for negative entropy regularizer can be used to remove the factor $\sum_{i\neq i^*} 1/\Delta_i$ in front of σ_{max} in the regret bound.

In the analysis of Algorithm 2 we divided the regret into stability and penalty expressions. Moreover, in each of the bounds for stability and penalty we have two terms which correspond to negative entropy and Tsallis parts of the hybrid regularizer. The terms related to negative entropy part in both stability and penalty bounds are

$$\underbrace{\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t}(\upsilon_{t+d_t}-1)\mathbb{E}[x_{t,i}]\Delta_i}_{B}+\underbrace{\sum_{i=1}^{K}\mathbb{E}[x_{t,i}\log(1/x_{t,i})](\gamma_t^{-1}-\gamma_{t-1}^{-1})}_{C},$$

where B and C, as we have seen in Section 3.5, are due to stability and penalty terms,respectively. The idea here is to scale-up γ_t to decrease C, however increasing γ_t increases B. Hence, we are facing a trade off here. To deal with this trade-off we change the learning rates for negative entropy from symmetric γ_t to asymmetric $\gamma_{t,i}$, and we expect this change only affect the parts of regret bound come from the negative entropy part of the regularizer, which are B and C. This change results in to having two following terms instead,

$$\underbrace{\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t,i}(\upsilon_{t+d_t}-1)\mathbb{E}[x_{t,i}]\Delta_i}_{B_{new}}+\underbrace{\sum_{i=1}^{K}\mathbb{E}[x_{t,i}\log(1/x_{t,i})](\gamma_{t,i}^{-1}-\gamma_{t-1,i}^{-1})}_{C_{new}}.$$

Here if we could choose $\gamma_{t,i}=\gamma_t/\sqrt{\Delta_i}$, then using the definition of γ_t we would be able to rewrite B_{new} and C_{new} as

$$B_{new}=\mathcal{O}\left(\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+d_t}(\upsilon_{t+d_t}-1)\mathbb{E}[x_{t,i}]\sqrt{\Delta_i}\right)$$
$$C_{new}=\mathcal{O}\left(\sum_{i=1}^{K}\frac{\sigma_t\gamma_t\mathbb{E}[x_{t,i}\log(1/x_{t,i})]\sqrt{\Delta_i}}{\sqrt{\log K}}\right).$$

Now we must see what is the result of applying the self-bounding technique on these new terms. For B_{new} and C_{new}, following the similar analysis as Lemma 3.5 and Lemma 3.6 we can get

$$4B_{new}-\overline{Reg}_T=\mathcal{O}(\upsilon_{max}\log K)=\mathcal{O}(d_{max}\log K)$$
$$4C_{new}-\overline{Reg}_T=\mathcal{O}(\frac{\sigma_{max}}{\log K}).$$

This implies that injecting $\sqrt{1/\Delta_i}$ in the negative entropy learning rates removes the factor $\sum_{i \neq i^*} \frac{1}{\Delta_i}$ in front of the σ_{max}. More interestingly this comes without having any significant changes in the other terms of regret bound.
As a result, we conjecture that replacing a good estimation of the suboptimal gaps namely $\hat{\Delta}_i$ in $\gamma_{t,i}$ as $\gamma_{t,i} = \gamma_t / \sqrt{\hat{\Delta}_i}$ might be also helpful to remove the multiplicative factors related to suboptimal gaps in front of the σ_{max}. We leave this problem to future work.

3.8.5 Lower bounds

Algorithm 3: Adversarial choice of ℓ

Input: x

Initialize $\mathcal{I} = \{\mathrm{argmax}_i \, x_i\}$ **while** $\sum_{i \in \mathcal{I}} x_i + \min_{i \in \bar{\mathcal{I}}} x_i \leq \frac{2}{3}$ **do**

  Update $\mathcal{I} \leftarrow \mathcal{I} \cup \{\mathrm{argmin}_{i \in \bar{\mathcal{I}}} \, x_i\}$

return $\ell_i = \begin{cases} \min\{1, \frac{\sum_{i \in \bar{\mathcal{I}}} x_i}{\sum_{i \in \mathcal{I}} x_i}\} & \text{for } i \in \mathcal{I} \\ \max\{-1, -\frac{\sum_{i \in \mathcal{I}} x_i}{\sum_{i \in \bar{\mathcal{I}}} x_i}\} & \text{for } i \in \bar{\mathcal{I}} \end{cases}$

Lemma 3.11. *For any $x \in \Delta([K])$, such that $\max_i x_i \leq \frac{2}{3}$, the vector ℓ returned by Algorithm 3 satisfies $\ell \in [-1, 1]$, $\langle x, \ell \rangle = 0$, and $\sum_{i=1}^K x_i \ell_i^2 \geq \frac{1}{2}$.*

Proof. The first two properties follow directly by construction. For the third property we bound the ratio of the two sets. Assume that $\sum_{i \in \mathcal{I}} x_i < \frac{1}{3}$, then $\mathrm{argmin}_{i \in \bar{\mathcal{I}}} \, x_i < \frac{1}{3}$ and the algorithm does not return yet, so at the end $\sum_{i \in \mathcal{I}} x_i \in [\frac{1}{3}, \frac{2}{3}]$. Let $p = \max\{\sum_{i \in \mathcal{I}} x_i, 1 - \sum_{i \in \mathcal{I}} x_i\}$, then $p \in [\frac{1}{3}, \frac{2}{3}]$ and the quantity in question is bounded by

$$\sum_{i=1}^K x_i \ell_i^2 = \sum_{i \in \mathcal{I}} x_i \ell_i^2 + \sum_{i \in \bar{\mathcal{I}}} x_i \ell_i^2 = p + (1-p) \left(\frac{p}{1-p} \right)^2 = \frac{p}{1-p} \geq \frac{1}{2} \, .$$

□

Claim 3.6. *For the negentropy potential $F(x) = \eta^{-1} \sum_{i=1}^K \log(x_i) x_i$, it holds that*

$$-\overline{F}^*(-L) - \min_i L_i = \eta^{-1} \log(\max_i \nabla \overline{F}^*(-L)_i) \, .$$

Proof. Denote $i^* = \operatorname{argmin}_{i\in[K]} L_i$. It is well known that the exponential weights distribution is $(\nabla\overline{F}^*(-L))_i = \exp(-\eta L_i)/(\sum_{j\in[K]}\exp(-\eta L_j)$. Therefore, the negentropy has an explicit form of the constrained convex conjugate:

$$\overline{F}^*(-L) = \left\langle \nabla\overline{F}^*(-L), -L \right\rangle - F(\nabla\overline{F}^*(-L)) = \eta^{-1}\log(\sum_{i=1}^{K}\exp(-\eta L_i))\,.$$

Hence

$$\begin{aligned}-\overline{F}^*(-L) - L_{i^*} &= -\eta^{-1}\log\left(\sum_{i=1}^{K}\exp(-\eta L_i)\right) + \eta^{-1}\log(\exp(-\eta L_{i^*})) \\ &= -\eta^{-1}\log\left(\frac{\sum_{i=1}^{K}\exp(-\eta L_i)}{\exp(-\eta L_{i^*})}\right) = \eta^{-1}\log\left(\nabla\overline{F}^*(-L)_{i^*}\right).\end{aligned}$$

□

Proof of Theorem 3.2. For ease of presentation, we will work with loss ranges $[-L_t/2, L_t/2]$, which is equivalent to loss ranges of $[0, L_t]$ in full-information games. Assume that

$$\frac{1}{2}\sum_{t=1}^{\lfloor\log_2(K)\rfloor} L_t \geq \frac{1}{32}\sqrt{\sum_{t=\lfloor\log_2(K)\rfloor}^{T} L_t^2\log(K)}\,.$$

Define the active set $\mathcal{A}_1 = [K]$. At any time t, if $\rho(t) \notin [\lfloor\log_2(K)\rfloor]$, we set ℓ_t to 0 and proceed with $\mathcal{A}_{t+1} = \mathcal{A}_t$. Otherwise, if $\rho(t) \in [\lfloor\log_2(K)\rfloor]$, we randomly select half of the arms in $\mathcal{A}_t$ to assign $\ell_{t,i} = -L_{\rho(t)}/2$, and the other half $\ell_{t,i} = L_{\rho(t)}/2$. (In case of an uneven number $|\mathcal{A}_t|$ we leave one arm at 0.) All other losses are 0. We reduce $\mathcal{A}_{t+1} = \{i \in \mathcal{A}_t \,|\, \ell_{t,i} < 0\}$ to the set of arms that were negative. The set $\mathcal{A}_n$ will not be empty since we can repeat halving the action set exactly $\lfloor\log_2(K)\rfloor$ many times. The expected loss of any player is always 0, while the loss of the best arm is $\min_a \sum_{t=1}^{T}\ell_{t,a} = -\sum_{t=1}^{\lfloor\log_2(K)\rfloor} L_t/2$, hence

$$\mathbb{R}^* \geq \sum_{t=1}^{\lfloor\log_2(K)\rfloor} L_t/2\,.$$

It remains to analyse the case

$$\frac{1}{2}\sum_{t=1}^{\lfloor\log_2(K)\rfloor} L_t < \frac{1}{32}\sqrt{\sum_{t=\lfloor\log_2(K)\rfloor}^{T} L_t^2\log(K)}\,.$$

In this case, note that we have

$$\sqrt{\sum_{t=\lfloor \log_2(K)\rfloor}^{T} L_t^2/\log(K)} > \frac{16}{\log(K)} \sum_{t=1}^{\lfloor \log_2(K)\rfloor} L_t > 16\frac{\lfloor \log_2(K)\rfloor}{\log(K)} L_{\lfloor \log_2(K)\rfloor} > 8L_{\lfloor \log_2(K)\rfloor}\,. \tag{3.47}$$

The high level idea is now to create a sequence of losses adapted to the choices of the algorithm. Let $x_{ti} = \mathbb{E}\left[I_t = i|\ell_{t-1},\ldots,\ell_1\right]$ be the expected trajectory of the algorithm and let $z_{ti} = \exp(-\eta L_{ti})/\sum_{j=1}^{K}\exp(-\eta L_{tj})$ for $L_t = \sum_{s=1}^{t-1}\ell_t$ be the trajectory of EXP3. Let the adversary follow Algorithm 4 for the selection of losses, then based on Lemma 3.11 we have $0 = \langle z_t, \ell_t\rangle$ and also it is easy to see $0 \le \langle x_t, \ell_t\rangle$, therefore we have $0 = \langle z_t, \ell_t\rangle \le \langle x_t, \ell_t\rangle$. This implies that the regret of the algorithm cannot be smaller than that of EXP3, so the regret of algorithm $\mathcal{A}$ can be bounded as

$$\begin{aligned} Reg_T(\mathcal{A}) &= \sum_{t=1}^{T}\langle x_t, \ell_t\rangle - \min_{a^*\in\Delta([K])}\langle a^*, L_{T+1}\rangle \\ &\ge \sum_{t=1}^{T}\langle z_t, \ell_t\rangle - \min_{a^*\in\Delta([K])}\langle a^*, L_{T+1}\rangle = -\min_{a^*\in\Delta([K])}\langle a^*, L_{T+1}\rangle\ , \end{aligned}$$

Let $F(x) = \eta^{-1}\sum_{i=1}^{K} x_i\log(x_i)$ then we have

$$\begin{aligned} -\min_{a^*\in\Delta([K])}\langle a^*, L_{T+1}\rangle =& \sum_{t=1}^{T}\left[\overline{F}^*(-L_{t+1}) - \overline{F}^*(-L_t)\right] + \overline{F}^*(-L_1) - \overline{F}^*(-L_{T+1}) \\ &- \min_{a^*\in\Delta([K])}\langle a^*, L_{T+1}\rangle \\ =& \sum_{t=1}^{T}\eta^{-1}\log\left(\sum_{i=1}^{K}\exp(-\eta L_{t+1,i})\right) - \eta^{-1}\log\left(\sum_{i=1}^{K}\exp(-\eta L_{t,i})\right) \\ &+ \eta^{-1}\log(K) + \eta^{-1}\log(\max_{i\in[K]} z_{T+1,i}) \\ =& \sum_{t=1}^{T}\eta^{-1}\log\left(\sum_{i=1}^{K} z_{ti}\exp(-\eta\ell_{ti})\right) + \eta^{-1}\log(K) \\ &+ \eta^{-1}\log(\max_{i\in[K]} z_{T+1,i}), \end{aligned}$$

where the second equality uses Claim 3.6 for $L = L_{T+1}$ and the fact for any $L \in \mathbb{R}^K$, $\overline{F}^*(-L) = \eta^{-1}exp(\sum_{i=1}^{K} -\eta L_i)$. Now we choose the learning rate for EXP3 to be $\eta = \sqrt{\log(K)/(\sum_{t=\lfloor \log_2(K)\rfloor}^{T} L_t^2)}$, that based on (3.47) together with the fact that we

set losses to zero for $\rho(t) \in [\lfloor \log_2 K \rfloor]$ in Algorithm 4 ensures $|\eta \ell_{ti}| \leq \frac{1}{2}\eta L_{\lfloor \log_2(K) \rfloor} \leq \frac{1}{2}$. Using that, by Taylor's theorem and the monotonicity of the second derivative of exp, we have for all $x \geq -\frac{1}{2}$: $\exp(x) \geq 1 + x + \frac{1}{2}\exp''(-\frac{1}{2})x^2 \geq 1 + x + \frac{3}{10}x^2$, as well as by concavity of log for all $0 \leq x \leq \frac{1}{4}$ we have $\log(1+x) \geq 4\log(5/4)x \geq \frac{5}{6}x$, we get for any $t \in [T]$ by Lemma 3.11

$$\begin{aligned}\eta^{-1}\log(\sum_{i=1}^{K} z_{ti}\exp(-\eta\ell_{ti})) &\geq \eta^{-1}\log(1+\eta^2\frac{3}{10}\sum_{i=1}^{K} z_{ti}\ell_{ti}^2) \\ &\geq \frac{\eta}{4}\sum_{i=1}^{K} z_{ti}\ell_{ti}^2 \geq \mathbb{I}\{\max_i z_{ti} \leq \frac{2}{3}\}\frac{\eta}{32}L_{\rho^{-1}(t)}^2 \,.\end{aligned}$$

Now we have two possible events, either $\forall t \in [T] : \max_i z_{ti} \leq \frac{2}{3}$ and

$$Reg_T(\mathcal{A}) \geq \frac{\eta}{32}\sum_{t=\lfloor \log_2(K) \rfloor} L_t^2 = \frac{1}{32}\sqrt{\sum_{t=\lfloor \log_2(K) \rfloor} L_t^2 \log(K)}\,,$$

or there exists $s \in [T]$ such that $\max_i z_{s,i} > \frac{2}{3}$, then from Algorithm 4 we infer that $\forall t \geq s : \ell_t = 0$ and consequently $\forall t \geq s : z_t = z_s$, so $\max_i z_{T+1,i} > \frac{2}{3}$ and

$$Reg_T(\mathcal{A}) \geq \eta^{-1}(\log(K) + \log(2/3)) \geq \frac{1}{32}\eta^{-1}\log(K) = \frac{1}{32}\sqrt{\sum_{t=\lfloor \log_2(K) \rfloor} L_t^2 \log(K)}\,.$$

□

Algorithm 4: Adversary

Input: Actor $\mathcal{A}$, learning rate η

for $t = 1, \ldots, n$ **do**
- Set $\forall i : z_{ti} = \exp(-\eta L_{ti}) / \sum_{j=1}^{K} \exp(-\eta L_{tj})$
- **if** $\max_{i \in [K]} z_{ti} > \frac{2}{3}$ *or* $\rho(t) \leq \lfloor \log_2(K) \rfloor$ **then**
  - $\ell_t = 0$
- **else**
  - Get ℓ from Algorithm 3 with $x = z_t$.
  - Determine $x_t = \mathbb{E}\left[\mathcal{A}((\ell_s)_{s=1}^{t-1})\right]$
  - Set $\ell_t = \text{sign}(\langle x_t, \ell \rangle) L_{\rho^{-1}(t)}\ell/2$

Chapter 4

An Improved Best-of-both-worlds Algorithm for Bandits with Delayed Feedback

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

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. An improved best-of-both-worlds algorithm for bandits with delayed feedback. `https://arxiv.org/abs/2308.10675`, 2023.

Abstract

We propose a new best-of-both-worlds algorithm for bandits with variably delayed feedback. The algorithm improves on prior work by Masoudian et al. (2022) by eliminating the need in prior knowledge of the maximal delay $d_{\max}$ and providing tighter regret bounds in both regimes. The algorithm and its regret bounds are based on counts of outstanding observations (a quantity that is observed at action time) rather than delays or the maximal delay (quantities that are only observed when feedback arrives). One major contribution is a novel control of distribution drift, which is based on biased loss estimators and skipping of observations with excessively large delays. Another major contribution is demonstrating that the complexity of best-of-both-worlds bandits with delayed feedback is characterized by the cumulative count of outstanding observations after skipping of observations with excessively large delays, rather than the delays or the maximal delay.

4.1 Introduction

Delayed feedback is an ubiquitous challenge in real-world applications. Study of multiarmed bandits with delayed feedback has started at least four decades ago in the context of adaptive clinical trials (Simon, 1977; Eick, 1988), the same problem that has earlier motivated introduction of the bandit model itself (Thompson, 1933).

Joulani et al. (2013) have studied multiarmed bandits with delayed feedback under the assumption that the rewards are stochastic and the delays are sampled from a fixed distribution. They provided a modification of the UCB1 algorithm, which was originally designed for stochastic bandits with non-delayed feedback (Auer et al., 2002a). Joulani et al. have shown that the regret of the modified algorithm is $O\left(\sum_{i:\Delta_i>0}\left(\frac{\log T}{\Delta_i}+\sigma_{\max}\Delta_i\right)\right)$, where i indexes the arms, Δ_i is the suboptimality gap of arm i, T is the time horizon (unknown to the algorithm), and $\sigma_{\max}$ is the maximal number of outstanding observations. The result implies that in the stochastic setting the delays introduce an additive term in the regret bound, proportional to the maximal number of outstanding observation.

In the adversarial setting, multiarmed bandits with delayed feedback were first analyzed under the assumption of uniform delays (Neu et al., 2010, 2014). For this setting Cesa-Bianchi et al. (2019) have shown an $\Omega(\sqrt{KT}+\sqrt{dT\log K})$ lower bound and an almost matching upper bound, where K is the number of arms, T is the time horizon, and d is the fixed delay. The algorithm of Cesa-Bianchi et al. is based on a modification of the EXP3 algorithm of (Auer et al., 2002b). Cesa-Bianchi et al. used

a fixed learning rate that is tuned based on the knowledge of d. The analysis is based on control of the drift of the distribution over arms played by the algorithm from round t to round $t + d$. Then Thune et al. (2019) and Bistritz et al. (2019) provided algorithms for variable adversarial delays, but under the assumption that the delays are known "at action time", meaning that the delay d_t is known at time t, when the action is taken, rather that at time $t + d_t$, when the observation arrives. The advanced knowledge of delays is necessary to tune the learning rate and control the drift of played distribution from round t, when an action is played, to round $t + d_t$, when the observation arrives. Alternatively, an advance knowledge of the cumulative delay up to the end of the game can be used for the same purpose. Finally, Zimmert and Seldin (2020) derived an algorithm for the adversarial setting that required no advance knowledge of delays and matched the lower bound of Cesa-Bianchi et al. (2019) within constants. The algorithm and analysis of Zimmert and Seldin are parametrized by running counts of the number of outstanding observations σ_t, an empirical quantity that is observed at time t ("at the time of action"), and avoids explicit control of the distribution drift.

Masoudian et al. (2022) attempted to extend the algorithm of Zimmert and Seldin (2020) to best-of-both-worlds setting. In best-of-both-worlds setting the goal is to derive algorithms that simultaneously provide an adversarial regret guarantee and a refined regret bound in case the environment happens to be stochastic, without prior knowledge of the nature of the environment. The stochastic part of the analysis of Masoudian et al. is based on a direct control of the distribution drift and, therefore, they had to go back and reintroduce an assumption that the maximal delay $d_{\max}$ is known. The maximal delay is used to tune the learning rate, to control the drift of playing distribution from round t to round $t + d_t$, and eventually shows up additively in both the stochastic and the adversarial regret bounds. Thus, in presence of just a single delay of order T, both the stochastic and the adversarial bounds could be linear in the time horizon.

We introduce a different best-of-both-worlds modification of the algorithm of Zimmert and Seldin (2020) that is fully parametrized by the running count of outstanding observations and requires no advance knowledge of delays or the maximal delay. Our algorithm is based on a careful augmentation of the algorithm of Zimmert and Seldin with implicit exploration (described below), followed by application of the skipping technique (also described below) as an alternative tool to limit the time span over which we need to control the distribution shift.

Implicit exploration was introduced by Neu (2015) as a tool to control the variance of importance-weighted loss estimates. Our application of implicit exploration was inspired by the work of Jin et al. (2022), who used it to control the variance of

Table 4.1: Comparison to state-of-the-art. The following notation is used: T is the time horizon, K is the number of arms, i indexes the arms, Δ_i is the suboptimality gap or arm i, $\sigma_{\max}$ is the maximal number of outstanding observations, $D = \sum_{t=1}^{T} d_t$ is the total delay, $\mathcal{S} \subseteq [T]$ is a subset of indexes of game rounds, in general it is a collection of rounds with excessively large delays that are skipped, $\bar{\mathcal{S}} = [T] \setminus \mathcal{S}$ is the complementary set of rounds, $D_{\bar{\mathcal{S}}} = \sum_{t \in \bar{\mathcal{S}}} d_t$ is the total delay in rounds that are *not* skipped, and $d_{\max}$ is the maximal delay. We have $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}} \right) \leq \sqrt{D}$ and $\sigma_{\max} \leq d_{\max}$, and in some cases $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}} \right) \ll \sqrt{D}$ and $\sigma_{\max} \ll d_{\max}$. Therefore, bounds that exploit skipping are generally tighter than the bounds without skipping, and terms involving $\sigma_{\max}$ are generally smaller than terms involving $d_{\max}$. The bounds of Masoudian et al. (2022) cannot benefit from skipping due to the $d_{\max}$ term (see Appendix 4.7.7). The S^* term in our stochastic bound is the number of rounds skipped by the algorithm. In Appendix 4.7.6 we show that S^* never exceeds $d_{\max}$.

Paper	Key results
Joulani et al. (2013)	Stochastic bound: $\mathcal{O}\left(\sum_{i:\Delta_i>0}\left(\frac{\log T}{\Delta_i} + \sigma_{\max}\Delta_i\right)\right)$
Zimmert and Seldin (2020)	Adversarial bound without skipping: $\mathcal{O}\left(\sqrt{KT} + \sqrt{D\log K}\right)$ Adversarial bound with skipping: $\mathcal{O}\left(\sqrt{KT} + \min_{\mathcal{S}}\left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}\log K}\right)\right)$ A matching lower bound is provided by Masoudian et al. (2022)
Masoudian et al. (2022)	Best-of-both-worlds bound, stochastic part $\mathcal{O}\left(\sum_{i\neq i^*}\left(\frac{\log T}{\Delta_i} + \frac{\sigma_{\max}}{\Delta_i \log K}\right) + d_{\max}K^{1/3}\log K\right)$
The results assume oracle knowledge of $d_{\max}$	Best-of-both-worlds bound, adversarial part $\mathcal{O}\left(\sqrt{TK} + \sqrt{D\log K} + d_{\max}K^{1/3}\log K\right)$
Our paper	Best-of-both-worlds bound, stochastic part $\mathcal{O}\left(\sum_{i\neq i^*}\left(\frac{\log T}{\Delta_i} + \frac{\sigma_{\max}}{\Delta_i \log K}\right) + K\sigma_{\max} + S^*\right)$, $S^* = \mathcal{O}\left(\min\left(d_{\max}, \min_{\mathcal{S}}\left\{|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}K^{\frac{2}{3}}\log K}\right\}\right)\right)$ Best-of-both-worlds bound, adversarial part $\mathcal{O}\left(\sqrt{KT} + \min_{\mathcal{S}}\left\{|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}K^{\frac{2}{3}}\log K}\right\} + K\sigma_{\max}\right)$

importance-weighted loss estimates in Markov decision processes with delayed feedback. However, our parametrization of implicit exploration is different from prior work, because we need to make it work in best-of-both-worlds setting, i.e., it should not deteriorate the bounds in either of the two settings, which makes it challenging.

Skipping was introduced by Thune et al. (2019) as a way to limit the dependence of an algorithm on a small number of excessively large delays. The idea is that the regret in every round is at most 1 and, therefore, it is "cheaper" to skip a round with an excessively large delay and bound the regret in the corresponding round by 1, rather than include it in the core analysis. As already mentioned, Thune et al. have assumed prior knowledge of delays, but Zimmert and Seldin (2020) have perfected the skipping technique by basing it on the running count of outstanding observations. In both prior works skipping was an optional add-on aimed to improve regret bounds in case of highly unbalanced delays. In our work skipping becomes an indispensable part of the algorithm, because, apart from making the algorithm robust to few excessively large delays, it also limits the time span over which the control over playing distribution drift is needed.

We compare our results to key results from prior work in Table 4.1. It has been shown by Joulani et al. (2013) and Masoudian et al. (2022) that $\sigma_{\max} \leq d_{\max}$, and that in some cases $\sigma_{\max} \ll d_{\max}$. For example, if the first observation has a delay of T, and all the remaining observations have delay zero, then $d_{\max} = T$, but $\sigma_{\max} = 1$. Therefore, bounds in terms of $\sigma_{\max}$ are preferable over bounds in terms of $d_{\max}$, and for some problem instances the improvement may be very significant. We also have that $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}} \log K}\right) \leq \sqrt{D \log K}$, where $\mathcal{S} \subseteq [T]$ is a subset of game rounds skipped by an algorithm, $\bar{\mathcal{S}} = [T] \setminus \mathcal{S}$ is the complementary set of rounds, $D = \sum_{t=1}^{T} d_t$ is the total delay, and $D_{\bar{\mathcal{S}}} = \sum_{t \in \bar{\mathcal{S}}} d_t$ is the total delay in rounds that are *not* skipped. Furthermore, in some cases $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}} \log K}\right) \ll \sqrt{D \log K}$. Thune et al. (2019) provided an example, where the delays in the first $\sqrt{T}$ rounds of the game are of order T, and the delays in the remaining rounds of the game are zero. In this case $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}} \log K}\right) = \mathcal{O}\left(\sqrt{T}\right)$, but $\sqrt{D \log K} = \mathcal{O}\left(T^{3/4}\right)$. Therefore, bounds that exploit skipping are preferable over bounds that do not, and for some problem instances the improvement may be very significant. In the supplementary material we show that bounds with an additive $d_{\max}$ term, including the results of Masoudian et al. (2022), cannot benefit from skipping, in contrast to our results.

The following list highlights our main contributions.

1. We provide a new technique to control the distribution drift that is independent of $d_{\max}$ and provides regret bounds that depend on $\sigma_{\max}$ rather than $d_{\max}$.

At the conceptual level it implies that the regret is affected by the amount of information missing at the time of decision making (which is bounded by $\sigma_{\max}$) rather than the time that the information is missing (which is bounded by $d_{\max}$).

2. We provide an implicit exploration scheme that works in best-of-both-worlds setting.

3. We improve both the stochastic and the adversarial part of best-of-both-worlds regret bounds relative to Masoudian et al. (2022) by replacing terms dependent on $d_{\max}$ by terms dependent on $\sigma_{\max}$.

4. We make skipping possible and useful, in contrast to Masoudian et al. (2022).

5. It has been shown in prior work that $d_{\max}$ does not appear in the regret bounds neither in the stochastic (Joulani et al., 2013) nor in the adversarial setting Zimmert and Seldin (2020) taken individually, and in general, when each of the two settings is considered in isolation, the regret is unaffected by presence of a small number of excessively large delays. However, the question of whether the same can be achieved in best-of-both-worlds setting was left open by Masoudian et al. (2022). We answer this question positively. The general message is that the delays per se are not the right quantity for characterizing the complexity of bandit learning with delayed feedback.

4.2 Problem setting

We study the problem of multi-armed bandit with variable delays. In each round $t = 1, 2, \ldots$, the learner picks an action I_t from a set of K arms and immediately incurs a loss ℓ_{t,I_t} from a loss vector $\ell_t \in [0,1]^K$. However, the incurred loss is observed by the learner only after a delay of d_t, at the end of round $t + d_t$. The delays are arbitrary and chosen by the environment. We use σ_t to denote the number of outstanding observations at time t defined as $\sigma_t = \sum_{s \le t} \mathbb{1}(s + d_s > t)$ and $\sigma_{\max} = \max_{t \in [T]} \sigma_t$ to be the maximal number of outstanding observations.

We consider two regimes for generation of losses by the environment: oblivious adversarial and stochastic.

We use pseudo-regret to compare the expected total loss of the learner's strategy to that of the best fixed action in hindsight. Specifically, the pseudo-regret is defined

as:

$$\overline{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,I_t}\right] - \min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_{t,I_t} - \ell_{t,i_T^*}\right)\right],$$

where $i_T^* = \min_{i\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,i}\right]$ is the best action in hindsight. In the oblivious adversarial setting, the losses are assumed to be deterministic and independent of the actions taken by the algorithm. As a result, the expectation in the definition of i_T^* can be omitted and the pseudo-regret definition coincides with the expected regret. Throughout the paper we assume that i_T^* is unique. This is a common simplifying assumption in best-of-both-worlds analysis (Zimmert and Seldin, 2021). Tools for elimination of this assumption can be found in Ito (2021).

4.3 Algorithm

The algorithm is a best-of-both-worlds modification of the adversarial FTRL algorithm with hybrid regularizer by Zimmert and Seldin (2020). It is provided in Algorithm 5 display. The modification includes biased loss estimators (implicit exploration) and adjusted skipping threshold. The algorithm maintains a set of skipped rounds $\mathcal{S}_t$ (initially empty), a cumulative count of "active" outstanding observations (those that have not been skipped yet), and a vector of cumulative observed loss estimates $\hat{L}_t^{obs}$ from non-skipped rounds. At round t the algorithm constructs an FTRL distribution x_t over arms using regularizer F_t defined in equation (4.2) below, and samples an arm according to x_t. Then it receives the observations that arrive at round t, except those that come from the skipped rounds, and updates the vector $\hat{L}_t^{obs}$ of cumulative loss estimates. The loss estimates $\hat{\ell}_t$ are defined below in equation (4.1). Then it counts the number of "active" outstanding observations $\hat{\sigma}_t$ (those that belong to non-skipped rounds), updates the cumulative count of outstanding observations $\mathcal{D}_t$, and computes the skipping threshold $d_{\max}^t = \sqrt{\frac{\mathcal{D}_t}{49K^{2/3}\log K}}$. Finally, it adds rounds s for which the observation has not arrived yet and the waiting time $(t-s)$ exceeds the skipping threshold $d_{\max}^t$ to the set of skipped rounds $\mathcal{S}_t$. Lemma 4.14, which is an adaptation of Zimmert and Seldin (2020, Lemma 5) to our skipping rule, shows that at most one round s is skipped at a time (at most one index s satisfies the if-condition for skipping in Line 16 of the algorithm for a given t).

We use implicit exploration to control importance-weighted loss estimates. The idea of using implicit exploration is inspired by the works of Neu (2015) and Jin et al. (2022), but its parametrization and application goal are different from prior work. To the best of our knowledge, it is the first time implicit exploration is used

Algorithm 5: Best-of-both-worlds algorithm for bandits with delayed feedback

Initialize $\mathcal{S}_0 = \emptyset$, $\mathcal{D}_0 = 0$, *and* $\hat{L}_0^{obs} = \mathbf{0}$, *where* $\mathbf{0}$ *is the zero vector in* $\mathbb{R}^K$
for $t = 1, 2, \ldots$ **do**
  // Playing an arm and receiving observations (except from skipped rounds)
  Set $x_t = \arg\min_{x \in \Delta^{K-1}} \langle \hat{L}_{t-1}^{obs}, x \rangle + F_t(x)$
    // F_t is defined in (4.2)
  Sample $I_t \sim x_t$
  for $s : (s + d_s = t) \wedge (s \notin \mathcal{S}_{t-1})$ **do**
    Observe (s, ℓ_{s,I_s})
    $\hat{L}_t^{obs} = \hat{L}_{t-1}^{obs} + \hat{\ell}_s$
      // $\hat{\ell}_s$ is defined in (4.1)
  // Counting "active" outstanding observations and updating the skipping threshold
  Set $\hat{\sigma}_t = \sum_{s \in [t-1] \setminus \mathcal{S}_{t-1}} \mathbb{1}(s + d_s > t)$ *// Count of "active" outstanding observations*
  Update $\mathcal{D}_t = \mathcal{D}_{t-1} + \hat{\sigma}_t$
  Set $d_{\max}^t = \sqrt{\mathcal{D}_t / \left(49K^{\frac{2}{3}} \log K\right)}$
  // Skipping observations with excessively large delays
  // By Lemma 4.14 at most one index s satisfies the if-condition for a given t
  for $s \in [t-1] \setminus \mathcal{S}_{t-1}$ **do**
    if $\min\{d_s, t - s\} \geq d_{\max}^t$ **then**
      $\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{s\}$ // If the waiting time $t - s$ exceeds $d_{\max}^t$, then s is skipped
    else
      $\mathcal{S}_t = \mathcal{S}_{t-1}$

for best-of-both-worlds bounds. For any $s, t \in [T]$ with $s \leq t$ we define implicit exploration terms $\lambda_{s,t} = e^{-\frac{\mathcal{D}_t}{\mathcal{D}_t - \mathcal{D}_s}}$. Our biased importance-weighted loss estimators are defined by

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}\mathbb{1}(I_t = i)}{\max\left\{x_{t,i}, \lambda_{t,t+\hat{d}_t}\right\}}, \tag{4.1}$$

where we use $\hat{d}_s = \min\left(d_s, \min\left\{(t-s) : t - s \geq d^t_{\max}\right\}\right)$ to denote the time that the algorithm waits for the observation from round s. It is the minimum of the delay d_s, and the time $(t-s)$ to the first round when the waiting time exceeds the skipping threshold $d^t_{\max}$.

Similar to Zimmert and Seldin (2020), we use a hybrid regularizer based on a combination of the negative Tsallis entropy and the negative entropy, with separate learning rates

$$F_t(x) = -2\eta_t^{-1}\left(\sum_{i=1}^{K}\sqrt{x_i}\right) + \gamma_t^{-1}\left(\sum_{i=1}^{K} x_i(\log x_i - 1)\right), \tag{4.2}$$

where the learning rates are $\eta_t^{-1} = \sqrt{t}$ and $\gamma_t^{-1} = \sqrt{\frac{49\mathcal{D}_t}{\log K}}$. The update rule for obtaining the distribution over arms is

$$x_t = \nabla \bar{F}_t^*(-\hat{L}_t^{obs}) = \arg\min_{x \in \Delta^{K-1}} \langle \hat{L}_t^{obs}, x\rangle + F_t(x), \tag{4.3}$$

where $\hat{L}_t^{obs} = \sum_{s=1}^{t-1} \hat{\ell}_s \mathbb{1}(s + d_s < t)\mathbb{1}(s \notin \mathcal{S}_{t-1})$ is the cumulative importance-weighted loss estimate of observations that have arrived by time t and have not been skipped.

In the analysis we use $\mathcal{S} = \mathcal{S}_T$ to denote the final set of skipped rounds at time T and $\bar{\mathcal{S}} = [T] \setminus \mathcal{S}$ to denote its complement.

4.4 Regret Bounds

The following theorem provides best-of-both-worlds regret bounds for Algorithm 5. A proof is provided in Section 4.5 and a bound on S^* can be found in Appendix 4.7.6.

Theorem 4.1. *The pseudo-regret of Algorithm 5 for any sequence of delays and losses (where the losses are bounded in the* $[0, 1]$ *interval) satisfies*

$$\overline{Reg}_T = \mathcal{O}\left(\sqrt{KT} + \min_{\mathcal{S} \subseteq [T]}\left\{|\mathcal{S}| + \sqrt{\mathcal{D}_{\bar{\mathcal{S}}} K^{\frac{2}{3}} \log K}\right\} + K\hat{\sigma}_{\max}\right).$$

Furthermore, if the losses are stochastic, the pseudo-regret also satisfies

$$\overline{Reg}_T = \mathcal{O}\left(\sum_{i \neq i^*}\left(\frac{\log T}{\Delta_i} + \frac{\hat{\sigma}_{\max}}{\Delta_i \log K}\right) + K\hat{\sigma}_{\max} + S^*\right),$$

where $\hat{\sigma}_{\max} = \max_{t\in[T]}\{\hat{\sigma}_t\}$ *is the maximal number of outstanding observations after skipping, and it satisfies* $\hat{\sigma}_{\max} \leq \sigma_{\max}$*, and* S^* *is the number of rounds skipped by the algorithm, and it satisfies*

$$S^* = \mathcal{O}\left(\min\left(d_{\max}, \min_{\mathcal{S}\subseteq[T]}\left\{|\mathcal{S}| + \sqrt{\mathcal{D}_{\bar{\mathcal{S}}} K^{\frac{2}{3}} \log K}\right\}\right)\right).$$

Masoudian et al. (2022) provide $\Omega\left(\sqrt{KT} + \min_{\mathcal{S}\subset[T]}\left\{|\mathcal{S}| + \sqrt{\mathcal{D}_{\bar{\mathcal{S}}} \log K}\right\}\right)$ regret lower bound for adversarial environments with variable delays, which is matched within constants by the algorithm of (Zimmert and Seldin, 2020) for adversarial environments. Our algorithm matches the lower bound within a multiplicative factor of $K^{\frac{1}{3}}$ on the delay-dependent term, which is the price we pay for obtaining a best-of-both-worlds guarantee. It is an open question whether this factor can be reduced.

In the stochastic regime, assuming that the delays in the first $\sigma_{\max}$ rounds are of order T, and that the losses come from Bernoulli distributions with bias close to $\frac{1}{2}$, we obtain a trivial regret lower bound $\Omega\left(\sigma_{\max}\frac{\sum_{i\neq i*}\Delta_i}{K} + \sum_{i\neq i^*}\frac{\log T}{\Delta_i}\right)$. This bound is almost matched by the algorithm of Joulani et al. (2013) for the stochastic regime only, which achieves $\mathcal{O}\left(\sum_{i\neq i^*}\left(\frac{1}{\Delta_i}\log(T) + \sigma_{\max}\Delta_i\right)\right)$ regret bound. Our bound has some extra terms, most notably $\sum_{i\neq i^*}\frac{\hat{\sigma}_{\max}}{\Delta_i \log K}$ and S^*. It is an open question whether these terms can be reduced or whether it is possible to derive a best-of-both-worlds lower bound showing that this price is inevitable.

Theorem 4.1 provides three major improvements relative to the results of Masoudian et al. (2022): (1) it requires no advance knowledge of $d_{\max}$; (2) it replaces terms dependent on $d_{\max}$ by terms dependent on $\hat{\sigma}_{\max}$, which never exceeds $d_{\max}$, and in some cases may be significantly smaller; and (3) it makes skipping possible and beneficial, making the algorithm robust to a small number of excessively large delays and replacing $\sqrt{D\log K}$ term with $\min_{\mathcal{S}\subseteq[T]}\left\{|\mathcal{S}| + \sqrt{\mathcal{D}_{\bar{\mathcal{S}}} K^{\frac{2}{3}} \log K}\right\}$, which is never much larger, but in some cases significantly smaller.

4.5 Analysis

In this section, we present a proof of Theorem 4.1. We begin with the stochastic analysis in Section 4.5.1, followed by the adversarial analysis in Section 4.5.2.

4.5.1 Stochastic Analysis

Our stochastic analysis is based on the drift control lemma (Lemma 4.1). This lemma enables us to control the drift of the playing distribution using the time-varying hybrid regularizer. Unlike the drift control lemma used by Masoudian et al. (2022), which injects $d_{\max}$ into the learning rates, our lemma relies on the implicit exploration terms introduced in the loss estimators defined in equation (4.1) and on skipping of large delays. We present our key lemma below.

Lemma 4.1 (Drift Control Lemma)**.** *Let $d^t_{\max}$ be the skipping threshold at time t. Then, for any $i \in [K]$ and $s, t \in [T]$, where $s \leq t$ and $t - s \leq d^t_{\max}$, we have*

$$x_{t,i} \leq 4 \max(x_{s,i}, \lambda_{s,t}).$$

The complete proof of the lemma is presented in Appendix 4.7.2. Below we provide a sketch of the proof.

Proof sketch. By the FTRL update rule we know that $x_t = \nabla \bar{F}^*_t(-\hat{L}^{obs}_{t-1})$ and $x_s = \nabla \bar{F}^*_s(-\hat{L}^{obs}_{s-1})$. We define an auxiliary variable $\tilde{x}_s = \nabla \bar{F}^*_s(-\hat{L}^{obs}_{t-1})$ to bridge between x_t and x_s. It is based on the regularizer from round s and the loss estimate from round t. More precisely, we use induction on the pair (s, t) and achieve the bound $\frac{x_{t,i}}{\max(x_{s,i}, \lambda_{s,t})} \leq 4$ through the following two steps. The first bounds the deviation between x_t and $\tilde{x}_s$ due to the change of regularizer from F_t to F_s, and the second bounds the deviation between $\tilde{x}_s$ and x_s due to the change of loss estimate from $\hat{L}^{obs}_{t-1}$ to $\hat{L}^{obs}_{s-1}$.

Deviation induced by the change of regularizer: This step keeps the cumulative loss vector fixed, and investigates the deviation caused by the change of regularizer. We show that

$$\frac{x_{t,i}}{\max(\tilde{x}_{s,i}, \lambda_{s,t})} \leq 2.$$

The proof relies on the implicit exploration term $\lambda_{s,t}$ in the loss estimates. This term plays a crucial role in controlling the deviation caused by the change of regularizer.

Deviation induced by the loss shift: This step bounds the deviation caused by the change of the cumulative loss estimate while keeping the regularizer fixed. We prove the following inequality:

$$\frac{\tilde{x}_{s,i}}{x_{s,i}} \leq 2.$$

The step is based on induction and the skipping procedure of the algorithm.

The second step establishes that $\frac{\max(\tilde{x}_i, \lambda_{s,t})}{\max(x_{s,i}, \lambda_{s,t})} \leq 2$. When combined with the result from the first step, it completes the proof. □

We then proceed by defining the drifted regret as

$$\overline{Reg}_T^{drift} = \mathbb{E}\left[\sum_{t=1}^{T}\left(\langle x_t, \hat{\ell}_t^{obs}\rangle - \hat{\ell}_{t,i_T^*}^{obs}\right)\right], \tag{4.4}$$

where $\hat{\ell}_t^{obs} = \sum_{s=1}^{t} \hat{\ell}_s \mathbb{1}(s+\hat{d}_s = t)\mathbb{1}(s \notin \mathcal{S}_t)$ is the cumulative vector of losses received at time t. To establish a relationship between $\overline{Reg}_T^{drift}$ and the actual regret $\overline{Reg}_T$, we provide Lemma 4.2 that measures the drift of the drifted regret from the actual one. Later in this section, we provide the proof, which is based on the drift control lemma (Lemma 4.1).

Lemma 4.2 (Drift of the Drifted Regret)**.** *Let $\sigma_{\max}^t = \max_{s\in[t]}\{\hat{\sigma}_s\}$. Then, we have the following lower bound for the drifted regret*

$$\overline{Reg}_T^{drift} \geq \frac{1}{4}\overline{Reg}_T - \frac{\sigma_{\max}}{4} - 2K\sum_{t=1}^{T}\left(\lambda_{t,t+\hat{d}_t} + \lambda_{t,t+\hat{d}_t+\sigma_{\max}^t}\right) - S^*.$$

We can use the standard FTRL analysis and the drift control lemma, similar to Masoudian et al. (2022), to obtain an upper bound for $\overline{Reg}_T^{drift}$. Specifically, in Appendix 4.7.1 we show that

$$\begin{aligned}\overline{Reg}_T^{drift} \leq\ & \mathbb{E}\left[a\sum_{t=1}^{T}\sum_{i\neq i^*}\eta_t x_{t,i}^{1/2} + b\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t}-1)x_{t,i}\Delta_i\right] \\ & + \mathbb{E}\left[c\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\hat{\sigma}_t\gamma_t x_{t,i}\log(1/x_{t,i})}{\log K}\right] + \mathcal{O}\left(K\sum_{t=1}^{T}\lambda_{t,t+\hat{d}_t} + S^*\right),\end{aligned} \tag{4.5}$$

where $a, b, c \geq 0$ are some constants, and for any time $t \in [T]$, $\upsilon_t = \sum_{s=1}^{t}\mathbb{1}\left(s+\hat{d}_s = t\right)$ is the number of arrivals at time t (if a round s is skipped

at time t it counts as an "empty" arrival with its loss estimate set to zero), and S^* is the total number of rounds skipped by the algorithm. By applying Lemma 4.2 to (4.5), we achieve the following regret bound

$$\begin{aligned}\overline{Reg}_T \leq \mathbb{E}\left[2a\sum_{t=1}^{T}\sum_{i\neq i^*}\eta_t x_{t,i}^{1/2} + 2b\sum_{t=1}^{T}\sum_{i\neq i^*}\gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t}-1)x_{t,i}\Delta_i\right] \\ + \mathbb{E}\left[2c\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\hat{\sigma}_t\gamma_t x_{t,i}\log(1/x_{t,i})}{\log K}\right] \\ + \mathcal{O}\left(K\sum_{t=1}^{T}\left(\lambda_{t,t+\hat{d}_t} + \lambda_{t,t+\hat{d}_t+\sigma^t_{\max}}\right) + \sigma_{\max} + S^*\right). \end{aligned} \tag{4.6}$$

Now we apply a self-bounding analysis, similar to Masoudian et al. (2022), and get

$$\overline{Reg}_T = \mathcal{O}\left(\sum_{i\neq i^*}\left(\frac{1}{\Delta_i}\log(T) + \frac{\sigma_{\max}}{\Delta_i\log K}\right) + \sigma_{\max} + K\sum_{t=1}^{T}\left(\lambda_{t,t+\hat{d}_t} + \lambda_{t,t+\hat{d}_t+\sigma^t_{\max}}\right) + S^*\right).$$

The details of the self-bounding analysis are provided in Appendix 4.7.3. To complete the analysis for the stochastic regime, we also need to bound the sum of the implicit exploration terms. This bound is provided in Lemma 4.3. This lemma is the second key result of the paper, because it shows that the bias introduced by implicit exploration does not deteriorate the bounds. The proof is based on a careful study of the evolution of $\mathcal{D}_t$ throughout the game, and is deferred to Appendix 4.7.4.

Lemma 4.3 (Summation Bound)**.** *For all $s \in [T]$, let $\mathcal{D}_s = \sum_{r=1}^{s}\hat{\sigma}_r$, then we have*

$$\sum_{t=1}^{T} e^{-\frac{\mathcal{D}_{t+\hat{d}_t}}{\mathcal{D}_{t+\hat{d}_t}-\mathcal{D}_t}} + e^{-\frac{\mathcal{D}_{t+\hat{d}_t+\sigma^t_{\max}}}{\mathcal{D}_{t+\hat{d}_t+\sigma^t_{\max}}-\mathcal{D}_t}} = \mathcal{O}(\hat{\sigma}_{\max}).$$

Proof of the Drifted Regret Lemma

We start with the definition of the drifted regret.

$$\begin{aligned}
\overline{Reg}_T^{drift} &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\langle x_t, \hat{\ell}_t^{obs}\rangle - \hat{\ell}_{t,i_T^*}^{obs}\right)\right] \\
&= \sum_{t=1}^{T}\sum_{s:s+\hat{d}_s=t}\sum_{i=1}^{K}\mathbb{E}\left[\left(\frac{\ell_{s,i}x_{s,i}x_{t,i}}{\max\{x_{s,i},\lambda_{s,t}\}} - \frac{\ell_{s,i_T^*}x_{s,i_T^*}x_{t,i}}{\max\{x_{s,i_T^*},\lambda_{s,t}\}}\right)\mathbb{1}(s\notin\mathcal{S}_t)\right] \\
&\geq \sum_{t=1}^{T}\sum_{s:s+\hat{d}_s=t}\sum_{i=1}^{K}\mathbb{E}\left[\left(\underbrace{\frac{\ell_{s,i}x_{s,i}x_{t,i}}{\max\{x_{s,i},\lambda_{s,t}\}}}_{\star} - \ell_{s,i_T^*}x_{t,i}\right)\mathbb{1}(s\notin\mathcal{S}_t)\right]. \qquad (4.7)
\end{aligned}$$

Note that when taking the expectation, we rely on the fact that $\hat{\ell}_s$ with $s+\hat{d}_s=t$ does not affect x_t. If $\max\{x_{s,i},\lambda_{s,t}\}=x_{s,i}$, then $\star=\ell_{s,i}x_{t,i}$, otherwise

$$\begin{aligned}
\star = \frac{\ell_{s,i}x_{s,i}x_{t,i}}{\lambda_{s,t}} &= \ell_{s,i}x_{t,i} - \frac{\ell_{s,i}x_{t,i}\left(\lambda_{s,t}-x_{s,i}\right)}{\lambda_{s,t}} \\
&\geq \ell_{s,i}x_{t,i} - \frac{4\lambda_{s,t}(\lambda_{s,t}-x_{s,i})}{\lambda_{s,t}} \geq \ell_{s,i}x_{t,i} - 4\lambda_{s,t}, \qquad (4.8)
\end{aligned}$$

where the first inequality uses $x_{t,i}\leq 4\max(x_{s,i},\lambda_{s,t})=4\lambda_{s,t}$ by Lemma 4.1, and $\ell_{s,i}\geq 1$, and the second inequality follows by $x_{s,i}\geq 0$. Plugging (4.8) into (4.7) gives

$$\begin{aligned}
\overline{Reg}_T^{drift} &\geq \sum_{t=1}^{T}\sum_{s:s+\hat{d}_s=t}\sum_{i=1}^{K}\mathbb{E}\left[\left(\ell_{s,i}x_{t,i} - 4\lambda_{s,t} - \ell_{s,i_T^*}x_{t,i}\right)\mathbb{1}(s\notin\mathcal{S}_t)\right] \\
&\geq \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\sum_{s:s+\hat{d}_s=t}\sum_{i=1}^{K}\Delta_i x_{t,i}\right]}_{R_T} - S^* - 4K\sum_{t=1}^{T}\sum_{s:s+\hat{d}_s=t}\mathbb{E}\left[\lambda_{s,t}\right]. \qquad (4.9)
\end{aligned}$$

It suffices to give a lower bound for R_T in terms of the actual regret $\overline{Reg}_T$. The difference between R_T and $\overline{Reg}_T$ is that in R_T the coefficient behind $\sum_{i=1}^{K}\Delta_i x_{t,i}$ is the number of arrivals $\upsilon_t=\sum_{s=1}^{t}\mathbb{1}\left(s+\hat{d}_s=t\right)$ at time t, and υ_t might be larger than one due to delays. Our main idea here is to leverage the drift control lemma to rearrange the arrivals. Specifically, by Lemma 4.1 for all $r\in[0,d_{\max}^t]$, we have

$\max(x_{t,i}, \lambda_{t,t+r}) \geq \frac{1}{4} x_{t+r,i}$, which implies $x_{t,i} \geq \frac{1}{4} x_{t+r,i} - \lambda_{t,t+r}$. Thus, we obtain the following bound for any $r \in [0, d^t_{\max}]$

$$\sum_{i=1}^{K} \Delta_i x_{t,i} \geq \frac{1}{4} \sum_{i=1}^{K} \Delta_i x_{t+r,i} - K\lambda_{t,t+r}. \tag{4.10}$$

Hence, if we observe more than one arrival at round t (i.e., $\upsilon_t \geq 1$), we can rearrange the arrivals to ensure that almost all rounds receive at least $\frac{1}{4}$ of an arrival with some additional cost from implicit exploration terms. To achieve this, we may push some of the arrivals forward to future rounds. When we push an arrival s from a current round t to round $t+r$ using (4.10), it is equivalent to replacing $\sum_{i=1}^{K} \Delta_i x_{t,i}$ by $\frac{1}{2}\sum_{i=1}^{K} \Delta_i x_{t+r,i} - K\lambda_{t,t+r}$ in R_T. Note that with this method, we may push an arrival to a round that is bigger than T which is equivalent to replacing $\sum_{i=1}^{K} \Delta_i x_{t,i}$ by zero. We propose Algorithm 6 that provides a greedy way to rearrange the arrivals. It pushes each arrival to the first available (unoccupied) round.

Algorithm 6: Greedy Rearrangement

```
Initialize υ_t^new = 0 for all t = 1, ..., T + d_max^T
for t = 1, ..., T do
    for s = 1, ..., t : s + d̂_s = t do
        Find the first round π(s) ∈ [t, t + d_max^t] such that υ_{π(s)}^new = 0
        Move the arrival from round s to round π(s) and update υ_{π(s)}^new = 1/4
```

Let υ_t^{new} for all $t \in [T + d^T_{\max}]$ be the total arrivals at time t after the rearrangement, and let $\pi(t)$ be the round to which we have mapped round t for all $t \in [T]$. Then the following inequality holds for any rearrangement

$$R_T = \sum_{t=1}^{T} \upsilon_t \sum_{i=1}^{K} \Delta_i x_{t,i} \geq \sum_{t=1}^{T} \upsilon_t^{new} \sum_{i=1}^{K} \Delta_i x_{t,i} - K \sum_{t=1}^{T} \lambda_{t,\pi(t)}. \tag{4.11}$$

We provide properties of the greedy rearrangement in Lemma 4.4.

Lemma 4.4. *Let $\sigma^t_{\max} = \max_{s\in[t]} \{\hat{\sigma}_s\}$. Then for any round t, Algorithm 6 keeps all arrivals at time t in the interval $[t, t + \sigma^t_{\max}]$, such that $\forall s \leq t : s + \hat{d}_s = t$, we have $\pi(s) - t \leq \sigma^t_{\max}$ and $\upsilon^{new}_{\pi(s)} \in \{0, \frac{1}{4}\}$.*

We provide a proof of this lemma in Appendix 4.7.5. Using this lemma, we have

$$\begin{aligned}\sum_{t=1}^{T} \upsilon_t^{new} \sum_{i=1}^{K} \Delta_i x_{t,i} &= \frac{1}{4}\sum_{t=1}^{T}\sum_{i=1}^{K} \Delta_i x_{t,i} - \frac{1}{4}\sum_{t=1}^{T} \mathbb{1}(\upsilon_t^{new}=0) \sum_{i=1}^{K} \Delta_i x_{t,i} \\ &\leq \frac{1}{4}\overline{Reg}_T - \frac{1}{4}\sigma_{\max}^T \leq \frac{1}{4}\overline{Reg}_T - \frac{1}{4}\sigma_{\max}, \end{aligned} \tag{4.12}$$

where the second inequality uses the fact after any rearrangement $\sum_{t=1}^{T} \upsilon_t = 4\sum_{t=1}^{T+\sigma_{\max}} \upsilon_t^{new}$, and as a result the number of rounds with zero arrivals will be $\sigma_{\max}$. Since $\forall t \in [T] : \pi(t) \leq t + \hat{d}_t + \sigma_{\max}^t$ then $\lambda_{t,\pi(t)} \leq \lambda_{t,t+\hat{d}_t+\sigma_{\max}^t}$. So, this together with (4.12), (4.11), and (4.9) completes the proof.

4.5.2 Adversarial Analysis

For the adversarial regret bound we have

$$\begin{aligned}\overline{Reg}_T &\leq 4\sqrt{KT} + \sum_{t=1}^{T} \gamma_t \hat{\sigma}_t + \gamma_T^{-1} \log K + S^* + K\sum_{t=1}^{T} \lambda_{t,t+\hat{d}_t} \\ &\leq \mathcal{O}\left(\sqrt{KT} + \sqrt{\mathcal{D}_T \log K} + S^* + K\hat{\sigma}_{\max}\right), \end{aligned} \tag{4.13}$$

where the first four terms on the right hand side of the first inequality is a bound on the regret in the non-skipped rounds, which follows by Zimmert and Seldin (2020, Theorem 3), since the structure of the algorithm is identical, the last term is the bias introduced by implicit exploration. We provide details in Appendix 4.7.8. The second inequality holds by the choice of the learning rates $\{\gamma_t\}_{t\in[T]}$ and Lemma 4.3. Since our skipping rule differs, we need to revise the bound of Zimmert and Seldin (2020) on $\sqrt{\mathcal{D}_T \log K} + S^*$.

Lemma 4.5. *We have*

$$S^* + \sqrt{\mathcal{D}_T \log(K)} \leq \mathcal{O}\left(\max\left\{K^{\frac{2}{3}}\log(K), \min_{\mathcal{S}\subset[T]}\left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}} K^{\frac{2}{3}}\log(K)}\right)\right\}\right).$$

A proof of the lemma can be found in Appendix 4.7.8. The lemma completes the proof of the adversarial bound in Theorem 4.1.

4.6 Discussion

We have presented a best-of-both-worlds algorithm for bandits with delayed feedback. The algorithm is based on a careful augmentation of the adversarial algorithm of Zimmert and Seldin (2020) with implicit exploration and adjusted skipping. The result improves on prior work of Masoudian et al. (2022) by eliminating the need in prior knowledge of $d_{\max}$, replacing terms dependent on $d_{\max}$ with terms dependent on $\hat{\sigma}_{\max}$, which is generally better, and benefiting from skipping. To the best of our knowledge, it is also the first use of implicit exploration in the context of best-of-both-worlds bounds. In particular, we manage to control the implicit exploration bias, so that it does not deteriorate neither the stochastic, nor the adversarial bound.

The work leads to several directions for future research. One is whether the best-of-both-worlds bounds could be improved further, in particular, whether it is possible to reduce the $K^{\frac{1}{3}}$ term in the adversarial bound and $\sum_{i \neq i^*} \frac{\hat{\sigma}_{\max}}{\Delta_i \log K}$ and S^* terms in the stochastic bound, or whether it is possible to derive lower bounds demonstrating that best-of-both-worlds bounds for bandits with delayed feedback must bear extra costs. Another interesting direction is to find additional applications for implicit exploration in the context of best-of-both-worlds bounds.

4.7 Appendix

4.7.1 Details of the Drifted Regret Analysis

In this section we prove the bound on drifted regret in equation (4.5). The derivation is same as the one by Masoudian et al. (2022), however, for the sake of completeness we reproduce it here. The analysis follows the standard FTRL approach, decomposing the drifted pseudo-regret into *penalty* and *stability* terms as

$$\overline{Reg}_T^{drift} = \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \langle x_t, \hat{\ell}_t^{obs}\rangle + \bar{F}_t^*(-\hat{L}_{t+1}^{obs}) - \bar{F}_t^*(-\hat{L}_t^{obs})}_{stability}\right] + \mathbb{E}\left[\underbrace{\sum_{t=1}^{T} \bar{F}_t^*(-\hat{L}_t^{obs}) - \bar{F}_t^*(-\hat{L}_{t+1}^{obs}) - \ell_{t,i_T^*}}_{penalty}\right].$$

The penalty term is bounded by the following inequality, derived by Abernethy et al. (2015)

$$penalty \leq \sum_{t=2}^{T} (F_{t-1}(x_t) - F_t(x_t)) + F_T(\mathrm{e}_{i_T^*}) - F_1(x_1), \tag{4.14}$$

where $\mathrm{e}_{i_T^*}$ represents the unit vector in $\mathbb{R}^K$ with the i_T^*-th element being one and zero elsewhere. This leads to the following bound for penalty term

$$penalty \leq \mathcal{O}\left(\sum_{t=2}^{T}\sum_{i\neq i^*} \eta_t x_{t,i}^{\frac{1}{2}} + \sum_{t=2}^{T}\sum_{i=1}^{K} \frac{\sigma_t \gamma_t x_{t,i} \log(1/x_{t,i})}{\log K}\right), \tag{4.15}$$

where we substitute the explicit form of the regularizer into (4.14) and exploit the properties $\eta_t^{-1} - \eta_{t-1}^{-1} = \mathcal{O}(\eta_t)$, $\gamma_t^{-1} - \gamma_{t-1}^{-1} = \mathcal{O}(\sigma_t\gamma_t/\log K)$, and $x_{t,i_T^*}^{\frac{1}{2}} - 1 \leq 0$.

For the stability term, following a similar analysis as presented by Masoudian et al. (2022, Lemma 5), but incorporating implicit exploration terms, for any $\alpha_t \leq \gamma_t^{-1}$ we obtain

$$stability \leq \sum_{t=1}^{T}\sum_{i=1}^{K} 2f_t''(x_{t,i})^{-1}(\hat{\ell}_{t,i}^{obs} - \alpha_t)^2.$$

Let $A_t = \left\{s \leq t : s + \hat{d}_s = t\right\}$, then due to the choice of skipping threshold, $\alpha_t = \sum_{s\in A_t} \bar{\ell}_{s,t}$ satisfies the condition $\alpha_t \leq \gamma_t^{-1}$, where $\bar{\ell}_{s,t} = \frac{\sum_{i=1}^K f_t''(x_{t,i})^{-1}\hat{\ell}_{s,i}}{\sum_{i=1}^K f_t''(x_{t,i})^{-1}} =$

$\frac{f_t''(x_{t,I_s})^{-1}\hat{\ell}_{s,I_s}}{\sum_{i=1}^K f_t''(x_{t,i})^{-1}}$. Thus we have

$$\begin{aligned}
\textit{stability} &\leq \sum_{t=1}^{T}\sum_{i=1}^{K} 2f_t''(x_{t,i})^{-1}\left(\sum_{s\in A_t}\hat{\ell}_{s,i}-\bar{\ell}_{s,t}\right)^2 \\
&= \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{s\in A_t} 2f_t''(x_{t,i})^{-1}\left(\hat{\ell}_{s,i}-\bar{\ell}_{s,t}\right)^2}_{S_1} \\
&\quad + \underbrace{\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{r,s\in A_t, r\neq s} 2f_t''(x_{t,i})^{-1}\left(\hat{\ell}_{s,i}-\bar{\ell}_{s,t}\right)\left(\hat{\ell}_{r,i}-\bar{\ell}_{r}\right)}_{S_2}
\end{aligned}$$

For brevity we define $z_{t,i} = f_t''(x_{t,i})^{-1}$ and $m_{s,i}^t = \max\{x_{s,i}, \lambda_{s,t}\}$ for any $s \leq t$ and $i \in [K]$. We begin bounding S_1 by replacing definition of loss estimators from (4.1) and get

$$\begin{aligned}
\mathbb{E}[S_1] &= \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{s\in A_t} 2\mathbb{E}\left[z_{t,i}\left(\frac{\ell_{s,I_s}\mathbb{1}(I_s=i)}{m_{s,i}^t} - \frac{z_{t,I_s}\ell_{s,I_s}}{m_{s,I_s}^t\sum_{j=1}^K z_{t,j}}\right)^2\right] \\
&\leq \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{s\in A_t} 2\mathbb{E}\left[z_{t,i}\left(\frac{\mathbb{1}(I_s=i)}{m_{s,i}^t} - \frac{z_{t,I_s}}{m_{s,I_s}^t\sum_{j=1}^K z_{t,j}}\right)^2\right] \\
&= \sum_{t=1}^{T}\sum_{s\in A_t}\underbrace{2\sum_{i=1}^{K}\mathbb{E}\left[z_{t,i}\left(\frac{\mathbb{1}(I_s=i)}{{m_{s,i}^t}^2} - \frac{z_{t,I_s}\mathbb{1}(I_s=i)}{m_{s,i}^t m_{s,I_s}^t\sum_{j=1}^K z_{t,j}}\right)\right]}_{S_1^1} \\
&\quad + \sum_{t=1}^{T}\sum_{s\in A_t}\underbrace{2\,\mathbb{E}\left[\left(\frac{z_{t,I_s}^2}{{m_{s,I_s}^t}^2(\sum_{j=1}^K z_{t,j})} - \sum_{i=1}^{K}\frac{z_{t,I_s}z_{t,i}\mathbb{1}(I_s=i)}{m_{s,i}^t m_{s,I_s}^t\sum_{j=1}^K z_{t,j}}\right)\right]}_{S_1^2}
\end{aligned}$$

Where the first inequality uses $\ell_{s,I_s} \leq 1$. We show that S_1^2 has negative contribution to S_1 by taking expectation w.r.t. I_s as the following

$$S_1^2 = \sum_{t=1}^{T}\sum_{s\in A_t}\mathbb{E}\left[\sum_{i=1}^{K}\frac{z_{t,i}^2 x_{s,i}}{{m_{s,i}^t}^2(\sum_{j=1}^K z_{t,j})} - \sum_{i=1}^{K}\frac{z_{t,i}^2 x_{s,i}}{{m_{s,i}^t}^2\sum_{j=1}^K z_{t,j}}\right] = 0$$

Thus we only need to bound S_1^1, for which we take expectation w.r.t. I_s and separate i^* from the other arms to get

$$\begin{aligned}
S_1^1 &= \sum_{i=1}^{K} \mathbb{E}\left[z_{t,i} \left(\frac{\mathbb{1}(I_s = i)}{{m_{s,i}^t}^2} - \frac{z_{t,I_s}\mathbb{1}(I_s = i)}{m_{s,i}^t m_{s,I_s}^t \sum_{j=1}^{K} z_{t,j}} \right) \right] \\
&\leq \sum_{i \neq i^*} \mathbb{E}\left[\frac{z_{t,i} x_{s,i}}{{m_{s,i}^t}^2} \right] + \mathbb{E}\left[\frac{z_{t,i^*} x_{s,i^*}}{{m_{s,i^*}^t}^2} - \frac{z_{t,i^*}^2 x_{s,i^*}}{{m_{s,i^*}^t}^2 \sum_{j=1}^{K} z_{t,j}} \right] \\
&\leq \sum_{i \neq i^*} \mathbb{E}\left[4\eta_t x_{s,i}^{1/2} \right] + \mathbb{E}\left[\frac{x_{s,i^*}}{{m_{s,i^*}^t}^2} \times z_{t,i^*} \left(1 - \frac{z_{t,i^*}}{\sum_{j=1}^{K} z_{t,j}} \right) \right] \\
&\leq \sum_{i \neq i^*} 4\mathbb{E}\left[\eta_t x_{s,i}^{1/2} \right] + \mathbb{E}\left[\frac{x_{s,i^*}}{{m_{s,i^*}^t}^2} \times \eta_t x_{t,i^*}^{3/2} \left(1 - \frac{x_{t,i^*}^{3/2}}{(1 - x_{t,i^*})^{3/2} + x_{t,i^*}^{3/2}} \right) \right] \\
&\leq \sum_{i \neq i^*} 4\mathbb{E}\left[\eta_t x_{s,i}^{1/2} \right] + \mathbb{E}\left[\frac{\eta_t x_{s,i^*} x_{t,i^*}^{3/2}}{{m_{s,i^*}^t}^2} \times \left(\frac{(1 - x_{t,i^*})^{3/2}}{2^{-1/2}} \right) \right] \\
&\leq \sum_{i \neq i^*} 4\mathbb{E}\left[\eta_t x_{s,i}^{1/2} \right] + \mathbb{E}\left[4\sqrt{2}\eta_t \sum_{i \neq i^*} x_{t,i} \right] \\
&\leq \sum_{i \neq i^*} 4\mathbb{E}\left[\eta_t x_{s,i}^{1/2} \right] + \mathbb{E}\left[16\sqrt{2}\eta_t \sum_{i \neq i^*} (x_{s,i} + \lambda_{s,t}) \right] \\
&\leq \mathcal{O}\left(\mathbb{E}\left[\eta_s \sum_{i \neq i^*} x_{s,i}^{1/2} \right] + \mathbb{E}\left[K\lambda_{s,t} \right] \right),
\end{aligned}$$

where the second inequality uses $z_{t,i} = f_t''(x_{t,i})^{-1} \leq \eta_t x_{t,i}^{3/2}$ along $x_{t,i} \leq m_{s,i}^t$ from Lemma 4.1, the third inequality is due the fact that $z_{t,i^*}\left(1 - \frac{z_{t,i^*}}{\sum_{j=1}^{K} z_{t,j}}\right)$ is an increasing function in terms of both z_{t,i^*} and $\sum_{i \neq i^*} z_{t,i}$ and we substitute $z_{t,i^*} \leq \eta_t x_{t,i^*}^{3/2}$ and $\sum_{j \neq i^*} z_{t,j} \leq \sum_{j \neq i^*} \eta_t x_{t,j}^{3/2} \leq \eta_t (1 - x_{t,i^*})^{3/2}$, the fourth inequality is due to $(1-a)^{3/2} + a^{3/2} \leq 2^{-1/2}$, the fifth and the sixth inequalities rely on Lemma 4.1, and finally the last inequality is followed by $\forall i : x_{s,i} \leq x_{s,i}^{1/2}$ and that $\eta_t \leq \eta_s$. Combining bounds for S_1^1 and S_1^2 gives the following bound for S_1

$$\mathbb{E}[S_1] \leq \mathcal{O}\left(\sum_{t=1}^{T} \sum_{i \neq i^*} \eta_t \mathbb{E}[x_{t,i}^{1/2}] + \sum_{t=1}^{T} K\lambda_{t,t+\hat{d}_t} \right) \tag{4.16}$$

For S_2, we take expectation with respect to I_s, I_r, and randomness of losses, all separately to get

$$\mathbb{E}[S_2] = \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{r,s\in A_t, r\neq s} 2\mathbb{E}\left[z_{t,i}\left(\hat{\ell}_{s,i}-\bar{\ell}_s\right)\left(\hat{\ell}_{r,i}-\bar{\ell}_s\right)\right]$$
$$= \sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{r,s\in A_t, r\neq s} 2\mathbb{E}\left[z_{t,i}\left(\frac{\mu_i x_{s,i}}{m^t_{s,i}} - \frac{\sum_{j=1}^K z_{t,j}\mu_j x_{s,j}/m^t_{s,j}}{\sum_{j=1}^K z_{t,j}}\right)\left(\frac{\mu_i x_{r,i}}{m^t_{r,i}} - \frac{\sum_{j=1}^K z_{t,j}\mu_j x_{r,j}/m^t_{r,j}}{\sum_{j=1}^K z_{t,j}}\right)\right]. \tag{4.17}$$

For simplicity we define $\epsilon^t_{s,i} = \mu_i - \frac{\mu_i x_{s,i}}{m^t_{s,i}}$ for any $s \le t$ and any $i \in [K]$, for which we have the following bounds

$$0 \le \epsilon^t_{s,i} \le \frac{\lambda_{s,t}}{m^t_{s,i}}.$$

We then continue from 4.17 and utilize the following decomposition

$$\sum_{i=1}^{K} 2\mathbb{E}\left[z_{t,i}\left(\mu_i - \frac{\sum_{j=1}^K z_{t,j}\mu_j}{\sum_{j=1}^K z_{t,j}} - \epsilon^t_{s,i} + \frac{\sum_{j=1}^K z_{t,j}\epsilon^t_{s,j}}{\sum_{j=1}^K z_{t,j}}\right)\left(\mu_i - \frac{\sum_{j=1}^K z_{t,j}\mu_j}{\sum_{j=1}^K z_{t,j}} - \epsilon^t_{r,i} + \frac{\sum_{j=1}^K z_{t,j}\epsilon^t_{r,j}}{\sum_{j=1}^K z_{t,j}}\right)\right]$$
$$\le 2\mathbb{E}\left[\underbrace{\sum_{i=1}^{K} z_{t,i}\left(\mu_i - \frac{\sum_{j=1}^K z_{t,j}\mu_j}{\sum_{j=1}^K z_{t,j}}\right)^2}_{S_2^1}\right]$$
$$+ 2\mathbb{E}\left[\underbrace{\sum_{i=1}^{K} z_{t,i}\epsilon^t_{s,i}\epsilon^t_{r,i} + 2z_{t,i}(\epsilon^t_{s,i}+\epsilon^t_{r,i})}_{S_2^2}\right]$$
$$+ 2\mathbb{E}\left[\underbrace{\frac{(\sum_{i=1}^K z_{t,i}\epsilon^t_{s,i})(\sum_{i=1}^K z_{t,i}\epsilon^t_{r,i})}{\sum_{i=1}^K z_{t,i}}}_{S_2^3}\right], \tag{4.18}$$

where the inequality holds because we ignore the negative terms after multiplication and that $|(\mu_i - \frac{\sum_{j=1}^K z_{t,j}\mu_j}{\sum_{j=1}^K z_{t,j}})| \le 1$. We need to bound each part from (4.18). We start

with S_2^1,

$$\begin{aligned}
S_2^1 &= \sum_{i=1}^{K} z_{t,i} \left(\mu_i - \frac{\sum_{j=1}^{K} z_{t,j}\mu_j}{\sum_{j=1}^{K} z_{t,j}} \right)^2 \\
&= \sum_{i=1}^{K} z_{t,i}\mu_i^2 - \frac{\left(\sum_{i=1}^{K} z_{t,i}\mu_i\right)^2}{\sum_{i=1}^{K} z_{t,i}} \\
&\leq \sum_{i=1}^{K} z_{t,i}\mu_i^2 - \frac{\left(\sum_{i=1}^{K} z_{t,i}\mu_{i^*}\right)^2}{\sum_{i=1}^{K} z_{t,i}} \\
&\leq \sum_{i=1}^{K} z_{t,i}(\mu_i^2 - \mu_{i^*}^2) \\
&\leq \sum_{i \neq i^*} 2\gamma_t x_{t,i}\Delta_i
\end{aligned} \tag{4.19}$$

We bound S_2^2 as

$$\begin{aligned}
S_2^2 &= \sum_{i=1}^{K} z_{t,i}\epsilon_{s,i}^t\epsilon_{r,i}^t + 2z_{t,i}(\epsilon_{s,i}^t + \epsilon_{r,i}^t) \\
&\leq \sum_{i=1}^{K} z_{t,i}\frac{\epsilon_{s,i}^t + \epsilon_{r,i}^t}{2} + 2z_{t,i}(\epsilon_{s,i}^t + \epsilon_{r,i}^t) \\
&\leq \frac{5}{2}\sum_{i=1}^{K} \frac{z_{t,i}\lambda_{s,t}}{m_{s,i}^t} + \frac{z_{t,i}\lambda_{r,t}}{m_{r,i}^t} \\
&\leq \frac{5}{2}K\gamma_t(\lambda_{s,t} + \lambda_{r,t}),
\end{aligned} \tag{4.20}$$

where the last inequality holds because $z_{t,i} \leq \gamma_t x_{t,i}$ and that $x_{t,i} \leq 4m_{s,i}^t, 4m_{r,i}^t$ from Lemma 4.1.
It remains to give upper bound for S_2^3 as

$$\begin{aligned}
S_2^3 &= \frac{(\sum_{i=1}^{K} z_{t,i}\epsilon_{s,i}^t)(\sum_{i=1}^{K} z_{t,i}\epsilon_{r,i}^t)}{\sum_{i=1}^{K} z_{t,i}} \\
&\leq \frac{(\sum_{i=1}^{K} z_{t,i}\lambda_{s,t}/m_{s,i}^t)(\sum_{i=1}^{K} z_{t,i}\lambda_{r,t}/m_{r,i}^t)}{\sum_{i=1}^{K} z_{t,i}} \\
&\leq \frac{1}{2}K\gamma_t(\lambda_{s,t} + \lambda_{r,t}),
\end{aligned} \tag{4.21}$$

where the second inequality rely on $z_{t,i} \leq \gamma_t x_{t,i}$, $\lambda_{s,t} \leq m^t_{s,i}$, $\lambda_{r,t} \leq m^t_{r,i}$, and $x_{t,i} \leq 4m^t_{s,i}, x_{t,i} \leq 4m^t_{r,i}$ from Lemma 4.1. It is suffices to plug bounds in (4.19), (4.20), and (4.21) to obtain

$$\begin{aligned}
\mathbb{E}[S_2] &\leq \sum_{t=1}^{T} \sum_{i \neq i^*} 4\Delta_i \gamma_t \mathbb{E}[x_{t,i}] \upsilon_t(\upsilon_t - 1) + 6 \sum_{t=1}^{T} K \gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t} - 1)\lambda_{t,t+\hat{d}_t} \\
&\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \sum_{s \in A_t} 4\Delta_i \gamma_t \mathbb{E}[x_{s,i} + \lambda_{s,t}](\upsilon_t - 1) + 6 \sum_{t=1}^{T} K \gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t} - 1)\lambda_{t,t+\hat{d}_t} \\
&\leq \sum_{t=1}^{T} \sum_{i \neq i^*} \sum_{s \in A_t} 4\Delta_i \gamma_t \mathbb{E}[x_{s,i}](\upsilon_t - 1) + 10 \sum_{t=1}^{T} K \gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t} - 1)\lambda_{t,t+\hat{d}_t} \\
&\leq \mathcal{O}\left(\sum_{t=1}^{T} \sum_{i \neq i^*} \gamma_{t+\hat{d}_t} \Delta_i \mathbb{E}[x_{t,i}](\upsilon_{t+\hat{d}_t} - 1) + K \sum_{t=1}^{T} \lambda_{t,t+\hat{d}_t} \right), \qquad (4.22)
\end{aligned}$$

where the third inequality uses Lemma 4.1 and the last inequality holds because of the skipping that ensures $\gamma_{t+\hat{d}_t}(\upsilon_{t+\hat{d}_t} - 1) \leq 1$. Now, it is sufficient to combine the bounds for S_1 and S_2 in (4.16) and (4.22) and get

$$\mathbb{E}[\textit{stability}] \leq \mathcal{O}\left(\sum_{t=1}^{T} \sum_{i \neq i^*} \eta_t \mathbb{E}[x^{1/2}_{t,i}] + \sum_{t=1}^{T} \sum_{i \neq i^*} \gamma_{t+\hat{d}_t} \mathbb{E}[x_{t,i}](\upsilon_{t+\hat{d}_t} - 1) + K \sum_{t=1}^{T} \lambda_{t,t+\hat{d}_t} \right). \qquad (4.23)$$

Combining the stability bound from (4.23) and the penalty bound from (4.15) concludes the proof.

4.7.2 Proof of the Drift Control Lemma

In this section we provide a proof of Lemma 4.1. We start with a few auxiliary results, and then prove the lemma.

4.7.2.1 Auxiliary results for the proof of the key lemma

For the proof we use two facts and a lemma from Masoudian et al. (2022), and a new lemma. Recall that $f_t(x) = -2\eta_t^{-1}\sqrt{x} + \gamma_t^{-1} x(\log x - 1)$.

Fact 4.2. *(Masoudian et al., 2022, Fact 15) $f_t'(x)$ is a concave function.*

Fact 4.3. *(Masoudian et al., 2022, Fact 16) $f_t''(x)^{-1}$ is a convex function.*

Lemma 4.6. *(Masoudian et al., 2022, Lemma 17) Fix t and s with $t \geq s$, and assume that there exists α, such that $x_{t,i} \leq \alpha \max(x_{s,i}, \lambda_{s,t})$ for all $i \in [K]$, and let $f(x) = \big(-2\eta_t^{-1}\sqrt{x} + \gamma_t^{-1}x(\log x - 1)\big)$, then we have the following inequality*

$$\frac{\sum_{j=1}^{K} f''(x_{t,j})^{-1}\hat{\ell}_{s,j}}{\sum_{j=1}^{K} f''(x_{t,j})^{-1}} \leq 2\alpha(K-1)^{\frac{1}{3}}.$$

Lemma 4.7. *If $t > s$ and $(t-s) \leq d_{\max}^t$, then*

$$d_{\max}^t \leq \sqrt{2} d_{\max}^s,$$

which is equivalent to $\mathcal{D}_t \leq 2\mathcal{D}_s$.

Proof. It suffices to prove that $\mathcal{D}_t \leq 2\mathcal{D}_s$, which is equivalent to proving that $(\mathcal{D}_t - \mathcal{D}_s) \leq \frac{1}{2}\mathcal{D}_t$. We have:

$$\mathcal{D}_t - \mathcal{D}_s = \sum_{r=s+1}^{t} \hat{\sigma}_r \leq (t-s) d_{\max}^t \leq \left(d_{\max}^t\right)^2 = \frac{\mathcal{D}_t}{49K^{\frac{2}{3}}\log K} \leq \frac{\mathcal{D}_t}{2},$$

where the first inequality holds because due to skipping, for all $r \leq t$ we have $\hat{\sigma}_r \leq d_{\max}^t$, and $(t-s) \leq d_{\max}^t$. □

4.7.2.2 Proof of the Drift Control Lemma

Now we are ready to provide a proof of Lemma 4.1. Similar to the analysis of Masoudian et al. (2022), the proof relies on induction on *valid* pairs (t, s), where a pair (t, s) is considered valid if $s \leq t$ and $(t-s) \leq d_{\max}^t$. The induction step for pair (t, s) involves proving that $x_{t,i} \leq 4\max(x_{s,i}, \lambda_{s,t})$ for all $i \in [K]$. To establish this, we use the induction assumption for all valid pairs (t', s') such that $s', t' < t$, as well as all valid pairs (t', s'), such that $t' = t$ and $s < s' \leq t$. The induction base encompasses all pairs (t', t') for all $t' \in [T]$, where the statement $x_{t',i} \leq 4x_{t',i}$ holds trivially.

To control $\frac{x_{t,i}}{\max(x_{s,i}, \lambda_{s,t})}$ we first introduce an auxiliary variable $\tilde{x} = \bar{F}_s^*(-\hat{L}_{t-1}^{obs})$. We then address the problem of drift control by breaking it down into two sub-problems:

1. $\frac{x_{t,i}}{\max(\tilde{x}_i, \lambda_{s,t})} \leq 2$: the drift due to change of regularizer,
2. $\frac{\tilde{x}_i}{x_{s,i}} \leq 2$: the drift due to loss shift.

Deviation induced by the change of regularizer

The regularizer at round r is defined as

$$F_r(x) = \sum_{i=1}^{K} f_r(x_i) = \sum_{i=1}^{K} \left(-2\eta_r^{-1}\sqrt{x_i} + \gamma_r^{-1} x_i(\log x_i - 1)\right).$$

We have $x_t = \nabla \bar{F}_t^*(-\hat{L}_{t-1}^{obs})$ and $\tilde{x} = \nabla \bar{F}_s^*(-\hat{L}_{t-1}^{obs})$. According to the KKT conditions, there exist Lagrange multipliers μ and $\tilde{\mu}$, such that for all i:

$$f_s'(\tilde{x}_i) = -\hat{L}_{t-1,i}^{obs} + \tilde{\mu},$$
$$f_t'(x_{t,i}) = -\hat{L}_{t-1,i}^{obs} + \mu.$$

We also know that there exists an index j, such that $\tilde{x}_j \geq x_{t,j}$. This leads to the following inequality:

$$-\hat{L}_{t-1,j}^{obs} + \mu = f_t'(x_{t,j}) \leq f_s'(x_{t,j}) \leq f_s'(\tilde{x}_j) = -\hat{L}_{t-1,j}^{obs} + \tilde{\mu},$$

where the first inequality holds because the learning rates are decreasing, and the second inequality is due to the fact that $f_s'(x)$ is increasing. This implies that $\mu \leq \tilde{\mu}$, which gives us the following inequality for all i:

$$f_t'(x_{t,i}) = -\frac{1}{\eta_t\sqrt{x_{t,i}}} + \frac{\log(x_{t,i})}{\gamma_t} \leq -\frac{1}{\eta_s\sqrt{\tilde{x}_i}} + \frac{\log(\tilde{x}_i)}{\gamma_s} = f_s'(\tilde{x}_i).$$

Thus, we have two cases, either $-\frac{1}{\eta_t\sqrt{x_{t,i}}} \leq -\frac{1}{\eta_s\sqrt{\tilde{x}_i}}$ or $\frac{\log(x_{t,i})}{\gamma_t} \leq \frac{\log(\tilde{x}_i)}{\gamma_s}$.

Case i: If $-\frac{1}{\eta_t\sqrt{x_{t,i}}} \leq -\frac{1}{\eta_s\sqrt{\tilde{x}_i}}$ holds, then we have $\frac{x_{t,i}}{\tilde{x}_i} \leq \frac{\eta_s^2}{\eta_t^2} = \frac{t}{s}$. On the other hand, we have

$$t - s \leq d_{\max}^t = \sqrt{\frac{\sum_{r=1}^{t}\hat{\sigma}_r}{K^{3/2}\log K}} \leq \sqrt{\frac{t^2/2}{K^{3/2}\log K}} \leq \frac{t}{2},$$

where the second inequality holds because trivially $\hat{\sigma}_r \leq r$. This implies that $\frac{x_{t,i}}{\tilde{x}_i} \leq 2$.

Case ii: If $\frac{\log(x_{t,i})}{\gamma_t} \leq \frac{\log(\tilde{x}_i)}{\gamma_s}$, it implies that $x_{t,i} \leq \tilde{x}_i^{\frac{\gamma_t}{\gamma_s}}$. Using $\tilde{x}_i \leq \max(\tilde{x}_i, \lambda_{s,t})$,

we get

$$\begin{aligned}
x_{t,i} &\leq \max(\tilde{x}_i, \lambda_{s,t})^{\frac{\gamma_t}{\gamma_s}} \\
&= \max(\tilde{x}_i, \lambda_{s,t}) \times \max(\tilde{x}_i, \lambda_{s,t})^{\frac{\gamma_t}{\gamma_s}-1} \\
&\leq \max(\tilde{x}_i, \lambda_{s,t}) \times \lambda_{s,t}^{\frac{\gamma_t}{\gamma_s}-1} \\
&= \max(\tilde{x}_i, \lambda_{s,t}) \times \lambda_{s,t}^{-\frac{\sqrt{\mathcal{D}_t}-\sqrt{\mathcal{D}_s}}{\sqrt{\mathcal{D}_t}}} \\
&= \max(\tilde{x}_i, \lambda_{s,t}) \times e^{\frac{\mathcal{D}_t}{\mathcal{D}_t-\mathcal{D}_s} \times \frac{\sqrt{\mathcal{D}_t}-\sqrt{\mathcal{D}_s}}{\sqrt{\mathcal{D}_t}}} \\
&= \max(\tilde{x}_i, \lambda_{s,t}) \times e^{\frac{\sqrt{\mathcal{D}_t}}{(\sqrt{\mathcal{D}_t}+\sqrt{\mathcal{D}_s})}} \leq \max(\tilde{x}_i, \lambda_{s,t}) \times e^{\frac{1}{1+\sqrt{\frac{1}{2}}}} \leq \max(\tilde{x}_i, \lambda_{s,t}) \times 2.
\end{aligned}$$

Therefore, in both cases we obtain

$$x_{t,i} \leq 2 \max(\tilde{x}_i, \lambda_{s,t}). \tag{4.24}$$

Deviation Induced by the Loss Shift

The initial steps of the proof of this part are the same as in Masoudian et al. (2022). However, for the sake of completeness, we restate them here.

Since we have $x_s = \nabla \bar{F}_s^*(-\hat{L}_{s-1}^{obs})$ and $\tilde{x} = \nabla \bar{F}_s^*(-\hat{L}_{t-1}^{obs})$, they both share the same regularizer $F_s(x) = \sum_{i=1}^K f_s(x_i)$. For brevity, we drop s from $f_s(x)$. By the KKT conditions $\exists \mu, \tilde{\mu}$ s.t. $\forall i$:

$$\begin{aligned}
f^{'}(x_{s,i}) &= -\hat{L}_{s-1,i}^{obs} + \mu, \\
f^{'}(\tilde{x}_i) &= -\hat{L}_{t-1,i}^{obs} + \tilde{\mu}.
\end{aligned}$$

Let $\tilde{\ell} = \hat{L}_{t-1}^{obs} - \hat{L}_{s-1}^{obs}$, then by the concavity of $f^{'}(x)$ from Fact 4.2, we have

$$(x_{s,i} - \tilde{x}_i) f^{''}(x_{s,i}) \leq \underbrace{f^{'}(x_{s,i}) - f^{'}(\tilde{x}_i)}_{\mu - \tilde{\mu} + \tilde{\ell}_i} \leq (x_{s,i} - \tilde{x}_i) f^{''}(\tilde{x}_i). \tag{4.25}$$

Since $f^{''}(x_{s,i}) \geq 0$, from the left side of (4.25) we get $x_{s,i} - \tilde{x}_i \leq f^{''}(x_{s,i})^{-1} \left(\mu - \tilde{\mu} + \tilde{\ell}_i\right)$. Taking summation over all i and using the fact that both vectors x_s and $\tilde{x}$ are probability vectors, we have

$$\begin{aligned}
0 = \sum_{i=1}^K (x_{s,i} - \tilde{x}_i) &\leq \sum_{i=1}^K f^{''}(x_{s,i})^{-1} \left(\mu - \tilde{\mu} + \tilde{\ell}_i\right), \\
\Rightarrow \tilde{\mu} - \mu &\leq \frac{\sum_{i=1}^K f^{''}(x_{s,i})^{-1} \tilde{\ell}_i}{\sum_{i=1}^K f^{''}(x_{s,i})^{-1}}.
\end{aligned} \tag{4.26}$$

Combining the right hand sides of (4.25) and (4.26) gives

$$(\tilde{x}_i - x_{s,i}) f''(\tilde{x}_i) \leq \tilde{\mu} - \mu - \tilde{\ell}_i \leq \frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1} \tilde{\ell}_j}{\sum_{j=1}^{K}, f''(x_{s,j})^{-1}},$$

and by rearrangement we get

$$\begin{aligned} \tilde{x}_i &\leq x_{s,i} + f''(\tilde{x}_i)^{-1} \times \frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1} \tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \\ &\leq x_{s,i} + \gamma_s \tilde{x}_i \times \frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1} \tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}}, \end{aligned} \tag{4.27}$$

where the last inequality holds because $f''(\tilde{x}_i)^{-1} = \left(\eta_s^{-1} \frac{1}{2} \tilde{x}_i^{-3/2} + \gamma_s^{-1} \tilde{x}_i^{-1}\right)^{-1}$. The next step for bounding $\tilde{x}_i$ is to bound $\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1} \tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}}$ in (4.27), where $\tilde{\ell}_j = \sum_{r \in A} \hat{\ell}_{r,j}$ and $A = \left\{ r : s \leq r + \hat{d}_r < t \right\}$.

If there exists $r \in A$, such that $r > s$ and $4 \max(x_{r,i}, \lambda_{r,r+\hat{d}_r}) \leq x_{s,i}$, then combining it with the induction assumption for $(r + \hat{d}_r, r)$, where we have $x_{r+\hat{d}_r,i} \leq 4 \max(x_{r,i}, \lambda_{r,r+\hat{d}_r})$, leads to $x_{r+\hat{d}_r,i} \leq x_{s,i}$. On the other hand, by the induction assumption for pair $(r + \hat{d}_r, t)$, we have

$$x_{t,i} \leq 4 \max(x_{r+\hat{d}_r,i}, \lambda_{r+\hat{d}_r,t}).$$

So using $x_{r+\hat{d}_r,i} \leq x_{s,i}$ and $\lambda_{r+\hat{d}_r,t} \leq \lambda_{s,t}$ we can derive $x_{t,i} \leq 4 \max(x_{s,i}, \lambda_{s,t})$. This inequality satisfies the condition we wanted to prove in the drift lemma. Therefore, we assume that for all $r \in A$ we have either $r \leq s$ or $x_{s,i} \leq 4 \max(x_{r,i}, \lambda_{r,r+\hat{d}_r})$. If $r \leq s$, using the the induction assumption for (s, r) together with the fact that $\lambda_{r,s} \leq \lambda_{r,r+\hat{d}_r}$, results in $x_{s,i} \leq 4 \max(x_{r,i}, \lambda_{r,s})$. Consequently, in either case, the following inequality holds for all $r \in A$

$$x_{s,i} \leq 4 \max(x_{r,i}, \lambda_{r,r+\hat{d}_r}). \tag{4.28}$$

Thus, inequality in (4.28) satisfies the condition of Lemma 4.6, and for all $r \in A$ we get:

$$\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1} \hat{\ell}_{r,j}}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \leq 8(K-1)^{\frac{1}{3}}. \tag{4.29}$$

We proceed by summing both sides of the inequality (4.29) over all $r \in A$ and obtain $\frac{\sum_{j=1}^{K} f''(x_{s,j})^{-1}\tilde{\ell}_j}{\sum_{j=1}^{K} f''(x_{s,j})^{-1}} \leq 4|A|(K-1)^{\frac{1}{3}}$. Now it suffices to plug this result into (4.27):

$$\begin{aligned}
&\tilde{x}_i \leq x_{s,i} + 8|A|\gamma_s \tilde{x}_i (K-1)^{\frac{1}{3}} \Rightarrow \\
&\tilde{x}_i \leq x_{s,i} \times \left(\frac{1}{1 - 8|A|\gamma_s (K-1)^{1/3}}\right) \qquad (4.30) \\
&\quad \leq x_{s,i} \times \left(\frac{1}{1 - 24\gamma_s d^s_{\max}(K-1)^{1/3}}\right) \\
&\quad \leq x_{s,i} \times \left(\frac{1}{1-1/2}\right) = 2x_{s,i}, \qquad (4.31)
\end{aligned}$$

where the third inequality uses $|A| \leq d^s_{\max} + t - s \leq d^t_{\max} + d^s_{\max}$, and that $d^t_{\max} \leq 2d^s_{\max}$ by Lemma 4.7, and for the last inequality we use the definitions of γ_s and $d^s_{\max}$.

Combining (4.31) and (4.24) completes the induction step.

4.7.3 Self-Bounding Analysis

In this section we show the details of how to apply self-bounding analysis to bound the right hand side of (4.6).

We start from (4.6) and decompose it as follows

$$\begin{aligned}
\overline{Reg}_T \leq \mathbb{E}\Bigg[&\underbrace{a \sum_{t=1}^{T} \sum_{i \neq i^*} \eta_t x_{t,i}^{1/2}}_{A} + \underbrace{b \sum_{t=1}^{T} \sum_{i \neq i^*} \gamma_{t+d_t} (\upsilon_{t+d_t} - 1) x_{t,i} \Delta_i}_{B} + \underbrace{c \sum_{t=2}^{T} \sum_{i=1}^{K} \frac{\hat{\sigma}_t \gamma_t x_{t,i} \log(1/x_{t,i})}{\log K}}_{C} \Bigg] \\
&+ \underbrace{\mathcal{O}\left(K \sum_{t=1}^{T} \left(\lambda_{t,t+\hat{d}_t} + \lambda_{t,t+\hat{d}_t+\sigma^t_{\max}} \right) + \sigma_{\max} + S^* \right)}_{D}.
\end{aligned}$$

We rewrite the pseudo-regret as $\overline{Reg}_T = 4\overline{Reg}_T - 3\overline{Reg}_T$, and then based on the decomposition above we have

$$\overline{Reg}_T \leq \mathbb{E}\left[4aA - \overline{Reg}_T\right] + \mathbb{E}\left[4bB - \overline{Reg}_T\right] + \mathbb{E}\left[4cC - \overline{Reg}_T\right] + 4D. \qquad (4.32)$$

Masoudian et al. (2022) provide the following three lemmas that give the bounds for the first three terms in (4.32).

Lemma 4.8. *(Masoudian et al., 2022, Lemma 6) For any $a \geq 0$, we have:*

$$4aA - \overline{Reg}_T \leq \sum_{i \neq i^*} \frac{4a^2}{\Delta_i} \log(T+1) + 1. \tag{4.33}$$

Lemma 4.9. *(Masoudian et al., 2022, Lemma 7) Let $\upsilon_{max} = \max_{t \in [T]} \upsilon_t$, then for any $b \geq 0$:*

$$4bB - \overline{Reg}_T \leq 64b^2 \upsilon_{max} \log K. \tag{4.34}$$

It is evident that $\upsilon_{max} \leq \sigma_{\max}$, so the bound in Lemma 4.9 is dominated by $\mathcal{O}(K\sigma_{\max})$ term in the regret bound.

Lemma 4.10. *(Masoudian et al., 2022, Lemma 8) For any $c \geq 0$:*

$$4cC - \overline{Reg}_T \leq \sum_{i \neq i^*} \frac{128c^2\sigma_{\max}}{\Delta_i \log K}. \tag{4.35}$$

By plugging (4.33),(4.34),(4.35) into (4.32) we get the desired bound.

4.7.4 A Proof of Lemma 4.3

First we provide two facts and two auxiliary lemmas.

Lemma 4.11. *For any t we have*

$$2\mathcal{D}_t \geq \sum_{s=1}^{t} \hat{d}_s.$$

Proof. We show that for any $t \in [T]$ we have $\sum_{s=1}^{t} \hat{d}_s - \mathcal{D}_t \leq \mathcal{D}_t$:

$$\begin{aligned}
\sum_{s=1}^{t} \hat{d}_s - \mathcal{D}_t &= \sum_{(s \leq t) \wedge (s + \hat{d}_s > t)} (\hat{d}_s - \hat{\sigma}_s) \\
&\leq \sum_{(s \leq t) \wedge (s + \hat{d}_s > t)} \hat{d}_s \\
&\leq \left(d^t_{\max}\right)^2 = \frac{\mathcal{D}_t}{49K^{\frac{2}{3}} \log K} \leq \mathcal{D}_t,
\end{aligned}$$

where the second inequality holds because $\hat{d}_s \leq d^t_{\max}$, and the total number of steps that satisfy $(s \leq t) \wedge (s + \hat{d}_s > t)$ is less than the skipping threshold at time t, which is again $d^t_{\max}$. Rearranging the inequality completes the proof. □

Lemma 4.12 ((Orabona, 2022, Lemma 4.13)). *Let $a_0 \geq 0$ and $f : [0; +\infty) \to [0; +\infty)$ be a nonincreasing function. Then*

$$\sum_{t=1}^{T} a_t f\left(a_0 + \sum_{i=1}^{t} a_i\right) \leq \int_{a_0}^{\sum_{t=0}^{T} a_t} f(x)dx.$$

Fact 4.4. *For any $x \geq 0$, we have $e^{-x} \leq \frac{1}{x}$.*

Fact 4.5. *For any $x \geq 1$, we have $e^{-x} \leq \frac{1}{x \log^2(x)}$.*

Proof of Lemma 4.3. We have two summations as

$$\sum_{t=1}^{T} e^{-\frac{\mathcal{D}_{t+\hat{d}_t}}{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}} + \sum_{t=1}^{T} e^{-\frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} - \mathcal{D}_t}},$$

where we show an upper bound of $\mathcal{O}(\hat{\sigma}_{\max})$ for each of them.

Bounding the First Summation: Let T_0 be the time satisfying $\sqrt{\mathcal{D}_{T_0}} = \frac{\hat{\sigma}_{\max}}{K^{1/3}\log(K)}$, then using Facts 4.4 and 4.5 we have

$$\sum_{t=1}^{T} e^{-\frac{\mathcal{D}_{t+\hat{d}_t}}{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}} \leq \underbrace{\sum_{t=1}^{T_0} \frac{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}{\mathcal{D}_{t+\hat{d}_t}}}_{A} + \underbrace{\sum_{t=T_0+1}^{T} \frac{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}{\mathcal{D}_{t+\hat{d}_t} \log^2\left(\frac{\mathcal{D}_{t+\hat{d}_t}}{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}\right)}}_{B}.$$

For A we give the following bound

$$\begin{aligned}
A = \sum_{t=1}^{T_0} \sum_{s=t+1}^{t+\hat{d}_t} \frac{\hat{\sigma}_s}{\mathcal{D}_{t+\hat{d}_t}} &= \sum_{s=1}^{T_0} \sum_{t=0}^{s-1} \frac{\hat{\sigma}_s \mathbb{1}(t + \hat{d}_t \geq s)}{\mathcal{D}_{t+\hat{d}_t}} \\
&\leq \sum_{s=1}^{T_0} \frac{\hat{\sigma}_s^2}{\mathcal{D}_s} \\
&\leq \sum_{s=1}^{T_0} \frac{\hat{\sigma}_s \sqrt{\mathcal{D}_s}}{K^{1/3}\log(K)\mathcal{D}_s} \\
&= \sum_{s=1}^{T_0} \frac{\hat{\sigma}_s}{K^{1/3}\log(K)\sqrt{\mathcal{D}_s}} \\
&\leq \mathcal{O}\left(\frac{\sqrt{D_{T_0}}}{K^{1/3}\log(K)}\right) = \mathcal{O}\left(\frac{\hat{\sigma}_{\max}}{K^{2/3}\log^2(K)}\right),
\end{aligned}$$

where the second equality is by swapping the summations, the first inequality holds because $\mathcal{D}_{t+\hat{d}_t} \geq \mathcal{D}_s$, the third inequality uses $\hat{\sigma}_s \leq d^s_{\max} \leq \frac{\sqrt{\mathcal{D}_s}}{K^{1/3}\log K}$, and the last inequality uses Lemma 4.12.

The bound for B is as follows

$$
\begin{aligned}
B = \sum_{t=T_0+1}^{T} \sum_{s=t+1}^{t+\hat{d}_t} \frac{\hat{\sigma}_s}{\mathcal{D}_{t+\hat{d}_t} \log^2\left(\frac{\mathcal{D}_{t+\hat{d}_t}}{\mathcal{D}_{t+\hat{d}_t} - \mathcal{D}_t}\right)} &\leq \sum_{t=T_0+1}^{T} \sum_{s=t+1}^{t+\hat{d}_t} \frac{\hat{\sigma}_s}{\mathcal{D}_{t+\hat{d}_t} \log^2\left(\frac{7K^{1/3}\log(K)\mathcal{D}_{t+\hat{d}_t}}{\hat{\sigma}_{\max}\sqrt{\mathcal{D}_{t+\hat{d}_t}}}\right)} \\
&= \sum_{s=T_0+1}^{T} \sum_{t=T_0+1}^{s-1} \frac{\hat{\sigma}_s \mathbb{1}(t+\hat{d}_t \geq s)}{\mathcal{D}_{t+\hat{d}_t} \log^2\left(\frac{\sqrt{7K^{1/3}\log(K)\mathcal{D}_{t+\hat{d}_t}}}{\hat{\sigma}_{\max}}\right)} \\
&= \sum_{s=T_0+1}^{T} \sum_{t=T_0+1}^{s-1} \frac{\hat{\sigma}_s \mathbb{1}(t+\hat{d}_t \geq s)}{4\mathcal{D}_{t+\hat{d}_t} \log^2\left(\frac{49K^{2/3}\log^2(K)\mathcal{D}_{t+\hat{d}_t}}{\hat{\sigma}^2_{\max}}\right)} \\
&\leq \sum_{s=T_0+1}^{T} \frac{\hat{\sigma}^2_s}{4\mathcal{D}_s \log^2\left(49K^{2/3}\log^2(K)\frac{\mathcal{D}_s}{\hat{\sigma}^2_{\max}}\right)} \\
&\leq \hat{\sigma}_{\max} \sum_{s=T_0+1}^{T} \frac{\hat{\sigma}_s}{4\mathcal{D}_s \log^2\left(\frac{49K^{2/3}\log^2(K)\mathcal{D}_s}{\hat{\sigma}^2_{\max}}\right)} \\
&\leq \hat{\sigma}_{\max} \int_{\mathcal{D}_{T_0}}^{\mathcal{D}_T} \frac{1}{4x\log^2(\frac{49K^{2/3}\log^2(K)x}{\hat{\sigma}^2_{\max}})} \\
&= \hat{\sigma}_{\max} \frac{-1}{4\log(\frac{49K^{2/3}\log^2(K)x}{\hat{\sigma}^2_{\max}})}\Bigg|_{\mathcal{D}_{T_0}}^{\mathcal{D}_T} = \mathcal{O}(\hat{\sigma}_{\max}),
\end{aligned}
$$

where the first inequality follows by $\hat{\sigma}_s \leq \hat{\sigma}_{\max}$ and our skipping procedure that ensures $\hat{d}_t \leq d^t_{\max} \leq \frac{\sqrt{\mathcal{D}_{t+\hat{d}_t}}}{K^{1/3}\log K}$, the second equality is by swapping the summations, the second inequality follows by $\mathcal{D}_{t+\hat{d}_t} \geq \mathcal{D}_s$ and $\sum_{t=1}^{s-1} \mathbb{1}(t+\hat{d}_t \geq s) = \hat{\sigma}_s$, the last inequality follows by Lemma 4.12 , and the last equality uses $\int \frac{1}{x\log^2(x/\hat{\sigma}^2_{\max})}dx = \frac{-1}{\log(x/\hat{\sigma}^2_{\max})}$.

Bound the Second Summation: The bound for the second summation follows the same approach, but it requires additional care due to existence of $\sigma^t_{\max}$ in it. Let

T_0 to be the time satisfying $\sqrt{\mathcal{D}_{T_0}} = \frac{\hat{\sigma}_{\max}}{K^{1/3}\log(K)}$, then using Facts 4.4 and 4.5 we have

$$\sum_{t=1}^{T} e^{-\frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}} \leq \underbrace{\sum_{t=1}^{T_0} \frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}}_{A} + \underbrace{\sum_{t=T_0+1}^{T} \frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}\log^2\left(\frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}\right)}}_{B}.$$

For A we give the following bound

$$\begin{aligned}
A = \sum_{t=1}^{T_0} e^{-\frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}} &\leq \sum_{t=1}^{T_0} \frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}-\mathcal{D}_t}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}} \\
&= \sum_{t=1}^{T_0} \sum_{s=t+1}^{t+\sigma^t_{\max}+\hat{d}_t} \frac{\hat{\sigma}_s}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}} \\
&\leq \sum_{s=1}^{T_0} \sum_{t=0}^{s-1} \frac{\hat{\sigma}_s \mathbb{1}(t+\sigma^t_{\max}+\hat{d}_t \geq s)}{\mathcal{D}_s} \\
&\leq \sum_{s=1}^{T_0} \frac{(2\sigma^s_{\max}+\hat{\sigma}_{s-\sigma^s_{\max}})\hat{\sigma}_s}{\mathcal{D}_s} \\
&\leq \sum_{s=1}^{T_0} \frac{3\sqrt{\mathcal{D}_s}\hat{\sigma}_s}{K^{1/3}\log(K)\mathcal{D}_s} \\
&= \sum_{s=1}^{T_0} \frac{3\hat{\sigma}_s}{K^{1/3}\log(K)\sqrt{\mathcal{D}_s}} \\
&\leq \mathcal{O}\left(\frac{\sqrt{\mathcal{D}_{T_0}}}{K^{1/3}\log(K)}\right) = \mathcal{O}(\frac{\hat{\sigma}_{\max}}{K^{2/3}\log^2(K)}),
\end{aligned}$$

where the first inequality is by Fact 4.4, the second inequality holds by swapping the summations and that $\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} \geq \mathcal{D}_s$, third inequality use the following derivation

$$\begin{aligned}
\mathbb{1}(t+\sigma^t_{\max}+\hat{d}_t \geq s) \leq& \mathbb{1}(t+\hat{d}_t \geq s) + \mathbb{1}(s > t+\hat{d}_t \geq s-\sigma^t_{\max}) \\
\leq& \mathbb{1}(t+\hat{d}_t \geq s) + \mathbb{1}(t \in [s-\sigma^t_{\max}, s-1])] \\
&+ \mathbb{1}(t < s-\sigma^t_{\max} \wedge t+\hat{d}_t \geq s-\sigma^t_{\max}), \qquad (4.36)
\end{aligned}$$

the third equality is by swapping the summations, the third inequality uses $\hat{\sigma}_s \leq d^s_{\max} \leq \frac{\sqrt{\mathcal{D}_s}}{K^{1/3}\log K}$, and finally the last inequality uses Lemma 4.12.

The bound for B is as follows

$$
\begin{aligned}
B &= \sum_{t=T_0+1}^{T} \frac{\sum_{s=t+1}^{t+\sigma^t_{\max}+\hat{d}_t} \hat{\sigma}_s}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} \log^2\left(\frac{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\sum_{s=t+1}^{t+\sigma^t_{\max}+\hat{d}_t}\hat{\sigma}_s}\right)} \\
&\leq \sum_{t=T_0+1}^{T} \sum_{s=t+1}^{t+\sigma^t_{\max}+\hat{d}_t} \frac{\hat{\sigma}_s}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} \log^2\left(\frac{7K^{1/3}\log(K)\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{2\hat{\sigma}_{\max}\sqrt{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}}\right)} \\
&= \sum_{s=T_0+1}^{T} \sum_{t=T_0+1}^{s-1} \frac{\hat{\sigma}_s \mathbb{1}(t+\sigma^t_{\max}+\hat{d}_t \geq s)}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} \log^2\left(\frac{3K^{1/3}\log(K)\sqrt{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}}{\hat{\sigma}_{\max}}\right)} \\
&= \sum_{s=T_0+1}^{T} \sum_{t=T_0+1}^{s-1} \frac{4\hat{\sigma}_s \mathbb{1}(t+\sigma^t_{\max}+\hat{d}_t \geq s)}{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t} \log^2\left(\frac{9K^{2/3}\log^2(K)\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}{\hat{\sigma}^2_{\max}}\right)} \\
&\leq \sum_{s=T_0+1}^{T} \frac{4(2\sigma^s_{\max}+\hat{\sigma}_{s-\sigma^s_{\max}})\hat{\sigma}_s}{\mathcal{D}_s \log^2\left(\frac{\mathcal{D}_s}{4\hat{\sigma}^2_{\max}}\right)} \\
&\leq \hat{\sigma}_{\max} \sum_{s=T_0+1}^{T} \frac{12\hat{\sigma}_s}{\mathcal{D}_s \log^2\left(\frac{9K^{2/3}\log^2(K)\mathcal{D}_s}{\hat{\sigma}^2_{\max}}\right)} \\
&\leq \hat{\sigma}_{\max} \int_{\mathcal{D}_{T_0}}^{\mathcal{D}_T} \frac{12}{x\log^2(\frac{9K^{2/3}\log^2(K)x}{\hat{\sigma}^2_{\max}})} \\
&= \hat{\sigma}_{\max} \frac{-12}{\log(\frac{9K^{2/3}\log^2(K)x}{\hat{\sigma}^2_{\max}})}\Bigg|_{\mathcal{D}_{T_0}}^{\mathcal{D}_T} = \mathcal{O}(\hat{\sigma}_{\max}),
\end{aligned}
$$

where the first inequality is due to our skipping procedure that ensures $\max\left\{\sigma^t_{\max}, \hat{d}_t\right\} \leq d^t_{\max} \leq \sqrt{\mathcal{D}_{t+\sigma^t_{\max}+\hat{d}_t}}$, the second equality is by swapping the summations, the second inequality follows by $\mathcal{D}_{t+\hat{d}_t} \geq \mathcal{D}_s$ and (4.36), the last inequality follows by Lemma 4.12, and the last equality uses $\int \frac{1}{x\log^2(x/\hat{\sigma}^2_{\max})}dx = \frac{-1}{\log(x/\hat{\sigma}^2_{\max})}$. □

4.7.5 A proof of Lemma 4.4

Proof. We use the term *free round* to refer to a round r such that υ_r^{new} is zero. By applying induction on the time step t, we show that if the algorithm is currently at time t and intends to rearrange the υ_t arrivals, there exist υ_t free rounds in the interval $[t, t+\sigma_{\max}^t - \hat{\sigma}_t + \upsilon_t]$ to which the algorithm can push the arrivals. This ensures that the arrival from round s, will be rearranged to round $\pi(s) \geq s + \hat{d}_s$, such that $\pi(s) - (s + \hat{d}_s) \leq \sigma_{\max}^t$. To this end, we assume the induction assumption holds for all $r < t$, and then proceed with induction step for t.

Induction Base:
The induction base corresponds to the first arrival time, denoted as t_0. At this time step, all υ_{t_0} arrivals can be rearranged to the free rounds in the interval $[t_0, t_0+\upsilon_{t_0}-1]$, which is a subset of $[t_0, t_0+\sigma_{\max}^{t_0} - \hat{\sigma}_{t_0} + \upsilon_{t_0} - 1]$. Therefore, the induction base holds.

Induction step:
Assume that we are at round t, and our aim is to rearrange the arrivals of round t. We define t_1 as the last occupied round, where $t_1 \geq t$. To prove that $t_1 - t \leq \sigma_{\max}^t - \hat{\sigma}_t$, we first note that since the algorithm is greedy, all rounds $t, t+1, \ldots, t_1 - 1$ must also be occupied by some arrivals from the past.

Let $t_0 < t$ be the first round where one of its arrivals has been rearranged to t, and let $\upsilon_{t_0}^{'}$ be the number of arrivals at time t_0 that are rearranged to some rounds before t. Then by induction assumption we know

$$t - t_0 \leq \sigma_{\max}^{t_0} - \hat{\sigma}_{t_0} + \upsilon_{t_0}^{'} + 1 = \sigma_{\max}^{t_0} - \sum_{r=1}^{t_0-1} \mathbb{1}(r + \hat{d}_r \geq t_0) + \upsilon_{t_0}^{'} + 1. \tag{4.37}$$

On the other hand, by the choice of t_0, each occupied round $t, t+1, \ldots, t_1$ must be occupied by exactly one arrival among the arrivals of rounds $t_0, \ldots, t-1$, except for the $\upsilon_t^{'}$ arrivals of t_0 that are rearranged to some rounds before t. So we have

$$\begin{aligned}
t_1 - t + 1 &\leq \sum_{r=1}^{t-1} \mathbb{1}(t_0 \leq r + \hat{d}_r \leq t-1) - \upsilon_{t_0}^{'} \\
&= \sum_{r=1}^{t_0-1} \mathbb{1}(t_0 \leq r + \hat{d}_r \leq t-1) + \sum_{r=t_0}^{t-1} \mathbb{1}(t_0 \leq r + \hat{d}_r \leq t-1) - \upsilon_{t_0}^{'} \\
&= \sum_{r=1}^{t_0-1} \mathbb{1}(t_0 \leq r + \hat{d}_r \leq t-1) + t - t_0 - \sum_{r=t_0}^{t-1} \mathbb{1}(r + \hat{d}_r \geq t) - \upsilon_{t_0}^{'},
\end{aligned}$$

where the second equality holds because $\sum_{r=t_0}^{t-1} \mathbb{1}(r + \hat{d}_r \geq t_0) = t - t_0$. We use (4.37)

to bound $t - t_0$ in the above inequality and get

$$\begin{aligned} t_1 - t &\leq \sigma_{\max}^{t_0} + \sum_{r=1}^{t_0-1} \mathbb{1}(t_0 \leq r + \hat{d}_r \leq t-1) - \sum_{r=1}^{t_0-1} \mathbb{1}(r + \hat{d}_r \geq t_0) - \sum_{r=t_0}^{t-1} \mathbb{1}(r + \hat{d}_r \geq t) \\ &= \sigma_{\max}^{t_0} - \sum_{r=1}^{t_0-1} \mathbb{1}(r + \hat{d}_r \geq t) - \sum_{r=t_0}^{t-1} \mathbb{1}(r + \hat{d}_r \geq t) \\ &= \sigma_{\max}^{t_0} - \sum_{r=1}^{t-1} \mathbb{1}(r + \hat{d}_r \geq t) \leq \sigma_{\max}^{t} - \hat{\sigma}_t, \end{aligned} \tag{4.38}$$

where the last inequality follows by the fact that $\{\sigma_{\max}^r\}_{r\in[T]}$ is a non-decreasing sequence. So if the algorithm rearranges the υ_t arrivals at round t to rounds $t_1 + 1, \ldots, t_1 + \upsilon_t$, then, using the inequality (4.38), we can conclude that these rounds fall within the interval $[t, t + \sigma_{\max}^t - \hat{\sigma}_t + \upsilon_t]$. □

4.7.6 A Bound on S^*

By Lemma 4.14 we know that Algorithm 5 does not not skip more than one outstanding observation per round. Let $(t_1, \ldots, t_{S^*})$ be an indexing of $\mathcal{S}$. By definition of skipping thresholds, we know that $d_{\max}^{t_s} = \sqrt{\frac{\mathcal{D}_{t_s}}{K^{1/3}\log K}}$ for all $s \in [S^*]$. If we only consider the contribution of outstanding observations for the skipped rounds in $\mathcal{D}_{t_s}$, we get the following inequality

$$\begin{aligned} d_{\max}^{t_s} = \sqrt{\frac{\mathcal{D}_{t_s}}{K^{1/3}\log K}} &\geq \sqrt{\frac{\sum_{i=1}^{s} \hat{d}_{t_i}}{K^{1/3}\log K}} \\ &= \sqrt{\frac{\sum_{i=1}^{s} d_{\max}^{t_i}}{K^{1/3}\log K}}, \end{aligned} \tag{4.39}$$

where the second inequality holds because the delay of skipped rounds must be equal to the skipping thresholds.

Let $c = K^{1/3}\log K$, then by (4.39) for any $s \in [S^*]$ we get

$$\left(c d_{\max}^{t_s}\right)^2 \geq \sum_{i=1}^{s} c d_{\max}^{t_i}.$$

By rearrangement we obtain the following recursive relation:

$$\sum_{i=1}^{s-1} c d_{\max}^{t_i} \leq \left(c d_{\max}^{t_s}\right)^2 - c d_{\max}^{t_s} \leq (c d_{\max}^{t_s} - \frac{1}{2})^2. \tag{4.40}$$

We use this recursive relation to show by induction that $cd_{\max}^{t_s} \geq s/2$ for any $s \in [S^*]$.

Induction base: We have $d_{\max}^{t_1} \geq 1$, so clearly $cd_{\max}^{t_1} \geq 1/2$ is satisfied.

Induction step: To prove the induction step for s, we use (4.40) to get

$$\begin{aligned} cd_{\max}^{t_s} &\geq \frac{1}{2} + \sqrt{\sum_{i=1}^{s-1} cd_{\max}^{t_i}} \\ &\geq \frac{1}{2} + \sqrt{\sum_{i=1}^{s-1} \frac{i}{2}} \\ &= \frac{1}{2} + \sqrt{\frac{s(s-1)}{4}} \\ &\geq \frac{1}{2} + \frac{s-1}{2} = \frac{s}{2}, \end{aligned}$$

where the second inequality follows by the induction assumption for all $i \in [s-1]$, and the last inequality uses the fact that $s \geq 1$. Thus, the induction step is satisfied.

We obtain that $S^* \leq 2c \times d_{\max}^{t_{S^*}} = \mathcal{O}(cd_{\max})$, which together with Lemma 4.5 completes the proof.

4.7.7 Adversarial bounds with $d_{\max}$ cannot benefit from skipping

In this section we show that adversarial regret bounds that involve terms that are linear in $d_{\max}$, such as the bounds of Masoudian et al. (2022), cannot benefit from skipping. We prove the following lemma.

Lemma 4.13.

$$\sqrt{D} \leq \min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}} \right) + d_{\max}.$$

Proof. For any split of the rounds $[T]$ into $\mathcal{S}$ and $\bar{\mathcal{S}}$ we have

$$D = D_{\bar{\mathcal{S}}} + D_{\mathcal{S}} \leq D_{\bar{\mathcal{S}}} + |\mathcal{S}| d_{\max} \leq D_{\bar{\mathcal{S}}} + |\mathcal{S}|^2 + d_{\max}^2.$$

Thus

$$\sqrt{D} \leq \sqrt{D_{\bar{\mathcal{S}}} + |\mathcal{S}|^2 + d_{\max}^2} \leq |\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}} + d_{\max},$$

and since the above holds for any $\mathcal{S}$, we obtain the statement of the lemma. □

We remind that skipping allows to replace a term of order $\sqrt{D}$ by a term of order $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}}\right)$ (for simplicity we ignore factors dependent on K). Thus, it may potentially replace a bound of order $\sqrt{D} + d_{\max}$ by a bound of order $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}}\right) + d_{\max}$, but since by the lemma $\min_{\mathcal{S}} \left(|\mathcal{S}| + \sqrt{D_{\bar{\mathcal{S}}}}\right) + d_{\max} = \Omega(\sqrt{D})$, this would not improve the order of the bound.

4.7.8 Details of the Adversarial Analysis

The only difference between our algorithm and the algorithm of Zimmert and Seldin (2020) is the implicit exploration and the slightly modified skipping rule. Let ℓ_t be the original loss sequence, then the adversary can create an adaptive sequence $\tilde{\ell}_t$ that forces the player to play according to the implicit exploration rule by simply down-scaling all the losses by

$$\tilde{\ell}_{ti} = \frac{x_{ti}\ell_{ti}}{\max\left\{x_{t,i}, \lambda_{t,t+\hat{d}_t}\right\}}.$$

Our regret bound decomposes now into

$$\begin{aligned}\overline{Reg}_T &= \max_{i_T^*} \mathbb{E}\left[\sum_{t=1}^{T} \langle x_t, \ell_t\rangle - \ell_{t,i_T^*}\right] \\ &\leq \max_{i_T^*} \mathbb{E}\left[\sum_{t=1}^{T} \left\langle x_t, \tilde{\ell}_t\right\rangle - \tilde{\ell}_{t,i_T^*}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left\langle x_t, \ell_t - \tilde{\ell}_t\right\rangle\right].\end{aligned}$$

The first term is bounded by Zimmert and Seldin (2020, Theorem 3) (since the player plays their algorithm on the modified loss sequence) by

$$4\sqrt{KT} + \sum_{t=1}^{T} \gamma_t \hat{\sigma}_t + \gamma_T^{-1} \log K + S^*$$

and the second term is

$$\sum_{t=1}^{T} \left\langle x_t, \ell_t - \tilde{\ell}_t\right\rangle \leq \sum_{i=1}^{K}\sum_{t=1}^{T}\left(1 - \frac{x_{ti}}{x_{ti} + \lambda_{t,t+\hat{d}_t}}\right)x_{ti} \leq K\sum_{t=1}^{T}\lambda_{t,t+\hat{d}_t},$$

which can be controlled via Lemma 4.3.

Next, we reason about the nature of skips. The following lemma is an adaptation of Zimmert and Seldin (2020, Lemma 5) to our skipping threshold.

Lemma 4.14. *Algorithm 5 will not skip more than 1 point at a time.*

Proof. We prove the lemma by contradiction. Assume that s_1, s_2 are both deactivated at time t. W.l.o.g. let $s_2 \leq s_1 - 1$. Skipping of s_1 at time t means $t - s_1 \geq \sqrt{\mathcal{D}_t/(K^{\frac{2}{3}}\log(K))} \geq \sqrt{\mathcal{D}_{t-1}/(K^{\frac{2}{3}}\log(K))}$. At the same time we assumed $t - 1 - s_2 \geq t - s_1$, which means that s_2 would have been deactivated at round $t-1$ or earlier. □

Next we bound the number of skips by quantities appearing in the proof.

Lemma 4.15. *The number of skips satisfies*

$$S^* \leq 2\sqrt{\mathcal{D}_T K^{\frac{2}{3}}\log(K)}\,.$$

Proof of Lemma 4.15. Recall that $\hat{d}_t$ is the contribution of a timestep t to the sum $\mathcal{D}_T$.

Let $(t_1, \ldots, t_{S^*})$ be an indexing of $\mathcal{S}$. By Lemma 4.14 we skip at most one outstanding observation per round. Thus, we have that

$$\hat{d}_{t_m} \geq \sqrt{\mathcal{D}_{t_m+\hat{d}_{t_m}}/(K^{\frac{2}{3}}\log(K))} \geq \sqrt{\sum_{i=1}^{m}\hat{d}_{t_i}/(K^{\frac{2}{3}}\log(K))} = \frac{\sqrt{\hat{d}_{t_m} + \sum_{i=1}^{m-1}\hat{d}_{t_i}}}{K^{\frac{1}{3}}\sqrt{\log(K)}}\,.$$

By solving the quadratic inequality in $\hat{d}_{t_m}$ we obtain

$$\hat{d}_{t_m} \geq \frac{1 + \sqrt{1 + 4K^{\frac{2}{3}}\log(K)\sum_{i=1}^{m-1}\hat{d}_{t_i}}}{2K^{\frac{2}{3}}\log(K)}\,.$$

Now we prove by induction that $\hat{d}_{t_m} \geq \frac{m}{2K^{\frac{2}{3}}\log(K)}$. The induction base holds since $\hat{d}_{t_1} = 1$. For the inductive step we have

$$\hat{d}_{t_m} \geq \frac{1 + \sqrt{1 + 4K^{\frac{2}{3}}\log(K)\sum_{i=1}^{m-1}\hat{d}_{t_i}}}{2K^{\frac{2}{3}}\log(K)} \geq \frac{1 + \sqrt{1 + m(m-1)}}{2K^{\frac{2}{3}}\log(K)} \geq \frac{m}{2K^{\frac{2}{3}}\log(K)}\,.$$

Finally, we have

$$\sqrt{\mathcal{D}_T\log(k)} \geq \sqrt{\sum_{m=1}^{S^*}\hat{d}_{t_m}\log(k)} \geq \sqrt{\frac{S^*(S^*+1)}{4K^{\frac{2}{3}}}} \geq \frac{1}{2K^{\frac{1}{3}}}S^*\,.$$

□

Proof of Lemma 4.5. When $\mathcal{D}_T < 16K^{\frac{2}{3}}\log(K)$, we have by Lemma 4.15

$$|\mathcal{S}| + \sqrt{\mathcal{D}_T \log(K)} \leq 12K^{\frac{2}{3}}\log(K)$$

and we are done. Otherwise, note that for any any $t \in [T] \setminus \mathcal{S}$, we have $\hat{d}_t \leq \sqrt{\mathcal{D}_T/(K^{\frac{2}{3}}\log(K))}$, hence for any $R \subset [T]$:

$$\begin{aligned}
\sum_{t\in[T]\setminus R} d_t \geq \sum_{t\in[T]\setminus R} \hat{d}_t &\geq D_T - |R|\sqrt{\mathcal{D}_T/(K^{\frac{2}{3}}\log(K))} - |\mathcal{S}| \\
&\geq D_T - |R|\sqrt{\mathcal{D}_T/(K^{\frac{2}{3}}\log(K))} - 2\sqrt{\mathcal{D}_T K^{\frac{2}{3}}\log(K)} \\
&\geq \frac{1}{2}D_T - |R|\sqrt{\mathcal{D}_T/(K^{\frac{2}{3}}\log(K))}
\end{aligned}$$

Hence,

$$\begin{aligned}
|R| + \sqrt{\sum_{s\in[T]\setminus R} d_s K^{\frac{2}{3}}\log(K)} &\geq \min_{r\in\left[0,\frac{1}{2}\sqrt{\mathcal{D}_T K^{\frac{2}{3}}\log(K)}\right]} r + \sqrt{\frac{1}{2}\mathcal{D}_T K^{\frac{2}{3}}\log(K) - r\sqrt{K^{\frac{2}{3}}\log(K)}} \\
&\geq \frac{1}{2}\sqrt{\mathcal{D}_T K^{\frac{2}{3}}\log(K)}.
\end{aligned}$$

Rearranging leads to

$$S^* + \sqrt{D_t \log(K)} \leq 6 \min_{R\subset[T]} \left(|R| + \sqrt{\sum_{s\in[T]\setminus R} d_s K^{\frac{2}{3}}\log(K)} \right).$$

□

Chapter 5

Delayed Bandits: When Do Intermediate Observations Help?

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

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Abstract

We study a K-armed bandit with delayed feedback and intermediate observations. We consider a model where intermediate observations have a form of a finite state, which is observed immediately after taking an action, whereas the loss is observed after an adversarially chosen delay. We show that the regime of the mapping of states to losses determines the complexity of the problem, irrespective of whether the mapping of actions to states is stochastic or adversarial. If the mapping of states to losses is adversarial, then the regret rate is of order $\sqrt{(K+d)T}$ (within log factors), where T is the time horizon and d is a fixed delay. This matches the regret rate of a K-armed bandit with delayed feedback and without intermediate observations, implying that intermediate observations are not helpful. However, if the mapping of states to losses is stochastic, we show that the regret grows at a rate of $\sqrt{\big(K+\min\{|\mathcal{S}|,d\}\big)T}$ (within log factors), implying that if the number $|\mathcal{S}|$ of states is smaller than the delay, then intermediate observations help. We also provide refined high-probability regret upper bounds for non-uniform delays, together with experimental validation of our algorithms.

5.1 Introduction

Delay is an ubiquitous phenomenon that many sequential decision makers have to deal with. For example, outcomes of medical treatments are often observed with delay, purchase events happen with delay after advertisement impressions, and acceptance/rejection decisions for scientific papers are observed with delay after manuscript submissions. The impact of delay on the performance of sequential decision makers, measured by regret, has been extensively studied under full information and bandit feedback, and in stochastic and adversarial environments. Yet, in many situations in real life *intermediate observations* may be available to the learner. For example, a health check-up might give a preliminary indication on the effect of a treatment, an advertisement click might be a precursor for an upcoming purchase, and preliminary reviews might provide some information regarding an upcoming acceptance or rejection decision. In this work we study when, and how, intermediate observations can be used to reduce the impact of delay in observing the final outcome of an action in a multi-armed bandit setting.

Online learning with delayed feedback and intermediate observations was studied by Mann et al. (2019) in a full-information setting, and then by Vernade et al. (2020) in a nonstationary stochastic bandit setting. In the paper of Vernade et al. (2020),

at each time step the learner chooses an action and immediately observes a signal (also called state) belonging to a finite set. The actual loss (i.e., feedback) incurred by the learner in that time step is only received with delay, which can be fixed or random. More formally, the observed state is drawn from a distribution that only depends on the chosen action, and the incurred loss is drawn from a distribution that only depends on the observed state (and not on the chosen action), forming a Markov chain. The work of Vernade et al. (2020) studies a setting, where s_t are nonstationary and ℓ_t are i.i.d. stochastic.

In this work, we consider two possible regimes for the mappings s_t from actions to states (stochastic and adversarial) and two possible regimes for the mappings ℓ_t from states to losses (also stochastic and adversarial). Altogether, we study four different regimes, defined by the combination of the first and the second mapping type.

$$\underset{A_t}{\text{Action}} \xrightarrow{\text{no delay}} \underset{S_t = s_t(A_t)}{\text{State}} \xrightarrow{\text{delay } d_t} \underset{\ell_t(S_t)}{\text{Loss}}$$

We characterize (within logarithmic factors) the minimax regret rates for all of them, by giving upper and lower bounds. Similar to Vernade et al., we assume that the states are observed instantaneously, and we assume that the losses are observed with delay d. We show that the minimax regret rate is fully determined by the regime of the states to losses mapping, regardless of the regime of the actions to states mapping. The results are informally summarized in the following table, where K denotes the number of actions, S denotes the number of states, and T denotes the time horizon. It is assumed that the losses belong to the $[0, 1]$ interval.

States to losses mapping	Regret (within log factors)
Adversarial	$\sqrt{(K+d)T}$
Stochastic	$\sqrt{\big(K+\min\{S,d\}\big)T}$

All of our upper bounds hold with high probability (with respect to the learner's internal randomization) irrespective of the regime of the action to states mapping.

We recall that (within logarithmic factors) the minimax regret rate in multi-armed bandits with delays without intermediate observations is of order $\sqrt{(K+d)T}$ (Cesa-Bianchi et al., 2019). Therefore, we conclude that if the mapping from states to actions is adversarial, then intermediate observations do not help (in the minimax sense), because the regret rates are the same irrespective of whether the intermediate observations are used or not, and irrespective of whether the mapping from actions to states is stochastic or adversarial. However, if the mapping from states to losses is stochastic, and the number S of states is smaller than the delay d, then

intermediate observations are helpful, and we provide an algorithm, `MetaAdaBIO`, which is able to exploit them. Our result improves on the $\widetilde{\mathcal{O}}\big(\sqrt{KST}\big)$ regret bound obtained by Vernade et al. (2020) for the case of stochastic and stationary action to states mapping. Our algorithm also applies to a more general setting of non-uniform delays $(d_t)_{t\in[T]}$ where we achieve a high-probability regret bound of order $\sqrt{KT + \min\{ST, \mathcal{D}_T\}}$ (ignoring logarithmic factors). This improves upon the total delay term $\mathcal{D}_T = d_1 + \cdots + d_T$ similarly to the respective term in the fixed delay setting.

Related work Adaptive clinical trials have served an inspiration for the multi-armed bandit model (Thompson, 1933), and, interestingly, they have also pushed the field to study the effect of delayed feedback (Simon, 1977; Eick, 1988). In the bandit setting Joulani et al. (2013) have studied a stochastic setting with random delays, whereas Neu et al. (2010, 2014) have studied a nonstochastic setting with constant delays. Cesa-Bianchi et al. (2019) have shown an $\Omega(\max\{\sqrt{KT}, \sqrt{dT \ln K}\})$ lower bound for nonstochastic bandits with uniformly delayed feedback, and an upper bound matching the lower bound within logarithmic factors by using an Exp3-style algorithm (Auer et al., 2002b), whereas Zimmert and Seldin (2020) have reduced the gap to the lower bound down to constants by using a Tsallis-INF approach (Zimmert and Seldin, 2021). Follow up works have studied adversarial multi-armed bandits with non-uniform delays (Thune et al., 2019; Bistritz et al., 2019, 2022; Gyorgy and Joulani, 2021; Van der Hoeven and Cesa-Bianchi, 2022) with Zimmert and Seldin (2020) providing a minimax optimal algorithm and Masoudian et al. (2022) deriving a matching lower bound and a best-of-both-worlds extension. Two key techniques for handling non-uniform delays are skipping, introduced by Thune et al. (2019), and algorithm parametrization by the number of outstanding observations (an observed quantity at action time), as opposed to the delays (an unobserved quantity at action time), introduced by Zimmert and Seldin (2020).

Paper structure In Section 5.2 we provide a formal problem definition. In Section 5.3 we introduce two algorithms, `MetaBIO` and `MetaAdaBIO`, for the model of bandits with intermediate observations. In Section 5.4 we analyze both algorithms and prove high-probability regret bounds for the setting of adversarial action-state mappings and stochastic losses. In Section 5.5 we provide the lower bounds, and in Section 5.6 experimental evaluation, concluding with a discussion in Section 5.7.

5.2 Problem definition

We consider an online learning setting with a finite set $\mathcal{A} = [K]$ of $K \geq 2$ actions and a finite set $\mathcal{S} = [S]$ of $S \geq 2$ states. In each round $t = 1, 2, \ldots$ the learner picks an action $A_t \in \mathcal{A}$ and receives a state $S_t = s_t(A_t) \in \mathcal{S}$ as an intermediate observation according to some mapping $s_t \in \mathcal{S}^{\mathcal{A}}$. The learner also incurs a loss $\ell_t(S_t) \in [0, 1]$, which is only observed at the end of round $t + d_t$, where the delay $d_t \geq 0$ is revealed to the learner only when the observation is received.

The difficulty of this learning task depends on three elements all initially unknown to the learner:

- the sequence of action-state mappings $s_1, \ldots, s_T \in \mathcal{S}^{\mathcal{A}}$;
- the sequence of loss vectors $\ell_1, \ldots, \ell_T \in [0, 1]^S$;
- the sequence of delays $d_1, \ldots, d_T \in \mathbb{N}$, where $d_t \leq T - t$ for all $t \in [T]$ without loss of generality.

Note that unlike standard bandits, here the losses are functions of the states instead of the actions. However, since actions are chosen without a-priori information on the action-state mappings, learners have no direct control on the losses they will incur and, because of the delays, they also have no immediate feedback on the loss associated with the observed states. Note also that, for all $t \geq 1$, the states $s_t(a)$ for $a \neq A_t$ and the losses $\ell_t(s)$ for $s \neq S_t$ are never revealed to the algorithm. For brevity, we refer to this setting as (delayed) Bandits with Intermediate Observations (BIO).

In the setting of stochastic losses, we assume the loss vectors $\ell_t \in [0, 1]^S$ are sampled i.i.d. from some fixed but unknown distribution Q, and let $\theta \in [0, 1]^{\mathcal{S}}$ be the unknown vector of expected losses for the states. That is, $\ell_t(s) \sim Q(\cdot \mid s)$ has mean $\theta(s)$ for each $t \in [T]$ and $s \in \mathcal{S}$. Note that we allow dependencies between the stochastic losses of distinct states in the same round, but require losses to be independent across rounds. In the setting of stochastic action-state mappings, we assume that each observed state S_t is independently drawn from a fixed but unknown distribution $P(\cdot \mid A_t)$. If both losses and action-state mappings are stochastic, then $\ell_t(S_t)$ is independent of A_t given S_t. When losses or action-state mappings are adversarial, we always assume oblivious adversaries.

Our main quantity of interest is the regret measured via the learner's cumulative loss $\sum_t \ell_t(S_t)$, where $S_t = s_t(A_t)$ and $(A_t)_{t \geq 1}$ is the sequence of learner's actions. In case of stochastic losses, we define the learner's performance by $\sum_t \theta(S_t)$. In case of stochastic action-state mappings, we average each instantaneous loss over the random choice of the state: $\sum_s \ell_t(s) P(s \mid A_t)$ for adversarial losses and $\sum_s \theta(s) P(s \mid A_t)$ for

stochastic losses. Regret is always computed according to the best action with respect to appropriate notion of cumulative loss. In particular, for stochastic state-action mappings, the cumulative losses of the best action are

$$\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \ell_t(s) P(s \mid a) \qquad \text{and} \qquad \min_{a\in\mathcal{A}} \sum_{t=1}^{T} \sum_{s\in\mathcal{S}} \theta(s) P(s \mid a) \ .$$

5.3 Algorithm

In this section we introduce `MetaBIO` (Algorithm 7) that transforms any algorithm $\mathcal{B}$ tailored for the delayed setting *without* intermediate observations into an algorithm for our setting. We then propose `MetaAdaBIO`, a modification of `MetaBIO` that delivers an improved regret bound for our setting.

Algorithm 7: `MetaBIO`

Input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0,1)$

Initialize $\mathcal{L}(s) = \emptyset$ *for all* $s \in \mathcal{S}$

for $t = 1, \ldots, T$ **do**

- Get A_t from $\mathcal{B}$
- Observe $S_t = s_t(A_t)$
- **for** $j : j + d_j = t$ **do**
  - Receive $(j, \ell_j(S_j))$
  - Update $\mathcal{L}(S_j) = \mathcal{L}(S_j) \cup \{(j, \ell_j(S_j))\}$
- Initialize feedback set $\mathcal{M} = \emptyset$
- Compute $n_t(S_t)$
- **if** $|\mathcal{L}(S_t)| \geq n_t(S_t)$ **then**
  - Add t to $\mathcal{M}$
- **for** $j : j + d_j = t \wedge |\mathcal{L}(S_j)| < n_j(S_j)$ **do**
  - Add j to $\mathcal{M}$
- **for** $j \in \mathcal{M}$ **do**
  - Compute $\tilde{\theta}_t(S_j)$ from $\mathcal{L}(S_j)$ `// using` δ
  - Feed $(j, A_j, \tilde{\theta}_t(S_j))$ to $\mathcal{B}$

The idea of `MetaBIO` is to reduce the impact of delays using the information we

get from intermediate observations. More precisely, if we have *enough* observations for the current state S_t at time t, we immediately feed to $\mathcal{B}$ the *estimate* of the mean loss of this state as if it were the actual loss at time t; otherwise, we wait for d_t time steps and refine our estimate using the additional loss observations.

The are two key steps in the design of our algorithm: *how* we construct the mean estimate and *when* we use it instead of waiting for the actual loss. They are the steps highlighted in green in Algorithm 7. For all $t \in [T]$ and $s \in \mathcal{S}$, we use $\tilde{\theta}_t(s)$ to denote the mean estimate of $\theta(s)$ at round t and $n_t(s)$ to denote the number of observations for state s that we want to observe before using $\tilde{\theta}_t(s)$. We add a subscript t to $\mathcal{L}(s)$ in Algorithm 7 to denote the set of observations we have collected at the end of round t. Thus, $\tilde{\theta}_t(s)$ uses $N_t(s) = |\mathcal{L}_t(s)|$ observations.

Fixed delay setting. When all rounds have delay d, we simply choose $n_t(s) = d$ for all $s \in \mathcal{S}, t \in [T]$. In other words, if we have at least d observations for some state, then we can compensate for the effect of delays and construct a well concentrated mean estimate around the actual mean. Let $\hat{\theta}_t(s) = \sum_{j \in \mathcal{L}_t(s)} \ell_j(s) / N_t(s)$. Then our mean loss estimate is a lower confidence bound for $\theta(s)$ defined by

$$\tilde{\theta}_t(s) = \max\left\{0, \hat{\theta}_t(s) - \frac{1}{2}\varepsilon_t(s)\right\} \tag{5.1}$$

for $\varepsilon_t(s) = \sqrt{\frac{2}{N_t(s)} \ln \frac{4ST}{\delta}}$.

Arbitrary delay setting. In the arbitrary delay setting, where we do not have preliminary knowledge of delays, we can not use the delays to set $n_t(s)$. Instead, at the *end* of time t, we have access to the number of outstanding observations $\sigma_t = \left|\{j \in [t] \,:\, j + d_j > t\}\right|$, which is different from prior works that consider outstanding observation at the *beginning* of the round. Then, for any $s \in \mathcal{S}$, we may set $n_t(s) = \sigma_t$. With this choice, incurring zero delay at some round implies that we received at least half of all the observations we could have received in the no-delay setting (see Section 5.8.2.4). In Section 5.4 we see that this ensures our mean estimate is well concentrated around its mean.

Since Algorithm 7 waits for the actual loss at time t only if $N_t(S_t) < \sigma_t$, then $\tilde{d}_t = d_t \, \mathbb{1}[N_t(S_t) < \sigma_t]$ is the actual delay incurred by the algorithm, and $\mathcal{L}_{t+\tilde{d}_t}(s)$ is the set of observations used to compute the estimate of the mean loss at time t. Because some observations may arrive at the same time, the high-probability analysis of `MetaBIO` requires these observations to be ordered. More precisely, we construct

our mean estimate at time $t + \tilde{d}_t$ for the feedback of round t using the set

$$\mathcal{L}'_t(s) = \left\{ (j, \ell_j(s)) \in \mathcal{L}_{t+\tilde{d}_t}(s) \,\middle|\, j + \tilde{d}_j = t + \tilde{d}_t \Rightarrow j < t \right\}.$$

Letting $N'_t(s) = |\mathcal{L}'_t(s)|$, we define the empirical mean

$$\hat{\theta}_t(s) = \sum_{j \in \mathcal{L}'_t(s)} \frac{\ell_j(s)}{N'_t(s)} \ . \tag{5.2}$$

Algorithm 8: `MetaAdaBIO`

Input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0, 1)$

Initialize $\mathfrak{D}_0 = 0$

for $t = 1, \ldots, T$ **do**

- Get A_t from $\mathcal{B}$
- **for** $j : j + d_j = t$ **do**
  - Receive $(j, \ell_j(S_j))$
  - Feed $(j, A_j, \ell_j(S_j))$ to $\mathcal{B}$
- Set $\sigma_t = \sum_{j=1}^{t-1} \mathbb{1}[j + d_j > t]$
- Update $\mathfrak{D}_t = \mathfrak{D}_{t-1} + \sigma_t$
- **if** $\mathfrak{D}_t(3 \ln K + \ln(6/\delta)) > 49ST \ln \frac{8ST}{\delta}$ **then**
  - **break**

if $t < T$ **then**

- Run `MetaBIO`$(\mathcal{B}, \delta/2)$ for the remaining rounds

Then, we set $\varepsilon_t(s) = \sqrt{\frac{2}{N'_t(s)} \ln \frac{4ST}{\delta}}$ and define the mean loss estimator similarly to Equation (5.1).

The `MetaAdaBIO` algorithm. As we said already, the goal of intermediate observations is to reduce the impact of delays. However, if the number of states is too large compared to the average delay, then the information we get from intermediate observations could be misleading. To address this issue, we introduce `MetaAdaBIO` (Algorithm 8). Given a horizon T,[1] this algorithm runs $\mathcal{B}$ (which is tailored for the setting *without* intermediate observations) until the total incurred delay exceeds ST,

[1]Note that we may remove the a-priori knowledge of T by using a doubling trick at the cost of a polylog factor in the regret. See Remark 5.5 for further details.

and then switches to `MetaBIO`. We precise that `MetaAdaBIO` computes $\mathfrak{D}_t$ as the sum of outstanding observation counts up to round t, which is then used in the switching condition.

5.4 Regret Analysis

We analyze `MetaBIO` and `MetaAdaBIO` in the setting of adversarial action-state mappings and stochastic losses where the regret is defined by $R_T = \sum_{t=1}^T \theta(S_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^T \theta(s_t(a))$. Our analysis guarantees a bound on R_T that holds with high probability (and not just in expectation). A related notion of regret is $\mathcal{R}_T = \sum_{t=1}^T \ell_t(S_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^T \ell_t(s_t(a))$ which considers the realized losses instead of their means. The two quantities are close with high probability: each inequality

$$-\sqrt{2T\ln(2K/\delta)} \le R_T - \mathcal{R}_T \le \sqrt{2T\ln(2/\delta)} \tag{5.3}$$

individually holds with probability at least $1-\delta$ for any given $\delta \in (0,1)$ (see Lemma 5.2).

Let $\mathcal{D}_T = \sum_{t=1}^T d_t$ be the total delay. We start by showing an upper bound on the total actual delay $\tilde{\mathcal{D}}_T = \sum_{t=1}^T d_t \mathbb{1}[N_t(S_t) < \sigma_t] \le \mathcal{D}_T$ incurred by `MetaBIO`. Then, we provide a high-probability regret analysis of both `MetaBIO` and `MetaAdaBIO`.

More precisely, we can show that `MetaBIO` incurs the delays of no more than $\min\{2S\sigma_{\max}, T\}$ rounds, where $\sigma_{\max} = \max_{t\in[T]} \sigma_t$. In the worst case, these rounds correspond with those from the set

$$\Phi \in \operatorname*{argmax}_{\mathcal{J}\subseteq[T]} \left\{ \mathcal{D}_{\mathcal{J}} : |\mathcal{J}| = \min\{2S\sigma_{\max}, T\} \right\} . \tag{5.4}$$

where we denote $\mathcal{D}_{\mathcal{J}} = \sum_{t\in\mathcal{J}} d_t$ for any $\mathcal{J} \subseteq [T]$. Note that the set Φ is fully determined by the delay sequence $d_1, \dots, d_T$. Moreover, the total delay incurred by `MetaBIO` cannot be worse than the sum of delays corresponding to the rounds in Φ, as stated in the lemma below.

Lemma 5.1 (Total actual delay)**.** *If* `MetaBIO` *is run with any algorithm* $\mathcal{B}$ *on delays* $(d_t)_{t\in[T]}$*, then* $\tilde{\mathcal{D}}_T \le \mathcal{D}_\Phi$*.*

Lemma 5.1 (proof in Section 5.8.2.1) implies that, if all delays are bounded by $d_{\max}$, then $\tilde{\mathcal{D}}_T \le 2S\sigma_{\max}d_{\max}$, which does not depend on T. In the fixed-delay setting with delay d, for example, we get a total effective delay of at most $2Sd^2$, rather than the total delay dT we would incur without access to intermediate observations (when T is large enough).

We now turn `MetaBIO` into a concrete algorithm by instantiating $\mathcal{B}$. Specifically, we use `DAda-Exp3` (Gyorgy and Joulani, 2021), a variant of `Exp3` which does not use intermediate observations and is robust to delays. `DAda-Exp3` has the following regret bound.

Theorem 5.1 (Gyorgy and Joulani (2021, Corollary 4.2))**.** *For any $\delta \in (0,1)$, the regret with respect to realized losses of* `DAda-Exp3` *in the adversarial bandits with arbitrary delays with probability at least $1-\delta$ satisfies*

$$\mathcal{R}_T \leq 2\sqrt{3(2KT+\mathcal{D}_T)\ln K} + \left(\sqrt{\frac{2KT+\mathcal{D}_T}{3\ln K}} + \frac{\sigma_{\max}}{2} + 1\right)\ln\frac{2}{\delta} \ .$$

While Theorem 5.1 shows a high-probability bound on $\mathcal{R}_T$, Equation (5.3) shows that a high-probability bound for one notion of regret ensures a high-probability bound for the other. Although the original bound by Gyorgy and Joulani (2021) was stated with $d_{\max}$ instead of $\sigma_{\max}$, we can replace the former with the latter by observing that, in the analysis of Gyorgy and Joulani (2021, Theorem 4.1), they only use $d_{\max}$ to upper bound the number of outstanding observations. Note that $\sigma_{\max}$ is never larger than $d_{\max}$, indicating it is a well-behaved term that is not vulnerable to a few large delays. See Masoudian et al. (2022, Lemma 3) for a refined quantification of the relation between $\sigma_{\max}$ and $d_{\max}$.

If we consider a fixed confidence level $\delta \in (0,1)$, then we can make the learning rate η_t and the implicit exploration term γ_t in `DAda-Exp3` depend on the specific value of δ so as to achieve an improved regret bound (see Appendix 5.8.2.2). This allows us to show that in the BIO setting with adversarial action-state mappings and stochastic losses, the regret $\mathcal{R}_T$ of `DAda-Exp3` is upper bounded by

$$2\sqrt{2KTC_{K,6\delta}} + 2\sqrt{D_T C_{K,6\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{2}{\delta} \tag{5.5}$$

with probability at least $1-\delta$, where $C_{K,\delta} = 3\ln K + \ln\frac{12}{\delta}$.

Next, we state the regret bound for `MetaBIO`. We remark that we initialize `DAda-Exp3` with confidence parameter $\delta/2$ so as to guarantee the high-probability bound as in (5.5) with probability at least $1-\delta/2$ as required.

Theorem 5.2. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *using* `DAda-Exp3`*, then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T \leq 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST\ln\frac{4ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{4}{\delta} \ . \tag{5.6}$$

We begin the analysis of Theorem 5.2 by decomposing the regret into two parts: (i) the regret $\mathcal{R}_T$ of `DAda-Exp3` with losses $\tilde{\theta}_t(S_t)$, and (ii) the gap $R_T - \mathcal{R}_T$, corresponding to the cumulative error of the estimates fed to `DAda-Exp3`. For the first part, we follow an approach similar to Gyorgy and Joulani (2021) and apply Neu (2015, Lemma 1) to obtain a concentration bound for the loss estimates defined using importance weighting along with implicit exploration. When using the actual losses, the application of Neu (2015, Lemma 1) is straightforward. However, when the mean loss estimate $\tilde{\theta}_t(S_t)$ is used rather than the actual loss, there is a potential dependency between the chosen action A_t and $\tilde{\theta}_t(S_t)$. In Section 5.8.2.3 we carefully design a filtration to show that we may indeed use the high-probability regret bound of `DAda-Exp3` in order to upper bound the first part (regret $\mathcal{R}_T$ defined in terms of the estimates $\tilde{\theta}_t$).

The second part requires to bound the cumulative error of our estimator in (5.2) for the observed states $\{S_t\}_{t\in[T]}$. To this end, we use the Azuma-Hoeffding inequality to control the error of these estimates. Doing so causes a $\tilde{\mathcal{O}}(\sqrt{ST})$ term to appear in the regret bound. The detailed proof of this part is in Section 5.8.2.4, together with the proof of Theorem 5.2.

The presence of the $\tilde{\mathcal{O}}(\sqrt{ST})$ term in the regret bound implies that, when $S \gg \max\{\mathcal{D}_T/T, K\}$, using intermediate feedback leads to no advantage over ignoring it. So we ideally want to recover the original bound in (5.5) when this happens. `MetaAdaBIO` solves this issue and gives the following regret guarantee. The proof of this result is deferred to Section 5.8.2.5. We remark that, to achieve this bound, before the eventual switch we use algorithm `DAda-Exp3` with confidence parameter set to $\delta/3$ so as to guarantee a high-probability bound on R_{t^*} with probability at least $1 - \delta/2$ over the first t^* rounds that `DAda-Exp3` runs by itself.

Theorem 5.3. *Let $\delta \in (0, 1)$. If we run* `MetaAdaBIO` *with* `DAda-Exp3`*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1 - \delta$ satisfies*

$$R_T \leq 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta}\,. \tag{5.7}$$

If we consider any upper bound $d_{\max}$ on the delays $(d_t)_{t\in[T]}$, we can further observe that the regret R_T of `MetaAdaBIO` (with `DAda-Exp3`) satisfies

$$R_T = \tilde{\mathcal{O}}\Big(\sqrt{KT} + \min\Big\{\sqrt{S}\big(\sqrt{T} + d_{\max}\big), \sqrt{d_{\max}T}\Big\}\Big)$$

with high probability. This also follows from the fact that, as previously mentioned, we can bound the total delay of `MetaBIO` by $\mathcal{D}_\Phi \leq 2Sd_{\max}^2$.

Given the previous regret bounds, we observe that we may further improve the dependency on the delays by adopting the idea of skipping rounds with large delays when computing the learning rates. This "skipping" idea was introduced by Thune et al. (2019) and has been leveraged by Gyorgy and Joulani (2021) to show that `DAda-Exp3` can achieve a refined high-probability regret bound—see Gyorgy and Joulani (2021, Theorem 5.1). As a consequence, we can indeed provide an improved bound in our setting by following similar steps as in the proof of Theorem 5.2. The only main change is the adoption of the version of `DAda-Exp3` that uses the skipping procedure.

Corollary 5.1. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *with* `DAda-Exp3` *with skipping (Gyorgy and Joulani, 2021, Theorem 5.1), then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T = \mathcal{O}\left(\sqrt{KTC_{K,\delta}} + \sqrt{ST\ln\frac{ST}{\delta}} + \ln\frac{1}{\delta} + \sqrt{C_{K,\delta}\ln K}\min_{R\subseteq\Phi}\left\{|R| + \sqrt{\mathcal{D}_{\Phi\setminus R}\ln K}\right\}\right) .$$

This result could also be extended in a similar way to `MetaAdaBIO`, so as to achieve the best result from the presence of intermediate feedback.

So far, we have provided some high-probability guarantees for the regret of both `MetaBIO` and `MetaAdaBIO`, by which we can derive some expectation bounds as well (e.g., by setting $\delta \approx 1/T$). However, using the empirical mean estimators $\hat{\theta}_t$ as the mean loss estimators at time t and working directly with the expected regret allows us to improve the achievable bound by a polylogarithmic factor. Hence, for the expected regret we use `Tsallis-INF` (Zimmert and Seldin, 2020), a learning algorithm for the standard delayed bandit problem that uses a hybrid regularizer to deal with delays and gives a minimax-optimal expected regret bound. The proof of this expected regret upper bound is in Appendix 5.8.2.6.

Proposition 5.4. *If we execute* `MetaAdaBIO` *with* `Tsallis-INF` *(Zimmert and Seldin, 2020), and use the switching condition $\sqrt{8\mathfrak{D}_t\ln K} > 6\sqrt{ST\ln(2ST)}$ at each round $t \in [T]$, where $\mathfrak{D}_t = \sum_{j=1}^{t}\sigma_j$, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\mathbb{E}[R_T] \leq 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi\ln K} + 2\min\left\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T\ln K}\right\} .$$

Remark 5.5. In `MetaBIO`, we can replace T by t^2 in the definition of the confidence intervals for (5.2) and remove the need for prior knowledge of the time horizon T. In `MetaAdaBIO`, we could use a doubling trick to avoid the prior knowledge of T in the switching condition. On the other hand, it is not required to know the number of states S for expectation bounds on the regret of `MetaBIO`. However, removing the prior knowledge of S in the high-probability regret bounds is challenging. Indeed, to the best of our knowledge, there is no result in BIO that avoids prior knowledge on the number of states. Lifting this requirement in the high-probability analysis is thus an interesting question for future work.

5.5 Lower Bounds

The lower bounds in this section are for the expected regret $\mathbb{E}[R_T]$. Since our algorithms provide high-probability guarantees, the upper bounds also apply to the expected regret. Throughout this section we will make use of constant delay i.e. $d_t = d$ for all $t \in [T]$. We will first prove a general $\sqrt{KT}$ lower bound for all algorithms in BIO, after which we specialize to particular cases.

We start by proving a $\Omega\big(\sqrt{KT}\big)$ lower bound for any algorithm in our setting and for any combination of stochastic or adversarial action-state mappings and loss vectors. The construction is a reduction to the standard bandits lower bound construction (see Section 5.8.3 for a complete proof).

Theorem 5.6. *Irrespective to whether the action-state mappings and loss vectors are stochastic or adversarial, there exists a sequence of losses such that any (possibly randomized) algorithm in BIO suffers regret* $\mathbb{E}[R_T] = \Omega\big(\sqrt{KT}\big)$.

Adversarial action-state mapping and stochastic losses. We first prove a lower bound $\sqrt{ST}$ for any number $K \geq 2$ of actions. However, we do need a minor generalization of our setting to allow correlation between unseen losses. Specifically, we allow all pairs of losses $\ell_j(s), \ell_{j'}(s')$ of distinct states $s \neq s'$ to be correlated if $j > j'$ and $j - j' \leq d$, while we guarantee the i.i.d. nature of losses for any fixed state. Since $\mathbb{E}[\ell_t(S_t)] = \mathbb{E}[\theta(S_t)]$, this does not affect the analysis for the upper bound on the regret of our algorithms since $\mathbb{E}[R_T] \leq \mathbb{E}[\mathcal{R}_T]$ (see Lemma 5.4). However, for a high-probability upper bound, we need to relate R_T and $\mathcal{R}_T$, which now leads to an additive $\tilde{\mathcal{O}}(\sqrt{ST})$ term rather than an additive $\tilde{\mathcal{O}}(\sqrt{T})$ term as in Equation (5.3).

In the proof of the $\sqrt{ST}$ lower bound, we leverage the fact that losses are independent only across time steps for a fixed state, while they may depend on the losses of the other states. Note that our lower bound holds even when the learner knows the action-state assignments beforehand.

Theorem 5.7. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that $d_t = d$ for all $t \in [T]$. If $T \geq \min\{S, d\}$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\sqrt{\min\{S, d\}T}\big)$.*

We provide a sketch of the proof of Theorem 5.7 (see Appendix 5.8.3 for the full proof). First, suppose that $S \leq 2d$. For the construction of the lower bound we only consider two actions and equally split the states over these two actions. Then, we divide the T time steps in blocks of length $S/2 \leq d$. In each block, each state has the same loss. Since the block length is smaller then the delay, we have effectively created a two-armed bandit problem with $T' = T/(S/2)$ rounds and loss range $[0, S/2]$, for which we can prove a $\Omega\big(S\sqrt{T'}\big) = \Omega\big(\sqrt{ST}\big)$ lower bound by showing an equivalent lower bound for the full information setting. If $S > 2d$, we use the same construction with only $2d$ states, and obtain a $\Omega\big(\sqrt{dT}\big)$ lower bound.

Finally, we can show the following lower bound, whose proof can be found in Section 5.8.3.

Theorem 5.8. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq d+1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big)$.*

This term is also present in the dynamic regret bound of `NSD-UCRL2`, but it is necessarily incurred from their analysis even in the stationary case (Vernade et al., 2020, Theorem 1).

This last lower bound implies that the regret of our algorithm is near-optimal. Since the lower bound of Theorem 5.6 applies to the case where the action-state mapping is adversarial and the losses are stochastic, we find the following result as a corollary of Theorem 5.6, Theorem 5.7, and Theorem 5.8.

Corollary 5.2. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq 1 + \min\{S, d\}$, then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{\min\{S, d\}T}, (d+1)\sqrt{S}\big\}\big)$.*

Stochastic action-state mappings and adversarial losses. In this case we recover the standard lower bound for adversarial bandits with bounded delay. The full proof of this result can be found in Section 5.8.3.

Theorem 5.9. *Suppose that the action-state mapping is stochastic, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists a stochastic action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret* $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.

Adversarial action-state mappings, adversarial losses. Since we can recover the construction of the lower bound in Theorem 5.9, we have the following result.

Corollary 5.3. *Suppose that the action-state mapping is adversarial, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists an action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret* $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.

5.6 Experiments

We empirically compare our algorithm `MetaBIO` with the following baselines: `DAda-Exp3` (Gyorgy and Joulani, 2021) for adversarial delayed bandits without intermediate observations (which we used to instantiate the algorithm $\mathcal{B}$), the standard `UCB1` algorithm (Auer et al., 2002a) for stochastic bandits without delays and intermediate observations, and `NSD-UCRL2` (Vernade et al., 2020) for nonstationary stochastic action-state mappings and stochastic losses. We run all experiments with a time horizon of $T = 10^4$. All our plots show the cumulative regret of the algorithms considered as a function of time. The performance of each algorithm is averaged over 20 independent runs in every experiment, and the shaded areas consider a range centered around the mean with half-width corresponding to the empirical standard deviation of these 20 repetitions. In the first two experiments, we consider both fixed delays $d \in \{50, 100, 200\}$ and random delays $d_t \sim \text{Laplace}(50, 25)$ sampled i.i.d. from the Laplace distribution with $\mathbb{E}[d_t] = 50$.

Experiment 1: stochastic action-state mappings. Here we use a stationary version of the experiments in (Vernade et al., 2020)—see Table 5.1 in Section 5.8.4 for details. We set $K = 4$ and $S = 3$, while we repeat this experiment for the previously mentioned values of delays. Figure 5.1 shows that, across all delay regimes, `MetaBIO` largely improves on the performance of `DAda-Exp3` by exploiting intermediate observations.

Experiment 2: adversarial action-state mappings. In this construction, we simulate the adversarial mapping using a construction adapted from (Zimmert and Seldin, 2021): we alternate between two stochastic mappings while keeping the

loss means fixed. We set $K = 4$, $S = 3$, and we consider multiple instances for the different values of delays as in the previous experiment. The interval between two consecutive changes in the distribution of action-state mappings grows exponentially. See Table 5.2 in Section 5.8.4 for details. Figure 5.2 shows that `MetaBIO` and `MetaBIO` with "skipping" outperform both `UCB1` and `NSD-UCRL2`.

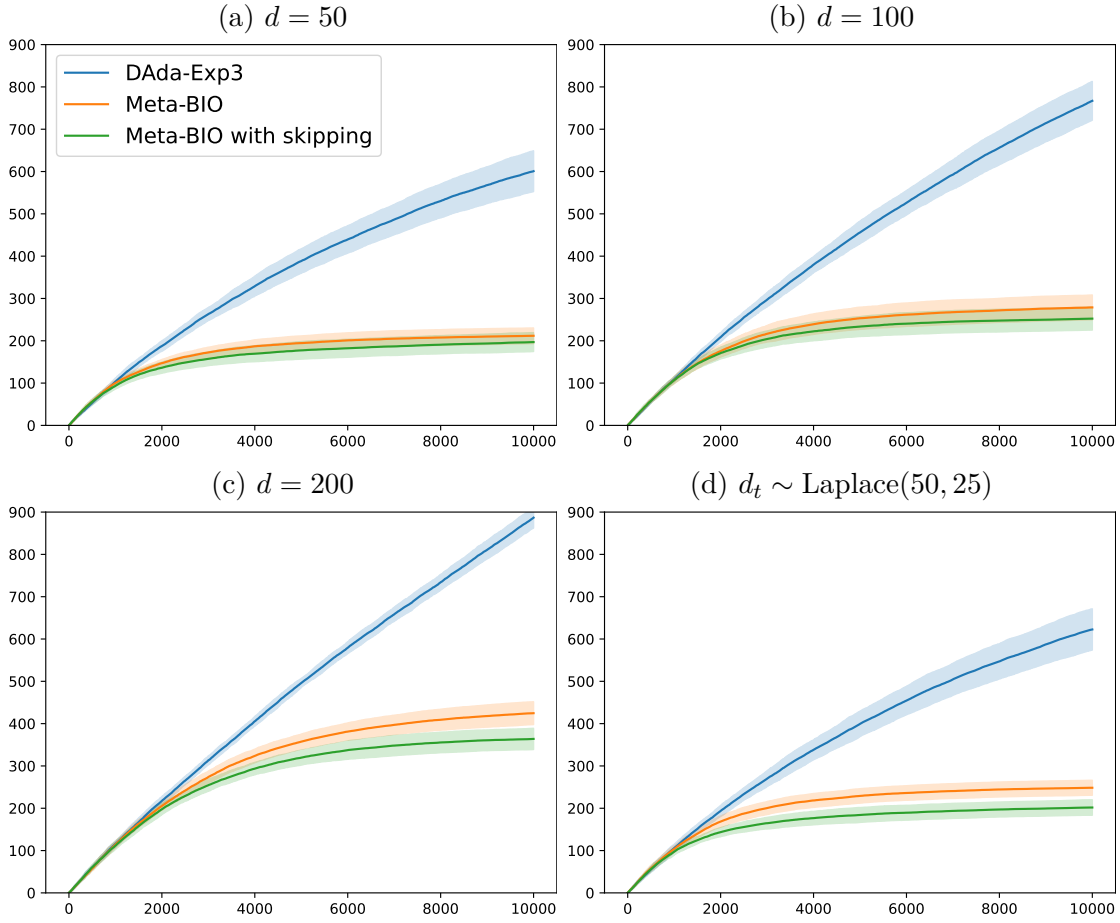

Figure 5.1: Cumulative regret over time for the stochastic action-state mapping when delays are fixed or random.

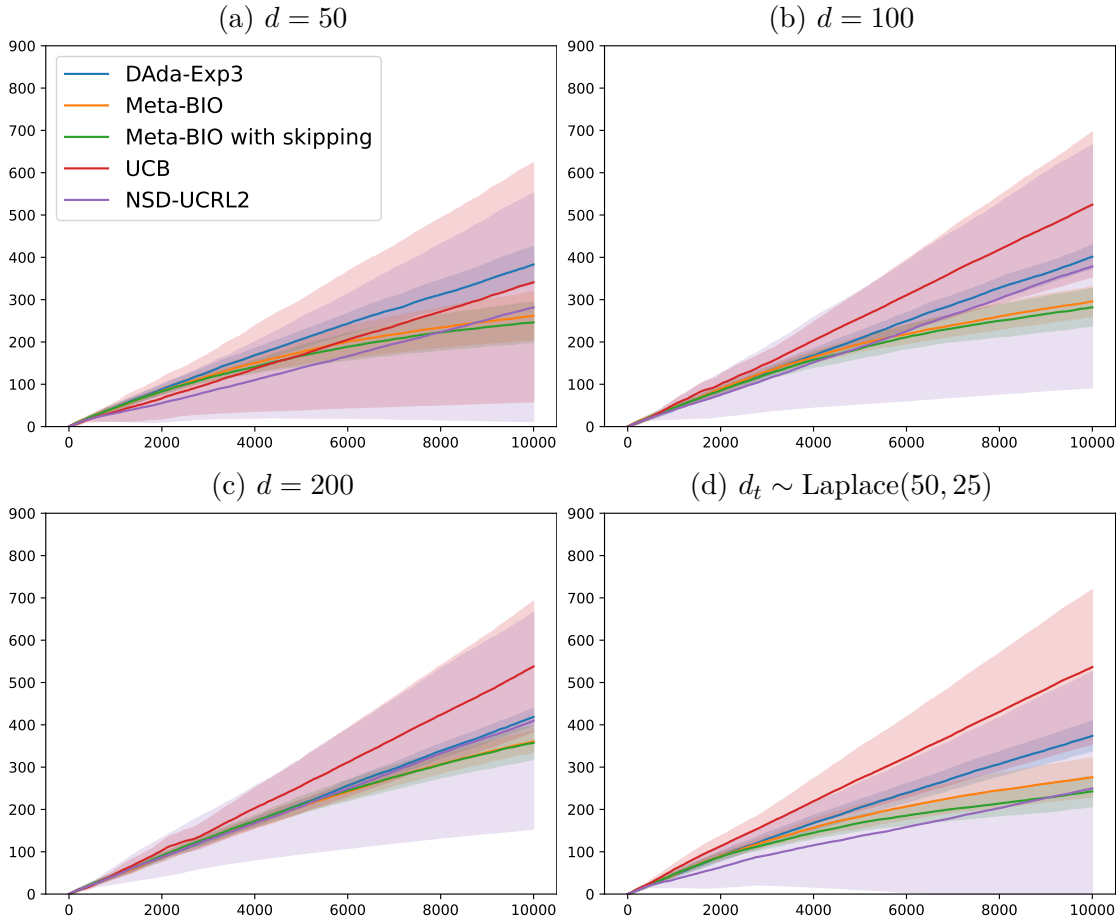

Figure 5.2: Cumulative regret over time for the adversarial action-state mapping when delays are fixed or random. All algorithms have small variance except for `UCB1` and `NSD-UCRL2`.

Experiment 3: utility of intermediate observations. Here we set $K = 8$, $d = 100$, and investigate how the performance of `MetaBIO` changes when the number S of states varies in $\{4, 6, 8, 10, 12\}$. The mean loss is always 0.2 for the optimal state and 1 for the others. The optimal action always maps to the optimal state. The suboptimal actions map to the optimal state with probability 0.6 and map to a random suboptimal state with probability 0.4. This implies that the expected loss of each arm remains constant when the number of states changes. Figure 5.3 shows that the regret gap between `MetaBIO` and `DAda-Exp3` shrinks as the number of states increases. This observation confirms our theoretical findings about the dependency

of the regret on the number of states, which lead to a larger improvement the fewer they are.

Experiment 4: performance of `MetaAdaBIO` when $S < d$. We use the same setting as in Experiment 1 with delay $d = 20$.[2] The first plot of Figure 5.4 shows the performance of `MetaAdaBIO` compared with both `DAda-Exp3` and `MetaBIO`. Before the switching point, `MetaAdaBIO` runs `DAda-Exp3` (up to independent internal randomization). Afterwards, `MetaAdaBIO` switches to `MetaBIO` (which in turn runs `DAda-Exp3` as a subroutine) and quickly aligns with its performance. Note that, at the switching time, `MetaAdaBIO` uses (via `MetaBIO`) the same instance of `DAda-Exp3` that was already running, rather than starting a new instance. It can be shown that our analysis of `MetaAdaBIO` applies to this variant as well without changes in the order of the bound.

Experiment 5: performance of `MetaAdaBIO` when $S > d$. We use a setting that is almost identical to that of Experiment 3 (Section 5.6), except we set $d = 4$ and $S = 14$. The performance of the three algorithms is shown in the second plot of Figure 5.4. We can observe that `MetaAdaBIO` does not switch to `MetaBIO` and its performance is thus the same as that of `DAda-Exp3`, whereas `MetaBIO` incurs a larger regret.

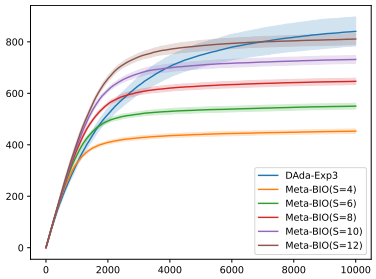

Figure 5.3: Cumulative regret over time of `DAda-Exp3` and `MetaBIO` with different numbers of states $S \in \{4, 6, 8, 10, 12\}$.

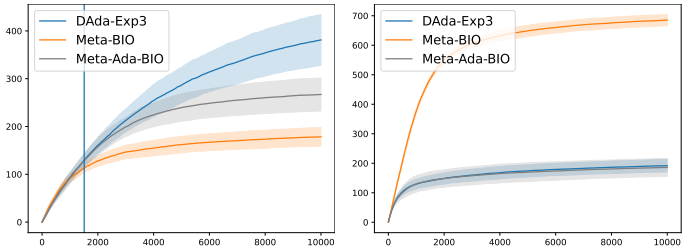

Figure 5.4: Cumulative regret over time of `DAda-Exp3`, `MetaBIO` and `MetaAdaBIO` when $S < d$ (left) and $S > d$ (right). The vertical line is the switching point of `MetaAdaBIO`.

[2]Compared to the switching condition used for the analysis of `MetaAdaBIO`, we replace $49ST \ln \frac{8ST}{\delta}$ with ST. This change allows the switching condition to be triggered more easily to provide a better visualization of the behaviour of `MetaAdaBIO`, while it only introduces a polylog factor in its regret bound.

5.7 Future Work

The work of Vernade et al. (2020) also considers a non-stationary action-state mapping and derive regret bounds for the switching regret. Preliminary results suggest that, as long as there is an algorithm that can provide bounds on the switching regret with delayed feedback, our ideas also transfer to this setting. Unfortunately, there is currently no algorithm that can provide bounds on the switching regret with delayed feedback and we leave this as a promising direction for future work.

5.8 Appendix

5.8.1 Auxiliary Results

Lemma 5.2. *Consider any algorithm that picks actions $(A_t)_{t\in[T]}$ in the adversarial delayed bandits problem with intermediate feedback with arbitrary action-state mappings $(s_t)_{t\in[T]}$ and i.i.d. loss vectors $(\ell_t)_{t\in[T]}$. Then, for any given $\delta \in (0,1)$,*

$$R_T - \mathcal{R}_T \leq \sqrt{2T\ln(2/\delta)} \qquad \text{and} \qquad \mathcal{R}_T - R_T \leq \sqrt{2T\ln(2K/\delta)}$$

individually hold with probability at least $1-\delta$.

Proof. First, observe that we can relate the two notions of regret as

$$R_T = \mathcal{R}_T + \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) + \underbrace{\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) - \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a))}_{(\triangle)} \ .$$

By Azuma-Hoeffding inequality, we can show that each side of

$$-\sqrt{\frac{T}{2}\ln\left(\frac{1}{\delta'}\right)} \leq \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) \leq \sqrt{\frac{T}{2}\ln\left(\frac{1}{\delta'}\right)} \tag{5.8}$$

holds with probability at least $1-\delta'$. Now, define

$$a^*_\ell \in \operatorname*{argmin}_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) \qquad \text{and} \qquad a^*_\theta \in \operatorname*{argmin}_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a)) \ .$$

On the one hand, observe that

$$(\triangle) \leq \sum_{t=1}^{T} \ell_t(s_t(a_\theta^*)) - \sum_{t=1}^{T} \theta(s_t(a_\theta^*)) \leq \sqrt{\frac{T}{2} \ln\left(\frac{1}{\delta'}\right)} \ ,$$

where the last inequality holds with probability at least $1 - \delta'$ by Azuma-Hoeffding inequality. On the other hand, we can show that

$$(\triangle) \geq \sum_{t=1}^{T} \ell_t(s_t(a_\ell^*)) - \sum_{t=1}^{T} \theta(s_t(a_\ell^*)) =: (\diamond) \ .$$

However, in this case a_ℓ^* depends on the entire sequence $\ell_1, \ldots, \ell_T$. We thus need to use a union bound in order to show that

$$\mathbb{P}\left((\diamond) \leq -\sqrt{\frac{T}{2} \ln\left(\frac{K}{\delta'}\right)}\right) \leq \sum_{a \in \mathcal{A}} \mathbb{P}\left(\sum_{t=1}^{T} \ell_t(s_t(a)) - \sum_{t=1}^{T} \theta(s_t(a)) \leq -\sqrt{\frac{T}{2} \ln\left(\frac{K}{\delta'}\right)}\right) \leq \delta' \ ,$$

where the last inequality follows by Azuma-Hoeffding inequality. We conclude the proof by setting $\delta' = \delta/2$. □

Lemma 5.3. *The estimates $(\hat{\theta}_t)_{t=1}^T$ defined in Equation (5.2) are such that $|\hat{\theta}_t(s) - \theta(s)| \leq \frac{1}{2}\varepsilon_t(s)$ simultaneously holds for all $t \in [T]$ and all $s \in \mathcal{S}$ with probability at least $1 - \delta/2$.*

Proof. In a similar way as in Vernade et al. (2020), define $X_m(s)$ to be the empirical mean estimate for $\theta(s)$ which uses the first $m \in [T]$ observed losses corresponding to state $s \in \mathcal{S}$. Notice that $\hat{\theta}_t(s) = X_{N_t'(s)}(s)$, while we define $\varepsilon'_m(s) = \sqrt{\frac{2}{m} \ln(\frac{4ST}{\delta})}$ so that $\varepsilon_t(s) = \varepsilon'_{N_t'(s)}(s)$. We can additionally observe that $\mathbb{E}[X_m(s)] = \theta(s)$. Then, we can use Azuma-Hoeffding inequality to show that

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{s \in \mathcal{S}} \bigcap_{t \in [T]} \left\{|\hat{\theta}_t(s) - \theta(s)| \leq \frac{1}{2}\varepsilon_t(s)\right\}\right) &\geq \mathbb{P}\left(\bigcap_{s \in \mathcal{S}} \bigcap_{m \in [T]} \left\{|X_m(s) - \theta(s)| \leq \frac{1}{2}\varepsilon'_m(s)\right\}\right) \\
&\geq 1 - 2 \sum_{s \in \mathcal{S}} \sum_{m=1}^{T} e^{-\frac{1}{2}\varepsilon'_m(s)^2 m} \\
&= 1 - \frac{\delta}{2} \ ,
\end{aligned}$$

where we also used a union bound in the second inequality. □

Lemma 5.4. *Consider any algorithm that picks actions $(A_t)_{t\in[T]}$ in the BIO setting with adversarial action-state mappings $(s_t)_{t\in[T]}$ and stochastic loss vectors $(\ell_t)_{t\in[T]}$. Assume that the losses for any fixed state are i.i.d., whereas pairs of losses $\ell_j(s), \ell_{j'}(s')$ of distinct states $s \neq s'$ might be correlated when $j > j'$ and $j - j' \leq d_{j'}$. Then, it holds that $\mathbb{E}[R_T] \leq \mathbb{E}[\mathcal{R}_T]$, where the expectation is with respect to the stochasticity of the losses and the randomness of the algorithm.*

Proof. We know that $\mathbb{E}[\ell_t(s_t(a))] = \theta(s_t(a))$ for any fixed $a \in \mathcal{A}$ and all $t \in [T]$. We further observe that

$$\mathbb{E}[\ell_t(S_t)] = \mathbb{E}\Big[\mathbb{E}[\ell_t(s_t(A_t)) \mid A_t]\Big] = \mathbb{E}[\theta(S_t)]$$

holds for all $t \in [T]$, as A_t is independent of losses that can be correlated with ℓ_t. Now, define

$$a^*_\ell \in \operatorname*{argmin}_{a\in\mathcal{A}} \sum_{t=1}^{T} \ell_t(s_t(a)) \qquad \text{and} \qquad a^*_\theta \in \operatorname*{argmin}_{a\in\mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a)) \ .$$

Then, we conclude the proof by showing that

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T] &= \sum_{t=1}^{T} \mathbb{E}[\ell_t(S_t)] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a^*_\ell))\right] \\ &\geq \sum_{t=1}^{T} \mathbb{E}[\ell_t(S_t)] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a^*_\theta))\right] = \sum_{t=1}^{T} \mathbb{E}[\theta(S_t)] - \sum_{t=1}^{T} \theta(s_t(a^*_\theta)) = \mathbb{E}[R_T] \ .\end{aligned}$$

□

5.8.2 High-Probability Regret Bound

5.8.2.1 Total delay bound

Lemma 5.1 (Total actual delay)**.** *If* `MetaBIO` *is run with any algorithm $\mathcal{B}$ on delays $(d_t)_{t\in[T]}$, then $\tilde{\mathcal{D}}_T \leq \mathcal{D}_\Phi$.*

Proof of Lemma 5.1. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s = \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s. Denote by t_s the last time step $t \in \mathcal{T}_s$ such that $N_t(s) < \sigma_t$ and let $\mathcal{C}_s = \{t \in \mathcal{T}_s : t \leq t_s\}$ be those rounds in $\mathcal{T}_s$ that come no later than t_s. According to the choice of t_s, all the rounds in $\mathcal{T}_s$ for which learner waits for the respective delayed loss, must belong to

$\mathcal{C}_s$, while the learner incurs $\tilde{d}_t = 0$ delay for rounds $t \in \mathcal{T}_s \setminus \mathcal{C}_s$. Now we partition $\mathcal{C}_s$ into two sets: the observed set $\mathcal{C}_s^{\text{obs}} = \{t \in \mathcal{C}_s : t + d_t \leq t_s\}$ and the outstanding set $\mathcal{C}_s^{\text{out}} = \{t \in \mathcal{C}_s : t + d_t > t_s\}$. From the choice of t_s, we can see that the number of rounds in $\mathcal{C}_s^{\text{obs}}$ is

$$|\mathcal{C}_s^{\text{obs}}| \leq N_{t_s}(s) < \sigma_{t_s} \leq \sigma_{\max} \ ,$$

and the number of rounds in C_s^{out} is

$$|\mathcal{C}_s^{\text{out}}| \leq \sigma_{t_s} \leq \sigma_{\max} \ .$$

Therefore, we have $|\mathcal{C}_s| \leq 2\sigma_{\max}$. So if we define $\mathcal{C}_{\text{all}} = \bigcup_{s \in \mathcal{S}} \mathcal{C}_s$, then $|\mathcal{C}_{\text{all}}| \leq \min\{2S\sigma_{\max}, T\} = |\Phi|$. This also implies that

$$\sum_{t=1}^{T} \tilde{d}_t \leq \sum_{t \in \mathcal{C}_{\text{all}}} d_t \leq \sum_{t \in \Phi} d_t$$

by definition of Φ. □

5.8.2.2 Improved Regret for `DAda-Exp3` for Fixed δ

We follow the analysis of Theorem 4.1 in Gyorgy and Joulani (2021, Appendix A) and our goal is to use the knowledge of $\delta \in (0,1)$ to tune the learning rates $(\eta_t)_{t \in [T]}$ and the implicit exploration terms $(\gamma_t)_{t \in [T]}$, accordingly. Let $d_1, \ldots, d_T$ be the sequence of delays perceived by `DAda-Exp3`, and let $D_T = \sum_{t=1}^{T} d_t$ be its total delay. Furthermore, let σ_t be the number of outstanding observations of `DAda-Exp3` at the beginning of round $t \in [T]$. Suppose that we take $\gamma_t = c\eta_t$ with $c > 0$ for all $t \in [T]$, then following the same analysis as in Gyorgy and Joulani (2021, Appendix A), we end up with the following regret bound that holds with probability at least $1 - 2\delta'$ for any $\delta' \in (0, 1/2)$:

$$\begin{aligned}\mathcal{R}_T &\leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T} \eta_t(\sigma_t + (c+1)K) + \frac{\ln(K/\delta')}{2c\eta_T} + \frac{\sigma_{\max} + c + 1}{2c} \ln(1/\delta') \\ &= \frac{1}{\eta_T}\left(\ln(K) + \frac{\ln(K/\delta')}{2c}\right) + \sum_{t=1}^{T} \eta_t(\sigma_{t-1} + (c+1)K) + \frac{\sigma_{\max} + 1}{2c} \ln(1/\delta') + \frac{\ln(1/\delta')}{2} \ .\end{aligned}$$

Therefore, by taking $\eta_t^{-1} = \sqrt{\frac{(c+1)Kt + \sum_{j=1}^{t} \sigma_j}{2\ln(K) + \frac{1}{c}\ln(K/\delta')}}$, we get the following bound with probability at least $1 - 2\delta'$:

$$\mathcal{R}_T \leq 2\sqrt{\left((c+1)KT + \sum_{t=1}^{T} \sigma_t\right)\left(2\ln(K) + \frac{\ln(K/\delta')}{c}\right)} + \frac{\sigma_{\max} + 1}{2c} \ln(1/\delta') + \frac{\ln(1/\delta')}{2} \ .$$

We know that $\sum_{t=1}^{T} \sigma_t = D_T$ by definition of σ_t. Then, we can set $c = 1$ to obtain that the regret $\mathcal{R}_T$ (as per the original notion of regret used in Gyorgy and Joulani (2021)) is

$$\mathcal{R}_T \leq 2\sqrt{2KT(3\ln(K) + \ln(1/\delta'))} + 2\sqrt{D_T(3\ln(K) + \ln(1/\delta'))} + \frac{\sigma_{\max} + 2}{2}\ln(1/\delta') \tag{5.9}$$

with probability at least $1 - 2\delta'$.

From Lemma 5.2, we have that

$$R_T \leq \mathcal{R}_T + \sqrt{2T\ln(2/\delta')} \tag{5.10}$$

holds with probability at least $1 - \delta'$. So, combining Equations (5.9) and (5.10), and setting $\delta = 3\delta'$, we can upper bound our notion of regret R_T as

$$R_T \leq 2\sqrt{2KT\left(3\ln K + \ln\frac{3}{\delta}\right)} + \sqrt{2T\ln\frac{6}{\delta}} + 2\sqrt{D_T\left(3\ln K + \ln\frac{3}{\delta}\right)} + \frac{\sigma_{\max} + 2}{2}\ln\frac{3}{\delta} \tag{5.11}$$

with probability at least $1 - \delta$.

5.8.2.3 Reduction to `DAda-Exp3` via `MetaBIO`

Based on the reduction via `MetaBIO`, we require that $\mathcal{B}$ guarantee a regret bound

$$\hat{\mathcal{R}}_T^{\mathcal{B}} = \sum_{t=1}^{T} \tilde{\theta}_t(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \tilde{\theta}_t(s_t(a)) \tag{5.12}$$

that holds with high probability when the losses experienced by $\mathcal{B}$ are of the form $\tilde{\theta}_t\big(s_t(a)\big)$. Note that, even though the action-state mappings $s_1, \ldots, s_T$ are unknown to the learner, we can provide those losses as long as $\mathcal{B}$ requires bandit feedback only. Indeed, we can compute $\tilde{\theta}_t(S_t)$ defined in Equations (5.1) and (5.2), while we cannot determine $s_t(a)$ for all actions $a \in \mathcal{A}$ that are not A_t. As mentioned in Section 5.4, in this work we consider `DAda-Exp3` (Gyorgy and Joulani, 2021) as algorithm $\mathcal{B}$ used by `MetaBIO`. In what follows, we refer to this specific choice for the algorithm $\mathcal{B}$.

The analysis of `DAda-Exp3` for the high-probability bound (Theorem 5.1) is such that most steps only require that the loss of each action is bounded in $[0, 1]$. Then, those steps apply for any such sequence of loss vectors. However, the crucial part of that analysis that requires attention is the application of Lemma 1 from Neu (2015). We restate it below for reference.

Before that, we introduce the notation required for stating the result. We consider a learner choosing actions $A_1, \dots, A_T$ according to probability distributions $p_1, \dots, p_T$ over actions. We denote by $\mathcal{F}_{t-1}$ the observation history of the learner until the beginning of round t. The result uses importance-weighted estimates for the losses $\ell_1, \dots, \ell_T$ with implicit exploration, where the implicit exploration parameter is $\gamma_t \geq 0$ for each time t. These loss estimates are defined as

$$\tilde{\ell}_t(a) = \frac{\mathbb{1}[A_t = a]}{p_t(a) + \gamma_t} \ell_t(a) \qquad \forall t \in [T], \forall a \in \mathcal{A} \ . \tag{5.13}$$

Lemma 5.5 (Neu (2015, Lemma 1))**.** *Let γ_t and $\alpha_t(a)$ be nonnegative $\mathcal{F}_{t-1}$-measurable random variables such that $\alpha_t(a) \leq 2\gamma_t$, for all $t \in [T]$ and all $a \in \mathcal{A}$. Let $\tilde{\ell}_t(a)$ be as in (5.13). Then,*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} \alpha_t(a) \big(\tilde{\ell}_t(a) - \ell_t(a) \big) \leq \ln(1/\delta)$$

holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

In our case, we require an analogous result that work when loss vectors correspond with our estimates $\tilde{\theta}_1, \dots, \tilde{\theta}_T$. However, these estimate have a dependency with the past actions chosen by the learner. This requires some nontrivial changes in the proof of Neu (2015, Lemma 1).

Before that, we introduce some crucial definitions for this proof. Let $\rho(t) = t + d_t$ be the arrival time for the realized loss $\ell_t(S_t)$ of the state S_t observed at time $t \in [T]$. Let $\tilde{\rho}(t) = t + \tilde{d}_t$ be instead the arrival time perceived by algorithm $\mathcal{B}$ relative to its choice of A_t at time t, i.e., when $\mathcal{B}$ receives $\tilde{\theta}_t(S_t)$. This also means that $\tilde{\theta}_t(S_t)$ is only defined at time $\tilde{\rho}(t) \leq \rho(t)$.

Let $\pi \colon [T] \to [T]$ be the permutation of $[T]$ that orders rounds according to their value of $\tilde{\rho}$. In other words, π satisfies the following property:

$$\pi(r) < \pi(t) \quad \iff \quad \tilde{\rho}(r) < \tilde{\rho}(t) \vee (\tilde{\rho}(r) = \tilde{\rho}(t) \wedge r < t) \qquad \forall r, t \in [T] \ . \tag{5.14}$$

This permutation allows us to sort rounds according to the order in which `MetaBIO` feeds $\mathcal{B}$ with a respective estimate for the mean loss. In particular, the r-th round in this order corresponds with the round $t_r = \pi^{-1}(r)$, for any $r \in [T]$. Hence, we can equivalently define the round t_r as the round such that its estimate $\tilde{\theta}_{t_r}(S_{t_r})$ for the mean loss $\theta(S_{t_r})$ is the r-th estimate received by $\mathcal{B}$.

Define

$$\mathcal{F}_r = \{(j, A_j, S_j, \ell_j(S_j)) \mid j \in [T], \pi(j) \leq r\} \qquad \forall r \in [T] \tag{5.15}$$

as the information observed by $\mathcal{B}$ by the end to the time step when we feed it the estimate relative to round t_r. Note that this defines a filtration, as $\mathcal{F}_{r-1} \subseteq \mathcal{F}_r$ for all $r \in [T]$, which has some desirable properties thanks to the ordering π we consider. In particular, we have that $\tilde{d}_{t_r}, \varepsilon_{t_r}, p_{t_r}, N'_{t_r}$ are $\mathcal{F}_{r-1}$-measurable random variables by the way we define them. This property is also due to the fact that N_{t_r} and $\mathcal{L}'_{t_r}$ are determined when conditioning on $\mathcal{F}_{r-1}$. Moreover, we are now interested in the following importance-weighted loss estimates with implicit exploration:

$$\tilde{\ell}_t(a) = \frac{\mathbb{1}[A_t = a]}{p_t(a) + \gamma_t} \tilde{\theta}_t(s_t(a)) \qquad \forall t \in [T], \forall a \in \mathcal{A} \ . \tag{5.16}$$

Corollary 5.4. *Let γ_{t_r} and $\alpha_{t_r}(a)$ be non-negative $\mathcal{F}_{r-1}$-measurable random variables such that $\alpha_{t_r}(a) \leq 2\gamma_{t_r}$, for all $r \in [T]$ and all $a \in \mathcal{A}$. Let $\tilde{\ell}_t(a)$ be as in (5.16). Then,*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} \alpha_t(a) \big(\tilde{\ell}_t(a) - \tilde{\theta}_t(s_t(a)) \big) \leq \ln(1/\delta)$$

holds with probability at least $1 - \delta$ for any $\delta \in (0, 1)$.

Proof. We follow the proof of Neu (2015, Lemma 1) by considering any realization $\ell_1, \dots, \ell_T$ of the losses. The main difference is that, when defining the supermartingale as in the original proof, we need to consider the terms of the sum in the order denoted by π instead of the increasing order of t. For this reason, we rewrite the sum from the statement by following the order given by π:

$$\sum_{r=1}^{T} \sum_{a=1}^{K} \alpha_{t_r}(a) \big(\tilde{\ell}_{t_r}(a) - \tilde{\theta}_{t_r}(s_{t_r}(a)) \big) \ .$$

At this point, we need prove that $\mathbb{E}\big[\tilde{\ell}_{t_r}(a) \,\big|\, \mathcal{F}_{r-1} \big] \leq \tilde{\theta}_{t_r}(s_{t_r}(a))$, where we recall that $t_r = \pi^{-1}(r)$. Also recall that ε_{t_r}, p_{t_r} and γ_{t_r} are $\mathcal{F}_{r-1}$-measurable. This property allows us to prove the inequality with the conditional expectation of $\hat{\theta}_t$ instead of the one with the actual optimistic estimates $\tilde{\theta}_t$, by the definition of the latter. In other words, we now need to prove that $\mathbb{E}\big[\hat{\ell}_{t_r}(a) \,\big|\, \mathcal{F}_{r-1} \big] \leq \hat{\theta}_{t_r}(s_{t_r}(a))$, where $\hat{\ell}_t(a) = \frac{\mathbb{1}[A_t = a]}{p_t(a) + \gamma_t} \hat{\theta}_t(s_t(a))$.

We can consider two cases depending on whether $\tilde{d}_{t_r} < d_{t_r}$ is true or not (and, thus, we are in the case $\tilde{d}_{t_r} = d_{t_r}$). In the first case, note that the realized losses used for computing $\hat{\theta}_{t_r}(s_{t_r}(a))$ correspond to time steps in $\mathcal{L}'_{t_r}(s_{t_r}(a))$, for which there is a corresponding tuple in $\mathcal{F}_{r-1}$. Therefore, we have that $\hat{\theta}_{t_r}(s_{t_r}(a))$ is $\mathcal{F}_{r-1}$-measurable,

and we can show that

$$\mathbb{E}\Big[\hat{\ell}_{t_r}(a)\mathbb{1}[\tilde{d}_{t_r} < d_{t_r}] \;\Big|\; \mathcal{F}_{r-1}\Big] = \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}} \;\middle|\; \mathcal{F}_{r-1}\right]\frac{\mathbb{1}[\tilde{d}_{t_r} < d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a)) \ .$$

In the second case, we have that $\tilde{d}_{t_r} = d_{t_r}$, which implies that $t_r \in \mathcal{L}'_{t_r}(s_{t_r}(a))$ in the case $A_{t_r} = a$. This means that we have a corresponding tuple in $\mathcal{F}_{r-1}$ only for rounds in $\mathcal{L}'_{t_r}(s_{t_r}(a)) \setminus \{t_r\}$. Nonetheless, this does not pose an issue since we have the indicator $\mathbb{1}[A_{t_r} = a]$, and thus $S_{t_r} = s_t(a)$. Indeed, we have that

$$\begin{aligned}
\mathbb{E}\Big[\hat{\ell}_{t_r}(a)\mathbb{1}[\tilde{d}_{t_r} = d_{t_r}] \;\Big|\; \mathcal{F}_{r-1}\Big] &= \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}}\cdot\frac{\mathbb{1}[\tilde{d}_{t_r} = d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a)) \;\middle|\; \mathcal{F}_{r-1}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}} \;\middle|\; \mathcal{F}_{r-1}\right]\frac{\mathbb{1}[\tilde{d}_{t_r} = d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{\substack{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))\\ j\neq t_r}}\ell_j(s_{t_r}(a)) \\
&\quad + \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}} \;\middle|\; \mathcal{F}_{r-1}\right]\frac{\mathbb{1}[\tilde{d}_{t_r} = d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\ell_{t_r}(s_{t_r}(a)) \\
&= \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}} \;\middle|\; \mathcal{F}_{r-1}\right]\frac{\mathbb{1}[\tilde{d}_{t_r} = d_{t_r}]}{N'_{t_r}(s_{t_r}(a))}\sum_{j\in\mathcal{L}'_{t_r}(s_{t_r}(a))}\ell_j(s_{t_r}(a))
\end{aligned}$$

and therefore the inequality

$$\mathbb{E}\Big[\hat{\ell}_{t_r}(a) \;\Big|\; \mathcal{F}_{r-1}\Big] = \mathbb{E}\left[\frac{\mathbb{1}[A_{t_r}=a]}{p_{t_r}(a)+\gamma_{t_r}} \;\middle|\; \mathcal{F}_{r-1}\right]\hat{\theta}_{t_r}(s_{t_r}(a)) \leq \hat{\theta}_{t_r}(s_{t_r}(a))$$

is true because $\mathbb{1}[\tilde{d}_t < d_t] + \mathbb{1}[\tilde{d}_t = d_t] = 1$ for all $t \in [T]$, and by definition of $\hat{\theta}_t$.

As already mentioned, this is equivalent to proving that $\mathbb{E}\big[\tilde{\ell}_{t_r}(a) \,\big|\, \mathcal{F}_{r-1}\big] \leq \tilde{\theta}_{t_r}(s_{t_r}(a))$ holds. By using a notation similar to the original proof, if we define $\tilde{\lambda}_r = \sum_{a=1}^{K}\alpha_{t_r}(a)\tilde{\ell}_{t_r}(a)$ and $\lambda_r = \sum_{a=1}^{K}\alpha_{t_r}(a)\tilde{\theta}_{t_r}(s_{t_r}(a))$, the process $(Z_r)_{r\in[T]}$ with $Z_r = \exp(\sum_{j=1}^{r}(\tilde{\lambda}_j - \lambda_j))$ is a supermartingale with respect to $(\mathcal{F}_r)_{r\in[T]}$ which has the same properties as in the proof of Neu (2015, Lemma 1). This concludes the current proof by following a similar reasoning as in the original one. □

Thanks to this result, we can conclude that the adoption of `DAda-Exp3` for the reduction via `MetaBIO` can guarantee a high-probability regret bound on $\hat{\mathcal{R}}_T^{\mathcal{B}}$ as stated in Theorem 5.1, but with total delay $\tilde{\mathcal{D}}_T = \sum_{t=1}^{T}\tilde{d}_t$ instead of $\mathcal{D}_T$.

5.8.2.4 Regret of `MetaBIO`

By Lemma 5.3, we have that

$$R_T \leq \sum_{t=1}^{T} \tilde{\theta}_t(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \tilde{\theta}_t(s_t(a)) + \sum_{t=1}^{T} \varepsilon_t(S_t) = \hat{\mathcal{R}}_T^{\mathcal{B}} + \sum_{t=1}^{T} \varepsilon_t(S_t) \tag{5.17}$$

with probability at least $1-\delta/2$, where $\hat{R}_T^{\mathcal{B}}$ (Equation (5.12)) is the regret of algorithm $\mathcal{B}$ when fed with $(\tilde{\theta}_t \circ s_t)_{t\in[T]}$ as losses.

Lemma 5.6. *Conditioning on the event as stated in Lemma 5.3, the sum of errors suffered from* `MetaBIO` *by using the loss estimates* $(\tilde{\theta}_t)_{t\in[T]}$ *from Equations* (5.1) *and* (5.2) *is*

$$\sum_{t=1}^{T} \varepsilon_t(S_t) \leq \left(4 + 2\sqrt{2}\right) \sqrt{ST \ln\left(\frac{4ST}{\delta}\right)} \ .$$

Proof. First, observe that we can rewrite the sum of errors as

$$\sum_{t=1}^{T} \varepsilon_t(S_t) = \sum_{t=1}^{T} \varepsilon_t(S_t) \mathbb{1}[\tilde{d}_t < d_t] + \sum_{t=1}^{T} \varepsilon_t(S_t) \mathbb{1}[\tilde{d}_t = d_t] \ .$$

We now provide an upper bound for the first sum of errors. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s = \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s. We can bound it as

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t) \mathbb{1}[\tilde{d}_t < d_t] &= \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}_s} \varepsilon_t(s) \mathbb{1}[\tilde{d}_t < d_t] \\
&= \sqrt{2 \ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}_s} \sqrt{\frac{1}{N'_t(s)}} \mathbb{1}[\tilde{d}_t < d_t] \\
&\leq 2\sqrt{\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}_s} \sqrt{\frac{1}{M_t(s)}} \mathbb{1}[\tilde{d}_t < d_t] \\
&\qquad \text{(because } N'_t(s) \geq \tfrac{1}{2} M_t(s)\text{)} \\
&\leq 4\sqrt{\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sqrt{M_T(s)} \\
&\qquad \text{(since } M_t(s) \text{ is increasing over } \mathcal{T}_s\text{)} \\
&\leq 4\sqrt{ST \ln\left(\frac{4ST}{\delta}\right)} \ ,
\end{aligned}$$

where the second inequality holds because $N'_t(S_t) = N_t(S_t) \geq \frac{1}{2}M_t(S_t)$ when $\tilde{d}_t < d_t$ since $M_t(S_t) \leq N_t(S_t) + \sigma_t$, while the last one follows by Jensen's inequality and the fact that $\sum_{s\in\mathcal{S}} M_T(s) = T$.

As a last step, we provide an upper bound to the second sum. Let $J_s = \{r \in \mathcal{T}_s : \tilde{d}_r = d_r\}$ and notice that $|J_s| \leq |\mathcal{T}_s| = M_T(s)$. Observe that $\rho(t) = \tilde{\rho}(t)$ for each round t such that $\tilde{d}_t = d_t$, and thus by Equation (5.14) we have that

$$\pi(r) < \pi(t) \iff \rho(r) < \rho(t) \vee (\rho(r) = \rho(t) \wedge r < t)$$

for all $r, t \in [T]$ such that $\tilde{d}_r = d_r$ and $\tilde{d}_t = d_t$. Define $\nu_s : J_s \to \big[|J_s|\big]$ by

$$\nu_s(t) = |\{r \in J_s \,:\, \pi(r) \leq \pi(t)\}| \qquad \forall t \in J_s \ .$$

Observe that $\nu_s(t) \leq N'_t(s) = |\mathcal{L}'_t(s)|$ for all $s \in \mathcal{S}$ and all $t \in J_s$. This is due to the fact that $\nu_s(t)$ counts a subset of $\mathcal{L}'_t(s)$; to be precise, we have that $\nu_s(t) = |\mathcal{L}'_t(s) \cap J_s|$. Moreover, notice that the condition $\pi(r) \leq \pi(t)$ defines a total order over J_s. Hence, $\nu_s(t)$ counts the number of elements of J_s preceding $t \in J_s$ (including t itself) in this total order. This implies that ν_s is a bijection between J_s and $\big[|J_s|\big]$. Then, using a similar reasoning as before, we show that

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{1}[\tilde{d}_t = d_t] &= \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in\mathcal{T}_s} \sqrt{\frac{1}{N'_t(s)}}\mathbb{1}[\tilde{d}_t = d_t] \\
&= \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in J_s} \sqrt{\frac{1}{N'_t(s)}} && \text{(by definition of } J_s\text{)} \\
&\leq \sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sum_{t\in J_s} \sqrt{\frac{1}{\nu_s(t)}} \\
& && \text{(since } \nu_s(t) \leq N'_t(s) \text{ for } t \in J_s\text{)} \\
&\leq 2\sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sqrt{|J_s|} && \text{(since } \nu_s(t) \text{ is bijective)} \\
&\leq 2\sqrt{2\ln\left(\frac{4ST}{\delta}\right)} \sum_{s\in\mathcal{S}} \sqrt{M_T(s)} && \text{(since } |J_s| \leq M_T(s)\text{)} \\
&\leq 2\sqrt{2ST\ln\left(\frac{4ST}{\delta}\right)} \ . && \text{(by Jensen's inequality)}
\end{aligned}$$

□

Theorem 5.2. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *using* `DAda-Exp3`*, then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T \leq 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST\ln\frac{4ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{4}{\delta}\ . \tag{5.6}$$

Proof of Theorem 5.2. By Equation (5.17), the regret R_T can be bounded as

$$R_T \leq \hat{\mathcal{R}}_T^{\mathcal{B}} + \sum_{t=1}^{T}\varepsilon_t(S_t) \leq \hat{\mathcal{R}}_T^{\mathcal{B}} + 7\sqrt{ST\ln\frac{4ST}{\delta}}$$

with probability at least $1-\delta/2$, where the last inequality follows by Lemma 5.6. From what we argued in Section 5.8.2.3, we can upper bound $\hat{\mathcal{R}}_T^{\mathcal{B}}$ using the high-probability regret bound of `DAda-Exp3`. Notice that the delays incurred by `DAda-Exp3` via `MetaBIO` are those given when providing the estimates $(\tilde{\theta}_t)_{t\in[T]}$. We denote these delays by $\tilde{d}_1,\dots,\tilde{d}_T$, and the total delay perceived by `DAda-Exp3` is thus $\tilde{\mathcal{D}}_T = \sum_{t=1}^{T}\tilde{d}_t$. Hence, from the improved bound for `DAda-Exp3` in Equation (5.9), we have that

$$\hat{\mathcal{R}}_T^{\mathcal{B}} \leq 2\sqrt{2KT(3\ln(K)+\ln(4/\delta))} + 2\sqrt{\tilde{\mathcal{D}}_T(3\ln(K)+\ln(4/\delta))} + \frac{\sigma_{\max}+2}{2}\ln(4/\delta)$$

holds with probability at least $1-\delta/2$. The combination of the above two inequalities, together with Lemma 5.1, concludes the proof. □

5.8.2.5 Regret of `MetaAdaBIO`

Theorem 5.3. *Let $\delta \in (0,1)$. If we run* `MetaAdaBIO` *with* `DAda-Exp3`*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses with probability at least $1-\delta$ satisfies*

$$R_T \leq 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta}\ . \tag{5.7}$$

Proof of Theorem 5.3. Let $t^* \in [T]$ be the last round before `MetaAdaBIO` switches from `DAda-Exp3` to `MetaBIO`, i.e., the last round that satisfies $\mathfrak{D}_{t^*}C_{K,4\delta} \leq$

$49ST \ln \frac{8ST}{\delta}$. Then, define $a^* \in \operatorname{argmin}_a \sum_{t=1}^{T} \theta(s_t(a))$. We may decompose regret as

$$\begin{aligned} R_T &= \sum_{t=1}^{t^*} \Big(\theta(S_t) - \theta(s_t(a^*)) \Big) + \sum_{t=t^*+1}^{T} \Big(\theta(S_t) - \theta(s_t(a^*)) \Big) \\ &\leq \underbrace{\sum_{t=1}^{t^*} \theta(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{t^*} \theta(s_t(a))}_{R_{t^*}} + \underbrace{\sum_{t=t^*+1}^{T} \theta(S_t) - \min_{a \in \mathcal{A}} \sum_{t=t^*+1}^{T} \theta(s_t(a))}_{R_{t^*:T}} \ . \end{aligned}$$

The incurred delay until time t^* is $\mathfrak{D}_{t^*}$. Thus, from Equation (5.11), we get that the following bound

$$R_{t^*} \leq 2\sqrt{2Kt^* C_{K,2\delta}} + \sqrt{2t^* \ln \frac{12}{\delta}} + 2\sqrt{\mathfrak{D}_{t^*} C_{K,2\delta}} + \frac{\sigma_{\max} + 2}{2} \ln \frac{6}{\delta} \tag{5.18}$$

holds with probability at least $1 - \delta/2$, where we recall that $C_{K,\delta} = 3 \ln K + \ln(12/\delta)$. If our algorithm never switches, then $t^* = T$ and we get the bound in (5.18) for R_T. Note that this is no greater than the upper bound in the statement as $\sqrt{\mathfrak{D}_T C_{K,2\delta}} \leq 7\sqrt{ST \ln(8ST/\delta)}$ by definition of t^* in this case.

Otherwise, we use the switching condition $\sqrt{\mathfrak{D}_{t^*} C_{K,2\delta}} \leq 7\sqrt{ST \ln(8ST/\delta)}$ along with the fact that $\sqrt{t^* \ln(12/\delta)} \leq \sqrt{Kt^* C_{K,2\delta}}$ to get

$$R_{t^*} \leq 3\sqrt{2Kt^* C_{K,2\delta}} + 14\sqrt{ST \ln \frac{8ST}{\delta}} + \frac{\sigma_{\max} + 2}{2} \ln \frac{6}{\delta} \ . \tag{5.19}$$

Furthermore, Theorem 5.2 directly gives us an upper bound for $R_{t^*:T}$ since `MetaAdaBIO` runs `MetaBIO` for $t > t^*$ with the confidence parameter set to $\delta/2$. We just need to bound the total incurred delays of these rounds, namely $\tilde{\mathcal{D}}_{t^*:T}$. Let σ'_t be the outstanding observations for any round $t > t^*$ as perceived by the execution of `MetaBIO` starting after round t^*, that is, when considering only delays $(d_t)_{t>t^*}$. It is immediate to observe that $\sigma'_t \leq \sigma_t$ and thus $\max_{t>t^*} \sigma'_t \leq \max_{t>t^*} \sigma_t$. Moreover, from Lemma 5.1 we have

$$\tilde{\mathcal{D}}_{t^*:T} \leq \mathcal{D}_{\Phi'} \ ,$$

where Φ' denotes a set of $\min\{T - t^*, 2S\sigma'_{\max}\}$ rounds with the largest delays among $(d_t)_{t>t^*}$, with $\sigma'_{\max} = \max_{t>t^*} \sigma'_t$. So we have

$$\mathcal{D}_{\Phi'} \leq \mathcal{D}_{\Phi}$$

due to the fact that $|\Phi'| = \min\{T - t^*, 2S\sigma'_{\max}\} \leq \min\{T, 2S\sigma_{\max}\} = |\Phi|$. Therefore, from Theorem 5.2 we obtain

$$R_{t^*:T} \leq 2\sqrt{2K(T-t^*)C_{K,3\delta}} + 7\sqrt{ST\ln\frac{8ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{8}{\delta} \quad (5.20)$$

with probability at least $1 - \delta/2$. We conclude the proof by combining Equations (5.19) and (5.20) along with the fact that $\sqrt{t^*} + \sqrt{T-t^*} \leq \sqrt{2T}$ to get that the bound

$$R_T \leq 6\sqrt{KTC_{K,2\delta}} + 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta}$$

holds with probability at least $1 - \delta$. □

5.8.2.6 Expected Regret Analysis of `MetaAdaBIO` with `Tsallis-INF`

Proposition 5.4. *If we execute* `MetaAdaBIO` *with* `Tsallis-INF` *(Zimmert and Seldin, 2020), and use the switching condition* $\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST\ln(2ST)}$ *at each round* $t \in [T]$*, where* $\mathfrak{D}_t = \sum_{j=1}^{t}\sigma_j$*, then the regret of* `MetaAdaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\mathbb{E}[R_T] \leq 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 2\min\left\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T \ln K}\right\} .$$

Proof of Proposition 5.4. We begin by studying of expected regret of `MetaBIO` and we then give a regret analysis of `MetaAdaBIO`. When running `MetaBIO`, we use the unbiased empirical mean estimators $(\hat{\theta}_t)_{t\in[T]}$ as the mean loss estimates, rather than the lower confidence bounds $(\tilde{\theta}_t)_{t\in[T]}$. The expected regret is defined as

$$\mathbb{E}[R_T] = \sum_{t=1}^{T}\mathbb{E}[\theta(S_t)] - \sum_{t=1}^{T}\theta(s_t(a^*)) \ ,$$

where $a^* = \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a))$. Here we use a version of `Tsallis-INF` that is tailored for the delayed bandits problem (Zimmert and Seldin, 2020), which guarantees a bound in expectation on the regret

$$\hat{\mathcal{R}}_T^{\text{Tsallis}}(a) = \sum_{t=1}^{T}\hat{\theta}_t(S_t) - \sum_{t=1}^{T}\hat{\theta}_t(s_t(a))$$

against any fixed action $a \in \mathcal{A}$, using the loss estimates $\{\hat{\theta}_t\}_{t\in[T]}$. Observe that this regret is defined in terms of our estimates, as required in our case. By Zimmert and Seldin (2020, Theorem 1), `Tsallis-INF` guarantees that its expected regret is

$$\begin{aligned}\mathbb{E}\Big[\hat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] &= \mathbb{E}\left[\sum_{t=1}^{T}\hat{\theta}_t(S_t) - \sum_{t=1}^{T}\hat{\theta}_t(s_t(a^*))\right] \\ &\leq 4\sqrt{KT} + \sqrt{8\tilde{\mathcal{D}}_T \ln K} \leq 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K}\ ,\end{aligned}$$

where the last inequality uses Lemma 5.1. Then, we can focus on our notion of regret and use the above regret bound to obtain that

$$\begin{aligned}\mathbb{E}[R_T] =& \mathbb{E}\Big[R_T - \hat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] + \mathbb{E}\Big[\hat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] \\ =& \mathbb{E}\left[\sum_{t=1}^{T}\big(\theta(S_t) - \hat{\theta}_t(S_t)\big)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\big(\hat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\big)\right] + \mathbb{E}\Big[\hat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\Big] \\ \leq& \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\big(\theta(S_t) - \hat{\theta}_t(S_t)\big)\right]}_{\Delta} + \mathbb{E}\left[\sum_{t=1}^{T}\big(\hat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\big)\right] \\ &+ 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K}\ .\end{aligned} \tag{5.21}$$

We know that our mean estimator is unbiased. Therefore, we have that $\mathbb{E}[\hat{\theta}_t(s_t(a^*))] = \theta(s_t(a^*))$ for any $t \in [T]$, meaning that the second term in the right-hand side of (5.21) is equal to zero.

On the other hand, we can apply Lemma 5.3 to get the following bound for Δ that holds with probability at least $1 - \delta/2$ for any $\delta \in (0, 1)$:

$$\Delta \leq \min\left\{\frac{1}{2}\sum_{t=1}^{T}\varepsilon_t(S_t), T\right\}\ , \tag{5.22}$$

where we recall that $\varepsilon_t(s) = \sqrt{\frac{2}{N_t'(s)}\ln\frac{4ST}{\delta}}$. In particular, the inequality $\Delta \leq T$ is true in general. By Lemma 5.6, we can bound the right-hand side of (5.22) as

$$\frac{1}{2}\sum_{t=1}^{T}\varepsilon_t(S_t) \leq \frac{7}{2}\sqrt{ST\ln\frac{4ST}{\delta}}$$

when conditioning on the event as in the statement of Lemma 5.3. If we denote such an event as $\mathcal{E}$, we have that $\mathbb{P}\big(\overline{\mathcal{E}}\big) \leq \delta/2$ and that $\mathbb{E}[\Delta \mid \mathcal{E}] \leq \frac{7}{2}\sqrt{ST\ln(4ST/\delta)}$. As

a consequence, we notice that

$$\mathbb{E}[\Delta] = \mathbb{E}[\Delta \mid \mathcal{E}]\mathbb{P}(\mathcal{E}) + \mathbb{E}\big[\Delta \mid \overline{\mathcal{E}}\big]\mathbb{P}\big(\overline{\mathcal{E}}\big) \leq \frac{7}{2}\sqrt{ST \ln \frac{4ST}{\delta}} + \frac{\delta}{2}T \leq 5\sqrt{ST \ln(2ST)} + 1$$

where in the last inequality we set $\delta = 2/T$. Since we assume that $S \geq 2$, we can easily observe that $\mathbb{E}[\Delta] \leq 6\sqrt{ST \ln(2ST)}$. Plugging this into Equation (5.21) gives us

$$\mathbb{E}[R_T] \leq 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST \ln(2ST)} \ . \tag{5.23}$$

At this point, we can proceed to the proof of the overall bound on the expected regret of `MetaAdaBIO`. The behaviour of `MetaAdaBIO` follows the same principle as before, but the switching condition is different:

$$\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST \ln(2ST)} \ .$$

Similar to the analysis of `MetaAdaBIO` in Section 5.8.2.5, we decompose the regret into

$$\mathbb{E}[R_T] \leq \underbrace{\sum_{t=1}^{t^*} \mathbb{E}[\theta(S_t)] - \min_{a \in \mathcal{A}} \sum_{t=1}^{t^*} \theta(s_t(a))}_{R_{t^*}} + \underbrace{\sum_{t=t^*+1}^{T} \mathbb{E}[\theta(S_t)] - \min_{a \in \mathcal{A}} \sum_{t=t^*+1}^{T} \theta(s_t(a))}_{R_{t^*:T}} \ ,$$

where t^* is the last round satisfying $\sqrt{8\mathfrak{D}_{t^*}} \leq 6\sqrt{ST \ln(2ST)}$. Then, we have

$$\mathbb{E}[R_{t^*}] \leq 4\sqrt{Kt^*} + \sqrt{8\mathfrak{D}_{t^*} \ln K} \ . \tag{5.24}$$

If $t^* = T$ then $R_{t^*} = R_T$ and we get the bound in (5.24), where we note that $\sqrt{8\mathfrak{D}_T \ln K} \leq 6\sqrt{ST \ln(2ST)}$ by definition of t^* in this case, and we can replace $\mathfrak{D}_T$ by $\mathcal{D}_T$. Otherwise, $t^* < T$ and we can apply the bound for `MetaBIO` from (5.23), along with the fact that the total incurred delay after round t^* is upper bounded by $\mathcal{D}_\Phi$, in order to derive an upper bound for $\mathbb{E}[R_{t^*:T}]$ that is

$$\mathbb{E}[R_{t^*:T}] \leq 4\sqrt{K(T - t^*)} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST \ln(2ST)} \ . \tag{5.25}$$

Finally, if we use the fact that $\sqrt{8\mathfrak{D}_{t^*}} \leq 6\sqrt{ST \ln(2ST)}$ (by definition of t^*) in (5.24), and combine it with (5.25), we conclude that

$$\mathbb{E}[R_T] \leq 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 2\min\Big\{6\sqrt{ST \ln(2ST)}, \sqrt{8\mathcal{D}_T \ln K}\Big\} \ ,$$

where we also used the fact that $\sqrt{t^*} + \sqrt{T - t^*} \leq \sqrt{2T}$. □

5.8.3 Proofs for the Lower Bounds

Theorem 5.6. *Irrespective to whether the action-state mappings and loss vectors are stochastic or adversarial, there exists a sequence of losses such that any (possibly randomized) algorithm in BIO suffers regret* $\mathbb{E}[R_T] = \Omega\big(\sqrt{KT}\big)$.

Proof. Our construction only uses two states h_1 and h_2. The loss vectors, which are deterministic and do not change over time, are defined as follows: $\ell_t(h_1) = 1$ and $\ell_t(h_2) = 0$ for all $t \geq 0$. The stochastic action-state mapping, which is also constant over time, is given by

$$s_t(a) = \begin{cases} h_1 & \text{with probability } p_a \\ h_2 & \text{with probability } 1 - p_a \end{cases}$$

for all $a \in \mathcal{A}$ and $t \geq 0$, where the probabilities p_a are to be determined. Thus, the loss of an arm a is $\ell_t(s_t(a)) = \ell_t(h_1) = 1$ with probability p_a and $\ell_t(s_t(a)) = \ell_t(h_2) = 0$ with probability $1 - p_a$. Since the loss is determined by the state, the learner receives bandit feedback without delay. We can then choose p_a for $a \in \mathcal{A}$ to mimic the standard $\Omega\big(\sqrt{KT}\big)$ distribution-free bandit lower bound—e.g., see Slivkins et al. (2019, Chapter 2). By Yao's minimax principle, the same lower bound also applies to the case with adversarial action-state mappings. Since the loss vectors are deterministic, this covers all possible cases in BIO. □

Theorem 5.7. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that* $d_t = d$ *for all* $t \in [T]$. *If* $T \geq \min\{S, d\}$ *then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret* $\mathbb{E}[R_T] = \Omega\big(\sqrt{\min\{S, d\}T}\big)$.

Proof of Theorem 5.7. Assume without loss of generality that $K = 2$ and let $\mathcal{S} = \{h_1, \dots, h_S\}$ be the finite set of possible states. Let $S' = \lfloor \min\{S/2, d\} \rfloor$ and let $I_1, \dots, I_T$ be the actions chosen by the considered algorithm. Split the T time steps into $m = \lfloor T/S' \rfloor$ blocks $B_1, \dots, B_m$ of equal size S', eventually leaving $\leq S' - 1$ extra time steps. We assume with no loss of generality that the last step corresponds to the end of the m-th block. The feedback formed by the losses of the actions chosen by the algorithm in a certain block is received only after the last time step of the same block since $S \leq 2d$. Define $b_i = (i-1)S' + 1$ for all $i \in [m]$. We assume that the learner receives *all* the realized losses $\ell_t(s_t(A))$ for all $t \in B_i$ and all $A \in \{1, 2\}$ at the end of each block, which means that we are in a full information setting, as this only helps the algorithm.

Now, we define a specific sequence of assignments from actions to states, and construct losses so that the expected regret becomes sufficiently large. Let $s_t(A) =$

$h_{2(t-b_i)+A}$ for all $t \in B_i$, all $i \in [m]$ and all $A \in \{1,2\}$; this means that, for the first time step of any block, actions 1 and 2 will be assigned to states h_1 and h_2 respectively, then to h_3 and h_4 respectively in the next time step of the same block, and so on. Let $\varepsilon = \frac{1}{4}\sqrt{\frac{S'}{2T\ln(4/3)}} \in [0, \frac{1}{4}]$ and let $\theta^{(A)} \in \mathbb{R}^2$ be a vector of mean losses such that $\theta_i^{(A)} = \frac{1}{2} - \mathbb{I}\{i = A\}\varepsilon$, for each $A \in \{1,2\}$. We simplify the notation with $\mathbb{E}_A[\cdot] = \mathbb{E}\big[\cdot \,\big|\, \theta^{(A)}\big]$ and $\mathbb{P}_A(\cdot) = \mathbb{P}\big(\cdot \,\big|\, \theta^{(A)}\big)$, where the conditioning on $\theta^{(A)}$ means that we sample losses for each state assigned to $i \in \{1,2\}$ such that they are Bernoulli random variables with mean $\theta_i^{(A)}$. In particular, conditioning on $\theta^{(A)}$, we sample independent Bernoulli random variables $X_1^i, \ldots, X_m^i$ with mean $\theta_i^{(A)}$, one for each block, for $i \in \{1,2\}$. Then, the losses are defined as $\ell_t(s_t(i)) = X_j^i$ for each $t \in B_j$ and each $j \in [m]$.

We can now proceed to show a lower bound for the expected pseudo-regret. Let T_i be the number of times the learner chooses action i over all T time steps. The expected pseudo-regret over the two instances determined by $\theta^{(k)}$ for $k \in \{1,2\}$ adds up to

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] = \varepsilon(2T - \mathbb{E}_1[T_1] - \mathbb{E}_2[T_2]) \ .$$

Following the standard analysis, we show that the difference $\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2]$ is such that

$$\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2] \leq T \cdot d_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_1) \leq T\sqrt{\frac{1}{2}D_{\mathrm{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2)} \ ,$$

where the last step follows by Pinsker's inequality.

Let $\lambda_i = \{(I_t, \ell_t(S_t(1)), \ell_t(S_t(2))) \mid t \in B_i\}$ be the feedback set known to the learner by the end of block B_i, and let $\lambda^i = (\lambda_1, \ldots, \lambda_i)$ be the tuple of all feedback sets up to the end of block B_i. Denote by $\mathbb{P}_{k,i}(\cdot)$ the probability measure of feedback tuples λ^i conditioned on $\theta^{(A)}$. By the chain rule for the relative entropy, we can observe that

$$\begin{aligned} D_{\mathrm{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) &= \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1(\lambda^{i-1}) D_{\mathrm{KL}}(\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1}) \| \, \mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1})) \\ &\leq \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1(\lambda^{i-1}) 16\varepsilon^2 \ln(4/3) \\ &= 16m\varepsilon^2 \ln(4/3) \ , \end{aligned}$$

where we used the fact that each relative entropy $D_{\mathrm{KL}}(\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1}) \,\|\, \mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1}))$ corresponds to the sum of the relative entropy between two Bernoulli distributions with means $1/2$ and $1/2 - \varepsilon$ and that between Bernoulli distributions with means

$1/2-\varepsilon$ and $1/2$, respectively, which is upper bounded by $16\varepsilon^2 \ln(4/3)$ for $\varepsilon \in [0, 1/4]$. This follows by an application of the chain rule for the relative entropy, as well as from the fact that the distribution of I_t is the same under both $\mathbb{P}_{1,i}(\cdot \mid \lambda^{i-1})$ and $\mathbb{P}_{2,i}(\cdot \mid \lambda^{i-1})$, for all $t \in B_i$ and any λ^{i-1}. Therefore, we have that

$$\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2] \leq 2\varepsilon T \sqrt{2m \ln(4/3)}$$

which also implies that

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] \geq \varepsilon T \left(1 - 2\varepsilon \sqrt{2\frac{T}{S'} \ln(4/3)}\right) = \frac{\varepsilon T}{2} \geq \frac{1}{8}\sqrt{\frac{\lfloor S/2 \rfloor T}{2\ln(4/3)}} \geq \frac{1}{8}\sqrt{\frac{ST}{6\ln(4/3)}} \,,$$

where we used the facts that $m \leq T/S'$ and that $\lfloor S/2 \rfloor \geq S/3$ for any integer $S \geq 2$. This means that the expected pseudo-regret of the learner has to be $\frac{1}{16}\sqrt{\frac{ST}{6\ln(4/3)}}$ at least in one of the two instances. Now, for $S > 2d$ we use the same construction, but now we only use $2d$ states, which leads to the promised $\Omega(\sqrt{\min\{S, d\}T})$ lower bound. □

Theorem 5.8. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq d+1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big)$.*

Proof of Theorem 5.8. Let $S' = \min\{\lfloor \frac{S}{2} \rfloor, \lfloor \frac{T}{d+1} \rfloor\} \geq 1$. We consider the first $(d+1)S'$ rounds of the game and divide them into S' blocks $B_1, \ldots, B_{S'}$ of same length $d+1$. In this way, we ensure that the feedback for any time step in some block is revealed to the learner only after its final round.

Without loss of generality, we can assume that the learner observes all the losses of one block immediately after its last time step; this only helps the learner since they would observe only the incurred losses at possibly later rounds otherwise. We can further simplify the problem by assuming that losses are deterministic functions of the states, i.e., $\ell_t \equiv \theta$ for every round t. This also means that the problem turns into an easier, full-information version of our problem with deterministic losses. Now, let the adversary choose the action-state mappings such that for each block index i and each action $a \in \mathcal{A}$, $S_t(a) = S_{t'}(a) \in \{s_{2i-1}, s_{2i}\}$ for all $t, t' \in B_i$. Furthermore, we assume that the losses are chosen such that $\theta(s_{2i-1}) \in \{0, 1\}$ and $\theta(s_{2i}) = 1 - \theta(s_{2i-1})$ for all $i \in [S']$. In this construction, the learner cannot obtain any useful information from the states of a block because of the delays. Moreover, the states observed in one block are not observed again in the other blocks.

It thus suffices to prove a lower bound for a standard full information game with S' rounds and loss range $[0, d+1]$. Hence, we can conclude that the expected regret of any algorithm has to be

$$\mathbb{E}[R_T] = \Omega\Big((d+1)\sqrt{S'}\Big) = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) \ .$$

□

Theorem 5.9. *Suppose that the action-state mapping is stochastic, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists a stochastic action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret $\mathbb{E}[R_T] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.*

Proof. Since by Theorem 5.6 we already know that any algorithm must suffer $\Omega\big(\sqrt{KT}\big)$ regret, we only need to show a $\Omega(\sqrt{dT})$ lower bound. We use two states, h_1 and h_2. Our action-state mapping is deterministic and, for all $t \geq 0$, assigns $s_t(a) = h_1$ to all but one action $a^\star$, to which the mapping assigns $s_t(a^\star) = h_2$. We now have constructed a two-armed bandit problem with delayed feedback and T rounds, for which a $\Omega(\sqrt{dT})$ lower bound is known (Cesa-Bianchi et al., 2019). □

5.8.4 Action-State Mappings and Loss Means Used in the Experiments

Table 5.1 and Table 5.2 describe the instances used to generate the data for the experiments of Section 5.6.

Mean loss	$s=1$	$s=2$	$s=3$
$\theta(s)$	0.2	0.4	0.8
Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	0.8	0.1	0.1
$a=2$	0.4	0.5	0.1
$a=3$	0.3	0.7	0.0
$a=4$	0.5	0.3	0.2

Table 5.1: Mean losses and stochastic action-state mapping for Experiment 1 in Section 5.6.

Mean loss	$s = 1$	$s = 2$	$s = 3$
$\theta(s)$	0	1	1

Environment 1

Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a = 1$	0.06	0.47	0.47
$a = 2$	0	0.50	0.50
$a = 3$	0	0.50	0.50
$a = 4$	0	0.50	0.50

Environment 2

Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a = 1$	1	0	0
$a = 2$	0.94	0.03	0.03
$a = 3$	0.94	0.03	0.03
$a = 4$	0.94	0.03	0.03

Table 5.2: Mean losses and stochastic action-state mappings for Experiment 2 in Section 5.6.

Chapter 6

Summary and Discussion

The thesis represents a significant advancement in the field of multi-armed bandits by addressing two critical challenges: the need for "best-of-both-worlds" guarantees and the effective handling of "delayed feedback" in real-world applications, as they closely mirror the complexities present in various practical scenarios.

In Chapter 2, we introduced an enhanced analysis of the Tsallis-INF algorithm within adversarial regimes with self-bounding constraint. With a corruption budget denoted as C, our enhancements culminates at $C = \mathcal{O}\big(\frac{T}{\log T}\big)$, resulting in a regret reduction by a factor of $\frac{\log T}{\log \log T}$. In stochastically constrained adversarial regime, it further refines regret by substituting $\log T$ with $\log\left(\frac{(K-1)T}{\left(\sum_{i\neq i^*}\frac{1}{\Delta_i}\right)^2}\right)$. Similar to Zimmert and Seldin (2021), our analysis relies on the uniqueness of the best arm. However, it's worth noting that Ito (2021) provide a new technique to remove this assumption. We conjecture that our derived bound has also a matching lower bound for adversarial regimes with a self-bounded constraint, as it is already optimal in extreme cases such as fully stochastic and fully adversarial regimes, but we leave it to future work.

In Chapter 3, we established the first-ever "best-of-both-worlds" guarantee for delayed bandits. We proposed adaptation to the algorithm of Zimmert and Seldin (2020) that relies on the knowledge of $d_{\max}$ to control the drift of arms distribution. The control of this distribution drift constitutes the core of our best-of-both-worlds analysis. This aspect presents the most challenging part of the analysis due to the dynamically changing learning rate. Furthermore, we established the optimality of the regret derived by Zimmert and Seldin (2020) for the adversarial regime, substantiating this claim with a corresponding lower bound.

In Chapter 4, we empowered the algorithm introduced in Chapter 3 with two

techniques: the *skipping technique* and *implicit exploration*, which allows us to eliminate the necessity for prior knowledge of $d_{\max}$. Furthermore, we showed that the contribution of delays in the regret, always appears as the *maximum amount* of observation we are missing ($\sigma_{\max}$), rather than the *maximum waiting time* for these missing information ($d_{\max}$). An intriguing question remains regarding the possibility of eliminating all multiplicative factors tied to $\sigma_{\max}$, especially $\sum_{i \neq i^*} \frac{1}{\Delta_i}$.

In Chapter 5, we explored "intermediate observations" as a means to mitigate the impact of delays in the problem of bandits with arbitrary delays. We demonstrated that the complexity of this problem lies only behind the nature of state-loss mappings. We proved that while intermediate observations offer no benefits in adversarial state-loss mappings, they bring significant advantages in scenarios involving stochastic state-loss mappings, where the dependence of the regret bound on delay can be replaced by the number of states. Our algorithm is based on a novel reduction strategy that could be extended to other regimes, such as non-stationary bandits if there exists an algorithm for the delayed non-stationary bandits. An interesting problem for future work is to provide an algorithm for delayed non-stationary bandits to be able to apply our reduction idea, enabling a comparative assessment against the regret bounds of the Vernade et al. (2020).

List of Publications

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

1. 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 Conference on Learning Theory (COLT)*, 2021.

2. 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.

3. Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. An improved best-of-both-worlds algorithm for bandits with delayed feedback. `https://arxiv.org/abs/2308.10675`, 2023.

4. Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Bibliography

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2015.

Jean-Yves Audibert and Sébastien Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, 11, 2010.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

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

Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32, 2002b.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online exp3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No weighted-regret learning in adversarial bandits with delays. *Journal of Machine Learning Research*, 2022.

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

Nicolo Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Journal of Machine Learning Research*, 2019.

Ioannis Chatzigeorgiou. Bounds on the lambert function and their application to the outage analysis of user cooperation. *IEEE Communications Letters*, 17, 2013.

Stephen G. Eick. The two-armed bandit with delayed responses. *The Annals of Statistics*, 1988.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk Van Der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2019.

Andras Gyorgy and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Dirk Van der Hoeven and Nicolò Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

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

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, Tal Lancewicki, Haipeng Luo, Yishay Mansour, and Aviv Rosenberg. Near-optimal regret for adversarial MDP with delayed bandit feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 2018.

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

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing*, 2018.

Timothy Arthur Mann, Sven Gowal, András György, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

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 Conference on 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.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. An improved best-of-both-worlds algorithm for bandits with delayed feedback. `https://arxiv.org/abs/2308.10675`, 2023.

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

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 2014.

Francesco Orabona. A modern introduction to online learning. `https://arxiv.org/abs/1912.13213v1`, 2019.

Francesco Orabona. A modern introduction to online learning. `https://arxiv.org/abs/1912.13213v5`, 2022.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 1952.

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

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

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the Conference on 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.

Richard Simon. Adaptive treatment assignment methods and clinical trials. *Biometrics*, 33, 1977.

Aleksandrs Slivkins et al. 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*, 25, 1933.

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*, 52, 1988.

Claire Vernade, András György, and Timothy A. Mann. Non-stationary delayed bandits with intermediate observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings on 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 on 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.