UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD Thesis

Tobias Sommer Thune

Exploiting Easiness and Overcoming Delays in Online Learning

Advisor: Yevgeny Seldin

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on May 9th 2020

Abstract

In recent decades the field of online machine learning has enjoyed an increasing interest both from applications and significant theoretical advances. Especially models based on the framework of multi-armed bandits and prediction with expert advice have been successful in uncovering central dynamics of sequential decision problems. In this thesis we explore two learning settings within this framework and design adaptive learning algorithms with theoretical performance guarantees.

One learning setting we explore falls within the general goal of exploiting structure or easiness when present without sacrificing performance if the problem is hard. We first show that one can circumvent the recently discovered impossibility of exploiting a small effective loss range in multi-armed bandits by having access to one additional point of feedback. We further show that our algorithm simultaneously exploit stochastic data achieving constant regret. Our algorithm requires no prior information of the learning scenario rendering it robust to the unconstrained adversarial setting.

The second setting we study is the within multi-armed bandits with delayed feedback. This model covers realistic streaming scenarios where the central timing assumption of multi-armed bandits is violated, including applications where the incoming data is batched. We prove a recent conjecture of the regret scaling when the delays are arbitrary and the losses are adversarially generated. We generalize the analytical approach of algorithmic stability for delayed feedback and uncover a data dependence that generalizes previous dependencies on the delay. We design a meta algorithm that skips excessively delayed feedback and alleviates the data dependence, thereby proving the conjectured regret bound for any delay sequence. Finally for a slightly easier setting we design a novel, adaptive tuning scheme with which our algorithm can perform much better than conjectured. We provide examples of settings where this is the case and the dependence of the problem parameters is polynomially better.

Resume in Danish

Online machine learning har som forskningfelt nydt stor interesse de seneste årtier både gennem anvendelser og i kraft af væsentlige teoretiske landvindinger. Især modeller baseret på multi-armed bandits og prediction with expert advice har bidraget til forståelsen af centrale strukturer indenfor sekventielle problemer. Denne afhandling udforsker to læringsscenarier dækket af sådanne modeller, og vi konstruerer læringsalgoritmer med tilhørende teoretiske garantier for deres præstationsevne.

Et af de udforskede læringsscenarier falder inden for det generelle mål at udnytte eventuelle strukturer, der gør læringsproblemer lette uden at opgive garantier for svære problemer. Vi viser først, at det er muligt at omgå en nyligt påvist umulighed inden for multi-armed bandits. Her er det umuligt at udnytte en lille effektiv fejlrækkevide, mens det er muligt med yderligere adgang til et punkts ekstra feedback. Vi viser endvidere, at vores algoritme kan udnytte stokastiske scenarier og opnå konstant regret. Vores algoritme kræver ikke kendskab til scenariet på forhånd, hvilket gør den robust overfor fuldt modarbejdende scenarier.

Det andet scenarie vi undersøger i afhandlingen er multi-armed bandits med forsinket feedback, hvilket er en forskningsretning, der særligt er opblomstret de seneste år. Modellen dækker realistiske online scenarier, hvor den centrale rækkefølge i multi-armed bandits ikke holder inklusiv anvendelser, hvor data processeres i partier. Vi påviser en formodning omkring skaleringen af regret, når forsinkelserne er arbitrære. Vi generaliserer algoritmisk stabilitet som en analytisk tilgang til forsinket feedback og opdager en dataafhængighed, der genereliserer tidliere afhængigheder af forsinkelserne. Vi designer en metaalgoritme, der forbigår udforholdmæssigt store forsinkelser og dermer løser dataafhængigheden. Dette påviser den formodede skalering for en vilkårlig sekvens af forsinkelser. Endeligt designer vi en ny fintuningsmetode når algoritmen har adgang til ekstra information. Med denne tuning opnår vi en langt bedre regret-skalering end den formodede. Vi designer eksempler på sådanne scenarier, hvor regret-skaleringen er polynomisk bedre.

Acknowledgements

My thanks go to my supervisor Yevgeny Seldin for guiding me through the process of becoming a researcher in my own right. The freedom, responsibility and ownership he has given me, though hard at times, has been indispensable for this journey. I would further like to thank Nicolò Cesa-Bianchi who kindly welcomed me in Milan as well as Giulia DeSalvo and Claudio Gentile. Finally my deepest appreciation goes to Monica and my friends and family for their unwavering distraction and support.

Contents

Chapter 1

Introduction

In online machine learning we are concerned with constructing algorithms that learn how to solve a task while performing it, in contrast to learning in a separate, offline phase. This direction of research is motivated by applications such as customer interaction or auctions where separate training data is expensive or the optimal planning of a training phase is difficult. The framework of multi-armed bandits has emerged as a central model in theoretical study of online machine learning. This framework is simple enough that provable results are possible, while it at the same time displays interesting, non-trivial dynamics. Combined with ease of extension this framework has opened up an avenue for theoretical understanding of learning in interaction with an environment. Through the use of this framework this thesis explores how a learning algorithm can exploit additional structure of a problem to perform better, and how to work with and overcome delayed feedback.

This chapter aims at introducing central ideas within machine learning and online learning and situating multi-armed bandits within the general field, thereby introducing the setting of the subsequent chapters.

1.1 Machine Learning

Machine learning has as its goal building systems capable of solving complex problems not by design but through learning how to solve them. Motivation range from automating tasks of high dimension or with huge data sets where humans traditionally perform poorly, to solving tasks that are easy for humans to solve but difficult to formalise algorithmically such as pattern recognition. A learning algorithm can further be seen as an abstract, general solution to a class of problems eliminating the need of solving these tasks individually. The potential benefits of this rapidly growing field include both relieving humans of labour-intensive tasks such as transportation and allowing us to solve previously impossible problems. For people in pursuit of general artificial intelligence, machine learning is seen as the candidate scientific approach. The approach – and name – of machine learning is inspired by the way humans acquire new skills by practising, seeking out new information and guidance, and reinforcing the understanding of the skill through feedback from applications.

In the sense that the focus is on the learning itself rather than what is learned this approach is closely connected to the field of statistics which also forms the mathematical foundations for machine learning. While machine learning can be though of as statistics it can also be seen as a more overarching model for the scientific method. In addition to empirical testing of a hypothesis, the formation of the hypothesis itself is also carried out. As similar data is being used for both processes it is necessary to carefully separate hypothesis formation and testing. This is not different from proper scientific conduct in general, but the process is compacted and potential issues exacerbated. In this scheme, where a model is learned directly from data, the subsequent empirical testing of said model is a test of how well the learner is able to *generalise* the understanding gained from seeing the data set onto new data. This is an important distinction from simply being able to describe the first data set well, which is rather a matter of flexible memorization. Proper learning requires capturing underlying behaviour and overlooking idiosyncrasies of the training data. From a statistical approach to learning this is viewed as being robust to noise and significant effort is being directed at designing techniques to this end.

While testing a learned model against a held out data set provides for an unbiased evaluation of that model's performance, this evaluation is specific to the concrete task at hand. In order to evaluate the learner and not just the learned model we need another approach. Empirically we could test learned models on a variety of data sets, which is also an approach used in practise, but ideally we would want to test the model on every data set (or every data set within some class) which is not practically feasible. The question of how well a learner would do on any hypothetical data set is a theoretical question rather than an empirical one, and this is the kind of questions that the field of learning theory is trying to answer.

Approaching machine learning theoretically affords us not only the ability to evaluate learning algorithms abstractly, it is also an avenue to understanding how these algorithms learn and understanding what learning means in general. As an example, the concept of model class complexities or dimensions captures the crucial trade-off between model capabilities and over-fitting pitfalls. While empirical evaluation is specific to the task on which the learner is evaluated, theoretical guarantees are learner specific in the sense that the analysis is carried out for a specified learner (or for a general class or heuristic of learning). In this way learning theory also guides the development of learning algorithms. In this thesis we will for instance see how the analytical approach of stability for delayed feedback guides algorithmic design.

1.2 Online Learning

Online learning is both a type of setting and algorithmic paradigm within machine learning, where data is accessed sequentially. This is relevant in modelling specific learning tasks such as trading, forecasting, medical diagnosis among others that are inherently sequential. As an algorithmic approach it is also used in traditional machine learning applications where the amount of data makes processing the entire data set at once infeasible, necessitating a sequential approach.

The online approach to learning dissolves the way data is traditionally split into training and testing sets – of course without dispensing of the methodological concerns of proper evaluation. Instead in online learning the learner's predictions are both tested and trained in an ongoing fashion on the same data. This can be achieved by first evaluating the prediction and then showing that evaluation to the learner for training. Crucially this means the learner is not only concerned with learning from the data and arriving at a good model, but doing so such that the intermediate modelling is also performant. One can think about this as caring about the performance on the training data or the learning trajectory.

This characterisation applies to a wide array of learning scenarios and scientific sub fields. Notable examples falling within this paradigm are reinforcement learning in the form of Markov decision processes [Kaelbling et al., 1996, Szepesvári, 2010], online convex optimisation [Shalev-Shwartz, 2011], and streaming formulations of supervised learning including active learning [Beygelzimer et al., 2009]. While reinforcement learning can be seen as a separate field seeing a large amount of applied research, there are overlaps with the more theoretical study of online learning focused on the frameworks of prediction with expert advice and multi-armed bandits.

1.3 Prediction with Expert Advice and Multi-armed Bandits

The model of online learning we are concerned with in this thesis is one of sequential prediction. In this model the learner must in each round make a prediction by choosing an action from a finite set. The problem of sequential prediction traces its roots back to Thompson [1933] and an accelerating growth has been seen in recent decades with the advent of notable algorithms such as Hedge [Vovk, 1990, Littlestone and Warmuth, 1994], UCB [Auer et al., 2002a] and Exp3 [Auer et al., 2002b]. Differing only in the feedback model, Prediction with expert advice and multi-armed bandits serve as two related frameworks for such problems. The simple model underlying both abstracts a wide array of sequential decision processes and has proven to be easily extendable to allow studying more specific scenarios. As the difference between the two settings falls within the myriad of variations hereof we consider them to be variations of the same underlying setting.

A general version of the learning setting is as follows: The game is played in rounds. In round t the learner first picks an action $A_t \in \{1, \dots, K\}$ and suffers the loss of that action $\ell_t^{A_t} \in [0, 1]$. Subsequently the learner observes some feedback. In the bandit setting this is the loss of the chosen action $\ell_t^{A_t}$, while losses ℓ_t^a of other actions $a \neq A_t$ remain hidden. In prediction with expert advice, also know as *full information*, losses for all actions are revealed.

Various models for how the losses are generated can be considered, but in general we would like to be able to work with any sequence of losses or any sequence within some generation scheme. This means that simply measuring the cumulative performance of the learner is not generally informative. Instead the learner is compared to a benchmark in order to normalize the performance with respect to the specific instance of losses. The canonical performance measure is the *regret* of the chosen action sequence compared to the best static strategy in hindsight:

Definition 1. *The expected regret or pseudo regret*[1] *of a learner at time T is*

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_{a\in\{1,\ldots,K\}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{a}\right],$$

where the expectation is with respect to the randomness of the learner and any randomness of the loss generation.

The regret definition is a central element to the dynamic of online learning where the learner is evaluated cumulatively and not just by the final model's performance. For bandits we further have a lack of counterfactual information as the learner only gets to know the loss of the chosen action.

The combination of measuring cumulative performance and not having access to counterfactual information leads to the central optimisation problem in online learning, the exploitation/exploration-trade off: As the learner desires a small cumulative loss there is in every round a desire to pick a good action (that is of small loss). In order to do so however, the learner needs to estimate the performance of every action for comparison. Due to the lack of counterfactual information, this requires playing every action some amount – even the bad ones. While displaying this non-trivial dynamic of trading off exploitation and exploration, theoretical analysis of multi-armed bandits is also tractable allowing for provable regret guarantees.

The above underlying model has proven to be a highly flexible framework that allows for theoretical understanding of a wide array of variations which the contributions of this thesis are examples of. From this point of view, multi-armed bandits can be seen as a framework that allows for isolation of interesting dynamics and providing an understanding of the impact of such dynamics, which can then guide the understanding of more complicated systems, where such an analysis might not be tractable.

Variations of multi-armed bandits occur in many different axes of the game. For instance different loss generation scheme such as stochastic losses [Bubeck and Cesa-Bianchi, 2012], corrupted schemes [Lykouris et al., 2018], bounded variations [Bubeck et al., 2019] and small effective loss ranges, as well as combined or intermediate regimes. The central difference between the full information and bandit setting is also an axis of variation such as prediction with limited advice where the loss of more than one, but not all, actions can be observed [Seldin et al., 2014]. Graph based feedback is another

[1] Note that the term *expected regret* would sometimes have the minimum taken under the expectation and the above expression would then be termed the *pseudo regret*. In this terminology the pseudo regret is typically the quantity of interest, even though the intuitive meaning is less clear. In the subsequent chapters of this thesis, the main focus will be on losses generated by an *oblivious* adversary, that is deterministically and prior to the game. In that case the second expectation can be omitted and the expected regret and pseudo regret coincides.

distinct interpolation between the two schemes [Alon et al., 2017]. The feedback can further be subject of delays or be composite [Mesterharm, 2005, Cesa-Bianchi et al., 2019, 2018]. Variations akin to active learning include both integrated schemes such as costly feedback [Seldin et al., 2014] and constrained settings such as label efficient prediction [Cesa-Bianchi et al., 2005]. A setting of much recent attention and wider applicability is contextual bandits, where the environment produces some side-information prior to the action being chosen [Langford and Zhang, 2007, Beygelzimer et al., 2011]. Other notable variations include combinatorial semi-bandits [Kveton et al., 2015, Zimmert et al., 2019], recharging bandits [Kleinberg and Immorlica, 2018], bandits with abstention [Cortes et al., 2017], bandits with switching cost [Dekel et al., 2014] and dueling bandits [Yue and Joachims, 2009] among others.

1.4 Outline and contributions

The following two chapters contain the enclosed papers Thune and Seldin [2018a] and Thune et al. [2019].

Chapter 2 corresponding to Thune and Seldin [2018a] consider exploiting additional structure in the loss generation under the framework of prediction with limited advice [Seldin et al., 2014]. This framework is an interpolation between bandit feedback and full feedback, allowing the learner to observe some variable number of arms in addition to the one played. We consider the loss structure of a small effective loss range, which a learner surprisingly cannot adapt to under bandit feedback [Gerchinovitz and Lattimore, 2016, Cesa-Bianchi and Shamir, 2018]. We design an algorithm that does so with one additional point of feedback and we further show that this algorithm also exploit stochastic loss generation simultaneously. This is done adaptively without requiring prior knowledge of the learning setting.

Chapter 3 corresponding to Thune, Cesa-Bianchi, and Seldin [2019] considers the problem of multi-armed bandits with delayed feedback. The approaches of Cesa-Bianchi et al. [2019] for adversarial losses and fixed delays are extended to the case of arbitrary delays. We uncover a generalized dependency on the delay sequence and design a meta-algorithm to circumvent this, by which a previously conjectured regret bound is proven for any delay sequence. For tuning the algorithm adaptively, we consider a case where the learner has access to the present delay immediately prior to making a prediction. For this tuning scheme we prove a regret bound with a much improved regret scaling in the delay sequence and design problem instances where this is the case.

1.4.1 Contributions

1. In prediction with limited advice we show that one additional point of feedback is sufficient for getting an expected regret scaling linear in the effective loss range ε. In contrast this is not generally possible with just bandit feedback.

2. We design an algorithm that is fully adaptive in the effective loss range, achieving an expected regret bound of $O\big(\varepsilon\sqrt{TK\ln K}\big)$, without requiring knowledge of either ε or T.

3. We provide a lower bound on the expected regret of order $O\big(\varepsilon\sqrt{TK}\big)$ for prediction with 2 points feedback. This shows that the expected regret upper bound of our algorithm is tight up to logarithmic and lower order factors.

4. We show that our algorithm adapts to stochastic loss generation achieving a regret bound constant in T while maintaining the linear scaling with the effective loss range. The algorithm does not need to know the setting for tuning.

5. We extend the analysis of Exp3 with delayed feedback to the case of arbitrary delays. For bounded delays we prove the conjecture of Cesa-Bianchi et al. [2019] that the expected regret scales as $O\big(\sqrt{(KT+D)\ln K}\big)$, where $D = \sum_{t=1}^{T} d_t$ is the sum of the delay sequence.

6. We extend the above result by introducing the Skipper wrapper algorithm, which maintains the conjectured regret bound for any delay sequence.

7. We design a doubling scheme for a slightly easier setting, where the learner has access to the delay d_t at the start of round t. With this the learner can maintain the conjectured regret bound without requiring knowledge of D or T.

8. In this new setting we show that it is possible to get a much improved regret upper bound of order $\min_\beta \big(|S_\beta| + \beta \ln K + (KT + D_\beta)/\beta\big)$, where $|S_\beta|$ is the number of observations with delay exceeding β, and D_β is the total delay of observations with delay below β. This relaxes to the previous bound but can be much better. We construct delay sequences where the improved regret bound is of order $O\big(T^{1/2}\big)$ while the previous, conjectured bound is of order $O\big(T^{3/4}\big)$.

Chapter 2

Adaptation to easy data in prediction with limited advice

This chapter is based on the following paper:

> Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems*, pages 2909–2918, 2018a

The text is presented here is an updated version corresponding to Thune and Seldin [2018b] as found on arXiv. The changes consist of a corrected scaling in the number of arms K in Theorem 3 and the proof thereof.

2.1 Abstract

We derive an online learning algorithm with improved regret guarantees for "easy" loss sequences. We consider two types of "easiness": (a) stochastic loss sequences and (b) adversarial loss sequences with small effective range of the losses. While a number of algorithms have been proposed for exploiting small effective range in the full information setting, Gerchinovitz and Lattimore [2016] have shown the impossibility of regret scaling with the effective range of the losses in the bandit setting. We show that just one additional observation per round is sufficient to circumvent the impossibility result. The proposed *Second Order Difference Adjustments* (SODA) algorithm requires no prior knowledge of the effective range of the losses, ε, and achieves an $O(\varepsilon\sqrt{KT\ln K}) + \tilde{O}(\varepsilon K\sqrt[4]{T})$ expected regret guarantee, where T is the time horizon and K is the number of actions. The scaling with the effective loss range is achieved under significantly weaker assumptions than those made by Cesa-Bianchi and Shamir [2018] in an earlier attempt to circumvent the impossibility result. We also provide a regret lower bound of $\Omega(\varepsilon\sqrt{TK})$, which almost matches the upper bound. In addition, we show that in the stochastic setting SODA achieves an $O\left(\sum_{a:\Delta_a>0}\frac{K^3\varepsilon^2}{\Delta_a}\right)$ pseudo-regret bound that holds simultaneously with the adversarial regret guarantee. In other words, SODA is safe against an unrestricted oblivious adversary and provides improved regret guarantees for at least two different types of "easiness" simultaneously.

2.2 Introduction

Online learning algorithms with both worst-case regret guarantees and refined guarantees for "easy" loss sequences have come into research focus in recent years. In our work we consider *prediction with limited advice* games [Seldin et al., 2014], which are an interpolation between *full information* games [Vovk, 1990, Littlestone and Warmuth, 1994, Cesa-Bianchi and Lugosi, 2006] and games with *limited* (a.k.a. *bandit*) *feedback* [Auer et al., 2002b, Bubeck and Cesa-Bianchi, 2012].[1] In prediction with limited advice the learner faces K unobserved sequences of losses $\{\ell_t^a\}_{t,a}$, where a indexes the sequence number and t indexes the elements within the a-th sequence. At each round t of the game the learner picks a sequence $A_t \in \{1, \dots, K\}$ and suffers the loss $\ell_t^{A_t}$, which is then observed. After that, the learner is allowed to observe the losses of M additional sequences in the same round t, where $0 \le M \le K - 1$. For $M = K - 1$ the setting is equivalent to a full information game and for $M = 0$ it becomes a bandit game.

For a practical motivation behind prediction with limited advice imagine that the loss sequences correspond to losses of K different algorithms for solving some problem, or K different parametrizations of one algorithm, or K different experts. If we had the opportunity we would have executed all the algorithms or queried all the experts before making a prediction. This would correspond to a full information game. But in reality we may be constrained by time, computational power, or monetary budget. In such case we are forced to select algorithms or experts to query. Being able to query just one expert or algorithm per prediction round corresponds to a bandit game, but we may have time or money to get a bit more, even though not all of it. This is the setting modeled by prediction with limited advice.

Our goal is to derive an algorithm for prediction with limited advice that is robust in the worst case and provides improved regret guarantees in "easy" cases. There are multiple ways to define "easiness" of loss sequences. Among them, loss sequences generated by i.i.d. sources, like the classical stochastic bandit model [Robbins, 1952, Lai and Robbins, 1985, Auer et al., 2002a], and adversarial sequences with bounded effective range of the losses within each round [Cesa-Bianchi et al., 2007]. For the former a simple calculation shows that in the full information setting the basic Hedge algorithm [Vovk, 1990, Littlestone and Warmuth, 1994] achieves an improved "constant" (independent of time horizon) pseudo-regret guarantee without sacrificing the worst-case guarantee. Much more work is required to achieve adaptation to this form of easiness in the bandit setting if we want to keep the adversarial regret guarantee simultaneously [Bubeck and Slivkins, 2012, Seldin and Slivkins, 2014, Auer and Chiang, 2016, Seldin and Lugosi, 2017, Wei and Luo, 2018, Zimmert and Seldin, 2018].

An algorithm that adapts to the second form of easiness in the full information setting was first proposed by Cesa-Bianchi et al. [2007] and a number of variations have followed [Gaillard et al., 2014, Koolen and van Erven, 2015, Luo and Schapire, 2015, Wintenberger, 2017]. However, a recent result by Gerchinovitz and Lattimore [2016] have shown that such adaptation is impossible in the bandit setting. Cesa-Bianchi and Shamir [2018] proposed a way to circumvent the impossibility result by either

[1] There exists an orthogonal interpolation between full information and bandit games through the use of feedback graphs Alon et al. [2017], which is different and incomparable with prediction with limited advice, see Seldin et al. [2014] for a discussion.

assuming that the ranges of the individual losses are provided to the algorithm in advance or assuming that the losses are smooth and an “anchor” loss of one additional arm is provided to the algorithm. The latter assumption has so far only lead to a substantial improvement when the “anchor” loss is always the smallest loss in the corresponding round.

We consider adaptation to both types of easiness in prediction with limited advice. We show that $M = 1$ (just one additional observation per round) is sufficient to circumvent the impossibility result of Gerchinovitz and Lattimore [2016]. This assumption is weaker than the assumptions in Cesa-Bianchi and Shamir [2018]. We propose an algorithm, which achieves improved regret guarantees both when the effective loss range is small and when the losses are stochastic (generated i.i.d.). The algorithm is inspired by the BOA algorithm of Wintenberger [2017], but instead of working with exponential weights of the cumulative losses and their second moment corrections it uses estimates of the loss differences. The algorithm achieves an $O(\varepsilon\sqrt{KT\ln K}) + \tilde{O}(\varepsilon K\sqrt[4]{T})$ expected regret guarantee with no prior knowledge of the effective loss range ε or time horizon T. We also provide regret lower bound of $\Omega(\varepsilon\sqrt{KT})$, which matches the upper bound up to logarithmic terms and smaller order factors. Furthermore, we show that in the stochastic setting the algorithm achieves an $O\left(\sum_{a:\Delta_a>0}\frac{K^3\varepsilon^2}{\Delta_a}\right)$ pseudo-regret guarantee. The improvement in the stochastic setting is achieved without compromising the adversarial regret guarantee.

The paper is structured in the following way. In Section 2.3 we lay out the problem setting. In Section 2.4 we present the algorithm and in Section 2.5 the main results about the algorithm. Proofs of the main results are presented in Section 2.6.

2.3 Problem Setting

We consider sequential games defined by K infinite sequences of losses $\{\ell_1^a, \ell_2^a, \dots\}_a$ for $a \in \{1, \dots, K\}$, where $\ell_t^a \in [0, 1]$ for all a and t. At each round $t \in \{1, 2, \dots\}$ of the game the learner selects an action (a.k.a. “arm”) $A_t \in [K] := \{1, \dots, K\}$ and then suffers and observes the corresponding loss $\ell_t^{A_t}$. Additionally, the learner is allowed to choose a second arm, B_t, and observe $\ell_t^{B_t}$. The loss of the second arm, $\ell_t^{B_t}$, is not suffered by the learner. (This is analogous to the full information setting, where the losses of all arms $a \neq A_t$ are observed, but not suffered). It is assumed that $\ell_t^{B_t}$ is observed *after* A_t has been selected, but other relative timing of events within a round is unimportant for our analysis.

The performance of the learner up to round T is measured by *expected regret* defined as

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{a}\right], \tag{2.1}$$

where the expectation is taken with respect to potential randomization of the loss generation process and potential randomization of the algorithm. We note that in the adversarial setting the losses are considered deterministic and the second expectation can be omitted, whereas in the stochastic setting the definition coincides with the definition of pseudo-regret [Bubeck and Cesa-Bianchi, 2012, Seldin and Lugosi, 2017]. In some literature $\mathcal{R}_T$ is termed *excess of cumulative predictive risk* [Wintenberger, 2017].

Below we define adversarial and stochastic loss generation models and effective range of loss sequences.

Adversarial losses

In the adversarial setting the loss sequences are selected arbitrarily by an adversary. We restrict ourselves to the *oblivious* model, where the losses are fixed before the start of the game and do not depend on the actions of the learner.

Stochastic losses

In the stochastic setting the losses are drawn i.i.d., so that $\mathbb{E}[\ell_t^a] = \mu_a$ independently of t. Since we have a finite number of arms, there exists a best arm $a^\star$ (not necessarily unique) such that $\mu_{a^\star} \leq \mu_a$ for all a. We further define the suboptimality *gaps* by

$$\Delta_a := \mu_a - \mu_{a^\star} \geq 0.$$

In the stochastic setting the expected regret can be rewritten as

$$\mathcal{R}_T = \sum_{a \in [K]:\Delta_a > 0} \Delta_a \, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(A_t = a)\right], \tag{2.2}$$

where $\mathbb{1}$ is the indicator function.

Effective loss range

For both the adversarial and stochastic losses, we define the *effective loss range* as the smallest number ε, such that for all $t \in [T]$ and $a, a' \in [K]$:

$$|\ell_t^a - \ell_t^{a'}| \leq \varepsilon \quad \text{almost surely.} \tag{2.3}$$

Since we have assumed that $\ell_t^a \in [0, 1]$, we have $\varepsilon \leq 1$, where $\varepsilon = 1$ corresponds to an unrestricted setting.

2.4 Algorithm

We introduce the *Second Order Difference Adjustments* (*SODA*) algorithm, summarized in Algorithm 1. SODA belongs to the general class of *exponential weights* algorithms. The algorithm has two important distinctions from the common members of this class. First, it uses cumulative *loss difference estimators* instead of cumulative loss estimators for the exponential weights updates. Instantaneous loss difference estimators at round t are defined by

$$\widetilde{\Delta\ell}_t^a = (K-1)\mathbb{1}(B_t = a)\left(\ell_t^{B_t} - \ell_t^{A_t}\right). \tag{2.4}$$

SODA samples the "secondary" action B_t (the additional observation) uniformly from $K-1$ arms, all except A_t, and the $(K-1)$ term above corresponds to importance weighting with respect to the sampling of B_t. The loss difference estimators scale with the effective range of the losses and they can be positive and negative. Both of these properties are distinct from the traditional loss estimators. The second difference is that we are using a second order adjustment in the weighting inspired by Wintenberger [2017]. We define the cumulative loss difference estimator and its second moment by

$$D_t(a) := \sum_{s=1}^{t} \widetilde{\Delta\ell}_s^a, \quad S_t(a) := \sum_{s=1}^{t} \left(\widetilde{\Delta\ell}_s^a\right)^2. \tag{2.5}$$

We then have the distribution $\boldsymbol{p_t}$ for selecting the primary action A_t defined by

$$p_t^a = \frac{\exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)}{\sum_{a=1}^{K} \exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)}, \tag{2.6}$$

where η_t is a learning rate scheme, defined as

$$\eta_t = \min\left\{\sqrt{\frac{\ln K}{\max_a S_{t-1}(a) + (K-1)^2}}, \frac{1}{2(K-1)}\right\}. \tag{2.7}$$

The learning rate satisfies $\eta_t \leq 1/(2\varepsilon(K-1))$ for all t, which is required for the subsequent analysis.

The algorithm is summarized below:

```
Algorithm 1: Second Order Difference Adjustments (SODA)
Initialize $\boldsymbol{p_1} \leftarrow (1/K, \ldots, 1/K)$.
for $t = 1, 2, \ldots$ do
    Draw $A_t$ according to $\boldsymbol{p_t}$;
    Draw $B_t$ uniformly at random from the remaining actions $[K] \setminus \{A_t\}$;
    Observe $\ell_t^{A_t}, \ell_t^{B_t}$ and suffer $\ell_t^{A_t}$;
    Construct $\widetilde{\Delta\ell}_t^a$ by equation (2.4);
    Update $D_t(a), S_t(a)$ by (2.5);
    Define $\boldsymbol{p_{t+1}}$ by (2.6);
end
```

2.5 Main Results

We are now ready to present the regret bounds for SODA. We start with regret upper and lower bounds in the adversarial regime and then show that the algorithm simultaneously achieves improved regret guarantee in the stochastic regime.

2.5.1 Regret Upper Bound in the Adversarial Regime

First we provide an upper bound for the expected regret of SODA against oblivious adversaries that produce loss sequences with effective loss range bounded by ε. Note that this result does not depend on prior knowledge of the effective loss range ε or time horizon T.

Theorem 1. *The expected regret of* SODA *against an oblivious adversary satisfies*

$$\mathcal{R}_T \leq 4\varepsilon\sqrt{(K-1)\ln K}\sqrt{T + (K-1)\sqrt{T}\left(2 + \sqrt{\ln\left(\sqrt{T}(K-1)\right)/2}\right)} \\ + 4(K-1)\ln K.$$

A proof of this theorem is provided in Section 2.6.1.[2] The upper bound scales as $O(\varepsilon\sqrt{KT\ln K}) + \tilde{O}(\varepsilon K\sqrt[4]{T})$, which nearly matches the lower bound provided below.

2.5.2 Regret Lower Bound in the Adversarial Regime

We show that in the worst case the regret must scale linearly with the effective loss range ε.

Theorem 2. *In prediction with limited advice with $M = 1$ (one additional observation per round or, equivalently, two observations per round in total), for loss sequences with effective loss range ε, we have for $T \geq 3K/32$:*

$$\inf\sup\mathcal{R}_T \geq 0.02\varepsilon\sqrt{KT},$$

where the infimum is with respect to the choices of the algorithm and the supremum is over all oblivious loss sequences with effective loss range bounded by ε.

The theorem is proven by adaptation of the $\Omega(\sqrt{KT})$ lower bound by Seldin et al. [2014] for prediction with limited advice with unrestricted losses in $[0, 1]$ and one extra observation. We provide it in Appendix 2.8. Note that the upper bound in Theorem 1 matches the lower bound up to logarithmic terms and lower order additive factors. In particular, changing the selection strategy for the second arm, B_t, from uniform to anything more sophisticated is not expected to yield significant benefits in the adversarial regime.

2.5.3 Regret Upper Bound in the Stochastic Regime

Finally, we show that SODA enjoys constant expected regret in the stochastic regime. This is achieved without sacrificing the adversarial regret guarantee.

[2] It is straightforward to extended the analysis to time-varying ranges, $\varepsilon_t : |\ell_t^a - \ell_t^{a'}| \leq \varepsilon_t$ for all a, a' a.s., which leads to an $O\left(\sqrt{\sum_{t=1}^T(\varepsilon_t^2)K\ln K}\right) + \tilde{O}\left(K\sqrt[4]{\sum_{t=1}^T\varepsilon_t^2}\right)$ regret bound . For the sake of clarity we restrict the presentation to a constant ε.

Theorem 3. *The expected regret of* SODA *applied to stochastic loss sequences with gaps Δ_a satisfies*

$$\mathcal{R}_T \leq \sum_{a:\Delta_a>0} \left[\left(\frac{16K^3}{\ln K} + 16K^2 \right) \frac{\varepsilon^2}{\Delta_a} + 4K^2 + \frac{\Delta_a}{K} \right]. \tag{2.8}$$

A brief sketch of a proof of this theorem is given in Section 2.6.2, with the complete proof provided in Appendix 2.10.

Note that ε is the effective range of realizations of the losses, whereas the gaps Δ_a are based on the expected losses. Naturally, $\Delta_a \leq \varepsilon$. For example, if the losses are Bernoulli then the range is $\varepsilon = 1$, but the gaps are based on the distances between the biases of the Bernoulli variables. When the losses are not $\{0, 1\}$, but confined to a smaller range ε, Theorem 3 yields a tighter regret bound. The scaling of the regret bound in K is suboptimal and it is currently unknown whether it could be improved without compromising the worst-case guarantee. Perhaps changing the selection strategy for B_t could help here. We leave this improvement for future work.

To summarize, SODA achieves adversarial regret guarantee that scales with the effective loss range and almost matches the lower bound and simultaneously has improved regret guarantee in the stochastic regime.

2.6 Proofs

This section contains the proof of Theorem 1 and a proof sketch for Theorem 3. The proof of Theorem 2 is provided in Appendix 2.8.

2.6.1 Proof of Theorem 1

The proof of the theorem is prefaced by two lemmas, but first we show some properties of the loss difference estimators. We use $\mathbb{E}_{B_t}$ to denote expectation with respect to selection of B_t conditioned on all random outcomes prior to this selection. For oblivious adversaries, the expected cumulative loss difference estimators are equal to the negative expect regret against the corresponding arm a:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\Delta\ell}_t^a\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathop{\mathbb{E}}_{B_t}\left[\widetilde{\Delta\ell}_t^a\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \left(\ell_t^a - \ell_t^{A_t}\right)\right] \\
&= \sum_{t=1}^{T} \ell_t^a - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] \\
&=: -\mathcal{R}_T^a,
\end{aligned}$$

where we have used the fact that $\widetilde{\Delta\ell}_t^a$ is an unbiased estimate of $\ell_t^a - \ell_t^{A_t}$ due to importance weighting with respect to the choice of B_t. Similarly, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\widetilde{\Delta\ell}_t^a\right)^2\right] = (K-1)\ \mathbb{E}\left[\sum_{t=1}^{T}\left(\ell_t^a - \ell_t^{A_t}\right)^2\right]. \tag{2.9}$$

Similar to the analysis of the anytime version of EXP3 in Bubeck and Cesa-Bianchi [2012], which builds on Auer et al. [2002b], we consider upper and lower bounds on the expectation of the incremental update. This is captured by the following lemma:

Lemma 4. *With a learning rate scheme η_t for $t = 1, 2, \ldots$, where $\eta_t \leq 1/2\varepsilon(K-1)$,* SODA *fulfills:*

$$-\sum_{t=1}^{T}\widetilde{\Delta\ell}_t^a \leq \frac{\ln K}{\eta_T} + \eta_T\sum_{t=1}^{T}\left(\widetilde{\Delta\ell}_t^a\right)^2 - \sum_{t=1}^{T}\mathop{\mathbb{E}}_{a\sim p_t}\left[\widetilde{\Delta\ell}_t^a\right] + \sum_{t}\left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right) \tag{2.10}$$

for all a, where we define the potential

$$\Phi_t(\eta) := \frac{1}{\eta}\ln\left(\frac{1}{K}\sum_{a=1}^{K}\exp\left(-\eta D_t(a) - \eta^2 S_t(a)\right)\right). \tag{2.11}$$

Note that unlike in the analysis of EXP3, here the learning rates η_t do not have to be non-increasing. A proof of this lemma is based on modification of standard arguments and is found in Appendix 2.9.1.

The second lemma is a technical one and is proven in Appendix 2.9.2.

Lemma 5. *Let σ_t with $t \in \mathbb{N}$ be an increasing positive sequence with bounded differences such that $\sigma_t - \sigma_{t-1} \leq c$ for a finite constant c. Let further $\sigma_0 = 0$. Then*

$$\sum_{t=1}^{T}\sigma_t\left(\frac{1}{\sqrt{\sigma_{t-1}+c}} - \frac{1}{\sqrt{\sigma_t + c}}\right) \leq 2\sqrt{\sigma_{T-1}+c}.$$

Proof of Theorem 1 We apply Lemma 4, which leads to the following inequality for any learning rate scheme η_t for $t = 1, 2, \ldots$, where $\eta_t \leq 1/2\varepsilon(K-1)$:

$$-\sum_{t=1}^{T}\widetilde{\Delta\ell}_t^a \leq \underbrace{\frac{\ln K}{\eta_T}}_{1^{\text{st}}} + \underbrace{\eta_T\sum_{t=1}^{T}\left(\widetilde{\Delta\ell}_t^a\right)^2}_{2^{\text{nd}}} - \underbrace{\sum_{t=1}^{T}\mathop{\mathbb{E}}_{a\sim p_t}\left[\widetilde{\Delta\ell}_t^a\right]}_{3^{\text{rd}}} + \underbrace{\sum_{t=1}^{T}\left(\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t)\right)}_{4^{\text{th}}}. \tag{2.12}$$

Note that in expectation, the left hand side of (2.12) is the regret against arm a. We are thus interested in bounding the expectation of the terms on the right hand side, where we note that the third term vanishes in expectation. We first consider the case where $\eta_t = \sqrt{\ln K/(\max_a S_t(a) + (K-1)^2)}$, postponing the initial value for now.

The first term becomes:

$$\frac{\ln K}{\eta_T} = \sqrt{\ln K}\sqrt{\max_a S_{T-1}(a) + (K-1)^2}. \tag{2.13}$$

The second term becomes:

$$\eta_T S_T(a) = \sqrt{\ln K}\frac{S_T(a)}{\sqrt{\max_a S_{T-1}(a) + (K-1)^2}} \leq \sqrt{\ln K}\sqrt{\max_a S_{T-1}(a) + (K-1)^2}, \tag{2.14}$$

where we use that $S_t(a) \leq S_{t-1}(a) + (K-1)^2$ for all t by design.

Finally, for the fourth term in equation (2.12), we need to consider the potential differences. Unlike in the anytime analysis of EXP3, where this term is negative [Bubeck and Cesa-Bianchi, 2012], in our case it turns to be related to the second moment of the loss difference estimators. We let

$$q_t^\eta = \frac{\exp\big(-\eta D_t(a) - \eta^2 S_t(a)\big)}{\sum_{a=1}^K \exp\left(-\eta D_t(a) - \eta^2 S_t(a)\right)} \tag{2.15}$$

denote the exponential update using the loss estimators up to t, but with a free learning rate η. We further suppress some indices for readability, such that $D_a = D_t(a)$ and $S_a = S_t(a)$ in the following. We have

$$\Phi_t'(\eta) = -\frac{1}{\eta^2}\ln\left(\frac{1}{K}\sum_a e^{-\eta D_a - \eta^2 S_a}\right) + \frac{1}{\eta}\frac{\sum_a e^{-\eta D_a - \eta^2 S_a}\cdot(-D_a - 2\eta S_a)}{\sum_a \exp\left(-\eta D_a - \eta^2 S_a\right)}$$

$$= \frac{\sum_a \left(e^{-\eta D_a - \eta^2 S_a}\cdot\left(-\eta D_a - 2\eta^2 S_a - \ln\left(\frac{1}{K}\sum_a e^{-\eta D_a - \eta^2 S_a}\right)\right)\right)}{\eta^2 \sum_a \exp\left(-\eta D_a - \eta^2 S_a\right)}.$$

By using $-\eta D_a - 2\eta^2 S_a = \ln\big(\exp(-\eta D_a - \eta^2 S_a)\exp(-\eta^2 S_a)\big)$ the above becomes

$$\Phi_t'(\eta) = \frac{1}{\eta^2}\mathop{\mathbb{E}}_{a\sim q_t^\eta}\left[\ln\left(\frac{q_t^\eta(a)}{1/K}\exp(-\eta^2 S_a)\right)\right] = \frac{1}{\eta^2}\mathrm{KL}\left(q_t^\eta \| \mathbf{1/K}\right) - \mathop{\mathbb{E}}_{a\sim q_t^\eta}\left[S_t(a)\right], \tag{2.16}$$

where we have used that $\mathbf{1/K}$ is the pmf. of the uniform distribution over K arms. Since the KL-divergence is always positive, we can rewrite the potential differences as

$$\Phi_t(\eta_{t+1}) - \Phi_t(\eta_t) = -\int_{\eta_{t+1}}^{\eta_t}\Phi_t'(\eta)d\eta \leq \int_{\eta_{t+1}}^{\eta_t}\mathop{\mathbb{E}}_{a\sim q_t^\eta}\left[S_t(a)\right]d\eta \leq \int_{\eta_{t+1}}^{\eta_t}\max_a S_t(a)d\eta$$

$$= \sqrt{\ln K}\max_a S_t(a)\left(\frac{1}{\sqrt{\max\limits_a S_{t-1}(a) + (K-1)^2}} - \frac{1}{\sqrt{\max\limits_a S_t(a) + (K-1)^2}}\right).$$

By Lemma 5 we then have

$$\sum_{t=1}^T \Phi_t(\eta_{t+1}) - \Phi_t(\eta_t) \leq 2\sqrt{\ln K}\sqrt{\max_a S_{T-1}(a) + (K-1)^2}. \tag{2.17}$$

Collecting the terms (2.13), (2.14) and (2.17) and noting that these bounds hold for all a, by taking expectations and using Jensen's inequality we get

$$\begin{aligned}\mathcal{R}_T &\leq \mathbb{E}\left[4\sqrt{\ln K}\sqrt{\max_a S_{T-1}(a) + (K-1)^2}\right] \\ &\leq 4\sqrt{\ln K}\sqrt{\mathbb{E}\left[\max_a S_{T-1}(a)\right] + (K-1)^2}. \end{aligned} \tag{2.18}$$

The remainder of the proof is to bound this inner expectation:

$$\mathbb{E}\left[\max_a S_{T-1}(a)\right] \leq (K-1)^2\varepsilon^2\,\mathbb{E}\left[\max_a \sum_{t=1}^{T-1} \mathbb{1}[B_t = a]\right].$$

Let $Z_t^a = \sum_{s=1}^t \mathbb{1}[B_s = a]$ and note that $Z_{T-1}^a \leq T-1$. We now consider a partioning of the probability for a cutoff $\alpha > 0$:

$$\begin{aligned}\mathbb{E}[\max_a Z_{T-1}^a] &\leq \alpha\,\mathbb{P}\left\{\max_a Z_{T-1}^a \leq \alpha\right\} + (T-1)\,\mathbb{P}\left\{\max_a Z_{T-1}^a > \alpha\right\} \\ &\leq \alpha + (T-1)K\,\mathbb{P}\left\{Z_{T-1}^a > \alpha\right\},\end{aligned}$$

using a union bound for the final inequality. To continue we need to address the fact that the B_t's are not independent. We can however note that $\mathbb{P}\{B_t = a\} \leq (K-1)^{-1}$ for all t and a. By letting x_t^a be Bernoulli with parameter $(K-1)^{-1}$ and $X_T^a = \sum_{t=1}^T x_t^a$ we then get

$$\mathbb{P}\left\{Z_{T-1}^a > \alpha\right\} \leq \mathbb{P}\left\{X_{T-1}^a > \alpha\right\}. \tag{2.19}$$

In the upper bound we can thus substitute X_{T-1}^a for Z_{T-1}^a and exploit the fact that the x_t^a's are independent by construction. Note further that $\mathbb{E}[X_{T-1}^a] = \frac{T-1}{K-1}$, so by choosing $\alpha = \frac{T-1}{K-1} + \delta$ for $\delta > 0$, we obtain by Hoeffding's inequality:

$$\begin{aligned}\mathbb{E}[\max_a Z_{T-1}^a] &\leq \frac{T-1}{K-1} + \delta + (T-1)K\,\mathbb{P}\left\{X_{T-1}^a - \frac{T-1}{K-1} > \delta\right\} \\ &\leq \frac{T-1}{K-1} + \delta + (T-1)K\exp\left(-\frac{2\delta^2}{T-1}\right).\end{aligned}$$

We now choose $\delta = \sqrt{\frac{T}{2}\ln\left(\sqrt{T}(K-1)\right)}$, which gives us

$$\mathbb{E}[\max_a Z_{T-1}^a] \leq \frac{T-1}{K-1} + \sqrt{\frac{T}{2}\ln\left(\sqrt{T}(K-1)\right)} + 2\sqrt{T}.$$

Inserting this in (2.18) gives us the desired bound.

For the case where the learning rate at T is instead given by $1/2(K-1)$ implying $4(K-1)^2\ln K \geq \max_a S_{T-1}(a) + (K-1)^2$, the first term is $\frac{\ln K}{\eta_T} = 2(K-1)\ln K$, and the second term is

$$\begin{aligned}\eta_T S_T(a) &= \frac{1}{2(K-1)} S_T(a) \\ &\le \frac{S_{T-1}(a) + (K-1)^2}{2(K-1)} \\ &\le \frac{4(K-1)^2 \ln K}{2(K-1)} \\ &\le 2(K-1) \ln K.\end{aligned}$$

Since the learning rate is constant the potential differences vanish, completing the proof. □

2.6.2 Proof sketch of Theorem 3

Here we present the key ideas used to prove Theorem 3. The complete proof is provided in Appendix 2.10.

Recall that the expected regret in the stochastic setting is given by (2.2), where $\mathbb{E}[\mathbb{1}(A_t = a)] = \mathbb{E}[p_t^a]$. Thus, we need to bound $\mathbb{E}[\sum_t p_t^a]$. The first step is to bound this as

$$\mathbb{E}\left[p_t^a\right] \le \sigma + \mathbb{P}\left\{p_t^a > \sigma\right\} \le \sigma + \mathbb{P}\left\{K e^{-\eta_t \sum_{i=1}^{t-1} X_i} > \sigma\right\} \tag{2.20}$$

for a positive threshold σ, where we show that $p_t^a \le K e^{-\eta_t \sum_{i=1}^{t-1} X_i}$ for $X_i := \widetilde{\Delta\ell}_i^{a} - \widetilde{\Delta\ell}_i^{a^\star}$. This approach is motivated by the fact that $\mathbb{E}_{B_i}[\widetilde{\Delta\ell}_i^{a} - \widetilde{\Delta\ell}_i^{a^\star}] = \Delta_a$, where the expectation is with respect to selection of B_i and the loss generation, conditioned on all prior randomness.

The next step is to tune $\sigma \propto \exp(\sum \mathbb{E}_i[X_i])$, which allows us to bound the second term using Azuma's inequality and balance the two terms. Finally, this bound is summed over t using a technical lemma for the limit of this sum.

2.7 Discussion

We have presented the SODA algorithm for prediction with limited advice with two observations per round (the "primary" observation of the loss of the action that was played and one additional observation). We have shown that the algorithm adapts to two types of simplicity of loss sequences simultaneously: (a) it provides improved regret guarantees for adversarial sequences with bounded effective range of the losses and (b) for stochastic loss sequences. In both cases the regret scales linearly with the effective range and the knowledge of the range is not required. In the adversarial case we achieve $O(\varepsilon\sqrt{KT \ln K}) + \tilde{O}(\varepsilon K \sqrt[4]{T})$ regret guarantee and in the stochastic case we achieve $O\left(\sum_{a:\Delta_a > 0} \frac{K^3 \varepsilon^2}{\Delta_a}\right)$ regret guarantee. Our result demonstrates that just one extra observation per round is sufficient to circumvent the impossibility result of Gerchinovitz and Lattimore [2016] and significantly relaxes the assumptions made by Cesa-Bianchi and Shamir [2018] to achieve the same goal.

There are a number of open questions and interesting directions for future research. One is to improve the regret guarantee in the stochastic regime. Another is to extend the results to bandits with limited advice in the spirit of Seldin et al. [2013], Kale [2014].

Acknowledgements

The authors thank Julian Zimmert for valuable input and discussion during this project. We further thank Chloè Rouyer for pointing out a mistake in the Proof of Theorem 3 which has been fixed in the present version.

Supplementary material for Chapter 2

2.8 Proof of Theorem 2

The lower bound is a straightforward adaptation of Theorem 2 in Seldin et al. [2014], which states that for prediction with limited advice where $M' = M + 1$ of K experts are queried, we have for $T \geq \frac{3}{16}\frac{K}{M'}$:

$$\inf \sup \mathcal{R}_T \geq 0.03\sqrt{\frac{K}{M'T}},$$

where the infimum is over learning strategies and the supremum over oblivious adversaries.

Our case of $M = 1$ additional expert corresponds to $M' = 2$. The proof of the above is based upon the standard technique for lower bounding, where Bernoulli losses with varying biases are constructed. As this is a stochastic setting, the regret of playing a suboptimal arm a is analysed as

$$(\nu_a - \nu_{a^\star})\, \mathbb{E}[N_T(a)],$$

where the ν's are the biases of the Bernoulli variables and $N_T(a)$ is the number of times an arm is played. The rest of analysis consists of lower bounding the expected number of plays and tuning the biases.

By changing the constructed losses to Bernoulli variables times ε (i.e. taking values in $\{0, \varepsilon\}$), the expected values become $\varepsilon\nu_a$, which means we get a factor of ε in the above expression. Since the bound on $\mathbb{E}[N_T(a)]$ does not depend on the values taken by the distributions, but only the ability to discern them, the proof follows directly from that in Seldin et al. [2014]. □

2.9 Supplement for the proof of Theorem 1 (Section 2.6.1)

2.9.1 Proof of Lemma 4

We first derive two inequalities, which are combined and rearranged into the statement of the lemma.

Consider the quantity

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\eta_t}\ln\mathop{\mathbb{E}}_{a\sim p_t}\left[\exp\left(-\eta_t\widetilde{\Delta\ell}_t^a-\eta_t^2\left(\widetilde{\Delta\ell}_t^a\right)^2\right)\right]&\leq\sum_{t=1}^{T}\frac{1}{\eta_t}\ln\mathop{\mathbb{E}}_{a\sim p_t}\left[1-\eta_t\widetilde{\Delta\ell}_t^a\right]\\&=\sum_{t=1}^{T}\frac{1}{\eta_t}\ln\left(1-\eta_t\mathop{\mathbb{E}}_{a\sim p_t}\left[\widetilde{\Delta\ell}_t^a\right]\right)\\&\leq-\sum_{t=1}^{T}\mathop{\mathbb{E}}_{a\sim p_t}\left[\widetilde{\Delta\ell}_t^a\right],\end{aligned}$$

where the first step is based on the inequality $e^{z-z^2}\leq 1+z$ for $z=-\eta_t\widetilde{\Delta\ell}_t^a\geq-1/2$ [Cesa-Bianchi et al., 2007]. The upper bound on $\eta_t\leq(2\varepsilon(K-1))^{-1}$ guarantees that the condition of the inequality holds. The last step is based on $\ln(1+z)\leq z$ for $z>-1$.

Using the potential (2.11) we can rewrite the same quantity as

$$\begin{aligned}&\frac{1}{\eta_t}\ln\mathop{\mathbb{E}}_{a\sim p_t}\left[\exp\left(-\eta_t\widetilde{\Delta\ell}_t^a-\eta_t^2\left(\widetilde{\Delta\ell}_t^a\right)^2\right)\right]\\&\qquad=\frac{1}{\eta_t}\ln\sum_{a=1}^{K}\exp\left(-\eta_t\widetilde{\Delta\ell}_t^a-\eta_t^2\left(\widetilde{\Delta\ell}_t^a\right)^2\right)\cdot p_t^a\\&\qquad=\frac{1}{\eta_t}\ln\frac{\sum_{a=1}^{K}\exp\left(-\eta_t D_t(a)-\eta_t^2S_t(a)\right)}{\sum_{a=1}^{K}\exp\left(-\eta_t D_{t-1}(a)-\eta_t^2S_{t-1}(a)\right)}\\&\qquad=\Phi_t(\eta_t)-\Phi_{t-1}(\eta_t).\end{aligned}$$

Summing over t and reindexing the sum we get

$$\sum_{t=1}^{T}\left(\Phi_t(\eta_t)-\Phi_{t-1}(\eta_t)\right)=\sum_{t=1}^{T-1}\left(\Phi_t(\eta_t)-\Phi_t(\eta_{t+1})\right)+\Phi_T(\eta_T)-\Phi_0(\eta_1).$$

Since by definition $D_0=0$ and $S_0=0$, we have $\Phi_0(\eta_1)=0$. Next, we lower bound the middle term:

$$\begin{aligned}\Phi_T(\eta_T)&=\frac{1}{\eta_T}\ln\left(\frac{1}{K}\sum_{a=1}^{K}\exp\left(-\eta_T D_T(a)-\eta_T^2S_T(a)\right)\right)\\&\geq-\frac{\ln K}{\eta_T}+\frac{1}{\eta_T}\ln\left(\exp\left(-\eta_T D_T(a)-\eta_T^2S_T(a)\right)\right)\\&=-\frac{\ln K}{\eta_T}-D_T(a)-\eta_T S_T(a),\end{aligned}$$

where we have used that the logarithm is monotonously increasing and all the terms in the inner sum are positive.

By using the lower and upper bounds simultaneously and moving everything except for $-D_T(a)$ from the left hand side, the proof is complete. □

2.9.2 Proof of Lemma 5

By the boundedness we have:

$$\begin{aligned}\sum_{t=1}^{T}\sigma_t\left(\frac{1}{\sqrt{\sigma_{t-1}+c}}-\frac{1}{\sqrt{\sigma_t+c}}\right) &\leq \sum_{t=1}^{T}(\sigma_{t-1}+c)\left(\frac{1}{\sqrt{\sigma_{t-1}+c}}-\frac{1}{\sqrt{\sigma_t+c}}\right)\\ &= \sum_{t=0}^{T-1}\frac{\sigma_t+c}{\sqrt{\sigma_t+c}}-\sum_{t=1}^{T}\frac{\sigma_{t-1}+c}{\sqrt{\sigma_t+c}}\\ &= \sum_{t=1}^{T-1}\frac{\sigma_t-\sigma_{t-1}}{\sqrt{\sigma_t+c}}+\frac{\sigma_0+c}{\sqrt{\sigma_0+c}}-\frac{\sigma_{T-1}+c}{\sqrt{\sigma_T+c}}.\end{aligned}$$

Here the second term is $\sqrt{c}$ and the third is negative and can thus be discarded in the upper bound. The first term is a lower Riemann sum of $x \mapsto 1/\sqrt{x+c}$, giving us:

$$\begin{aligned}\sum_{t=1}^{T}\sigma_t\left(\frac{1}{\sqrt{\sigma_{t-1}+c}}-\frac{1}{\sqrt{\sigma_t+c}}\right) &\leq \sqrt{c}+\int_{\sigma_1}^{\sigma_{T-1}}\frac{1}{\sqrt{x+c}}\,dx\\ &= \sqrt{c}+2\sqrt{x+c}\,\Big|_{\sigma_1}^{\sigma_{T-1}}\\ &\leq 2\sqrt{\sigma_{T-1}+c},\end{aligned}$$

where the final inequality uses $2\sqrt{\sigma_1+c} > \sqrt{c}$. □

2.10 Proof of Theorem 3

Before proving the theorem we need the following technical lemma:

Lemma 6. *For $c > 0$ we have*

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

Proof. For the first part, note that

$$\int e^{-c\sqrt{t}}dt = -\frac{2}{c}\sqrt{t}e^{-c\sqrt{t}}-\frac{2}{c^2}e^{-c\sqrt{t}},$$

which is confirmed by differentiation. Then

$$\sum_{t=1}^{\infty}e^{-c\sqrt{t}} \leq \int_0^{\infty}e^{-c\sqrt{t}}dt = -\frac{2}{c}\sqrt{t}e^{-c\sqrt{t}}-\frac{2}{c^2}e^{-c\sqrt{t}}\Big|_0^{\infty} = \frac{2}{c^2},$$

where we use that the summand is decreasing, making the series a lower Riemann sum of the intergral. For the second part we use the exact limit and that $e^x - 1 \geq x$ with the same sign for all x:

$$\sum_{t=1}^{\infty}e^{-ct} = \frac{1}{e^c-1} \leq \frac{1}{c}.$$ □

Proof of Theorem 3 Recall that the expected regret in the stochastic setting is given by

$$\mathcal{R}_T = \sum_{a:\Delta_a>0} \Delta_a \, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(A_t = a)\right],$$

where we identify $\mathbb{E}[\mathbb{1}(A_t = a)] = \mathbb{E}[p_t^a]$. Since $p_1^a = 1/K$ by definition, we need to bound

$$\mathbb{E}\left[\sum_{t=2}^{T} p_t^a\right] = \mathbb{E}\left[\sum_{t=2}^{T} \mathbb{E}[p_t^a]\right].$$

Consider first the case where the learning rate is $\eta_t = \sqrt{\frac{\ln K}{\max_a S_{t-1}(a)+(K-1)^2}}$. We bound the individual probabilities as :

$$\begin{aligned}
p_t^a &= \frac{\exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)}{\sum_{a=1}^{K} \exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)} \\
&= \frac{\exp\left(-\eta_t (D_{t-1}(a) - D_{t-1}(a^\star)) - \eta_t^2 (S_{t-1}(a) - S_{t-1}(a^\star))\right)}{\sum_{a=1}^{K} \exp\left(-\eta_t (D_{t-1}(a) - D_{t-1}(a^\star)) - \eta_t^2 (S_{t-1}(a) - S_{t-1}(a^\star))\right)} \\
&\leq \exp\left(-\eta_t (D_{t-1}(a) - D_{t-1}(a^\star)) - \eta_t^2 (S_{t-1}(a) - S_{t-1}(a^\star))\right) \\
&\leq \exp\left(-\eta_t (D_{t-1}(a) - D_{t-1}(a^\star)\right) \exp\left(\eta_t^2 S_{t-1}(a^\star)\right) \\
&\leq K \exp\left(-\eta_t \sum_{i=1}^{t-1} X_i\right), \qquad (2.21)
\end{aligned}$$

where we have defined $\sum X_i = D_{t-1}^a - D_{t-1}^{a^\star}$ and used $\eta_t^2 S_{t-1}(a^\star) \leq \ln K$.

Next we split up the expectation in two parts around a threshold $\sigma > 0$, using $p_t^a \leq 1$ and (2.21):

$$\mathbb{E}[p_t^a] \leq \sigma \, \mathbb{P}\left\{p_t^a \leq \sigma\right\} + 1 \cdot \mathbb{P}\left\{p_t^a > \sigma\right\} \leq \sigma + \mathbb{P}\left\{K \exp\left(-\eta_t \sum_{i=1}^{t-1} X_i\right) > \sigma\right\}, \qquad (2.22)$$

Since η_t is a random variable correlated with the X_i's, we cannot directly bound this expression. We can however split the event under the probability into two separate cases, and upper bound the expression using upper and lower bounds on η_t in the cases where

$\sum X_i$ is negative or positive:

$$\begin{aligned}\mathbb{P}\left\{K\exp\left(-\eta_t\sum_{i=1}^{t-1}X_i\right)>\sigma\right\} &= \mathbb{P}\left\{K\exp\left(-\eta_t\sum_{i=1}^{t-1}X_i\right)>\sigma \;\&\; \sum_{i=1}^{t-1}X_i\leq 0\right\}\\ &\quad+\mathbb{P}\left\{K\exp\left(-\eta_t\sum_{i=1}^{t-1}X_i\right)>\sigma \;\&\; \sum_{i=1}^{t-1}X_i> 0\right\}\\ &\leq \mathbb{P}\left\{K\exp\left(-\bar{\eta}_t\sum_{i=1}^{t-1}X_i\right)>\sigma\right\}\\ &\quad+\mathbb{P}\left\{K\exp\left(-\frac{\sum_{i=1}^{t-1}X_i}{2(K-1)}\right)>\sigma\right\},\end{aligned}$$

where we have introduced $\bar{\eta}_t := \sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}\frac{1}{K-1}$, which is a lower bound on η_t. Introducing $E=\sum_i \mathbb{E}_{B_i}[X_i]=(t-1)\Delta_a$ and the shorthand $V=\sum_{i=1}^{t-1}X_i - E$, we can rewrite the probabilities, resulting in

$$\mathbb{E}[p_t^a]\leq \sigma + \mathbb{P}\left\{V<-\frac{\ln(\sigma/K)}{\bar{\eta}_t}-E\right\}+\mathbb{P}\left\{V<-2(K-1)\ln(\sigma/K)-E\right\}.$$

Since V is the sum of martingale difference sequences we want to use Azuma's inequality, which requires that the right hand sides are negative. Choosing a positive splitting point σ as

$$\sigma = K\exp\left(-\frac{(t-1)\Delta_a}{2(K-1)}\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}\right), \tag{2.23}$$

the two right hand sides become

$$-\frac{\ln(\sigma/K)}{\bar{\eta}_t}-E=-\frac{E}{2}, \tag{2.24}$$

$$-2(K-1)\ln(\sigma/K)-E=E\left(\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}-1\right)\leq -\frac{E}{2}, \tag{2.25}$$

using$\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}=(K-1)\bar{\eta}_t\leq 1/2$ for the final inequality. As these are negative, we can use Azuma's inequality which since the range of the X_i's is $2(K-1)\varepsilon$ gives us

$$\mathbb{E}[p_t^a]\leq K\exp\left(-\frac{(t-1)\Delta_a}{2(K-1)}\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}\right)+2\exp\left(-\frac{E^2/4}{2(t-1)(K-1)^2\varepsilon^2}\right), \tag{2.26}$$

where the inequality comes from substitution of (2.23), (2.24) and (2.25), and the two probabilities becomes one expression using the final inequality of (2.25).

We now consider two cases of the first term in (2.26). If $(t-1)\varepsilon^2\geq 1$, then

$$\exp\left(-\frac{(t-1)\Delta_a}{2(K-1)}\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}\right)\leq \exp\left(-\frac{1}{2(K-1)}\sqrt{\frac{\ln K}{2}}\frac{\Delta_a}{\varepsilon}\sqrt{t-1}\right).$$

If instead $(t-1)\varepsilon^2 \leq 1$, then

$$\exp\left(-\frac{(t-1)\Delta_a}{2(K-1)}\sqrt{\frac{\ln K}{(t-1)\varepsilon^2+1}}\right) \leq \exp\left(-\frac{\Delta_a}{2(K-1)}\sqrt{\frac{\ln K}{2}t}\right).$$

For both cases the second term in (2.26) becomes

$$2\exp\left(-\frac{1}{8}\frac{\Delta_a^2}{\varepsilon^2}\frac{t-1}{(K-1)^2}\right).$$

For $\eta_t = \frac{1}{2(K-1)}$, we first note that $\eta_t \leq \sqrt{\frac{\ln K}{\max_a S_{t-1}(a)+(K-1)^2}}$, so the bound used for p_t^a in (2.22) still applies. Since η_t is no longer a random variable, we have

$$\mathbb{E}[p_t^a] \leq \sigma + \mathbb{P}\left\{K\exp\left(-\frac{\sum X_i}{2(K-1)}\right) > \sigma\right\}.$$

Rewriting this as before and choosing $\sigma = K\exp\left(-\frac{(t-1)\Delta_a}{4(K-1)}\right)$, we get by Azuma's inequality

$$\mathbb{E}[p_t^a] \leq K\exp\left(-\frac{(t-1)\Delta_a}{4(K-1)}\right) + \exp\left(-\frac{1}{8}\frac{\Delta_a^2}{\varepsilon^2}\frac{t-1}{K-1}\right).$$

We now have three cases of bounds on $\mathbb{E}\left[p_t^a\right]$. For each of these the analysis is completed by summing over $t = 2$ to ∞, using Lemma 6 and then summing over the arms times the gaps. For all cases, the result is smaller than the right hand side in Theorem 3. □

Chapter 3

Nonstochastic multiarmed bandits with unrestricted delays

This chapter is based on the following paper:

> Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pages 6538–6547, 2019

The paper appears as in the proceeding and is followed by the supplementary material.

3.1 Abstract

We investigate multiarmed bandits with delayed feedback, where the delays need neither be identical nor bounded. We first prove that "delayed" Exp3 achieves the $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$ regret bound conjectured by Cesa-Bianchi et al. [2019] in the case of variable, but bounded delays. Here, K is the number of actions and D is the total delay over T rounds. We then introduce a new algorithm that lifts the requirement of bounded delays by using a wrapper that skips rounds with excessively large delays. The new algorithm maintains the same regret bound, but similar to its predecessor requires prior knowledge of D and T. For this algorithm we then construct a novel doubling scheme that forgoes the prior knowledge requirement under the assumption that the delays are available at action time (rather than at loss observation time). This assumption is satisfied in a broad range of applications, including interaction with servers and service providers. The resulting oracle regret bound is of order $\min_\beta \big(|S_\beta| + \beta \ln K + (KT + D_\beta)/\beta\big)$, where $|S_\beta|$ is the number of observations with delay exceeding β, and D_β is the total delay of observations with delay below β. The bound relaxes to $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$, but we also provide examples where $D_\beta \ll D$ and the oracle bound has a polynomially better dependence on the problem parameters.

3.2 Introduction

Multiarmed bandits is an algorithmic paradigm for sequential decision making with a growing range of industrial applications, including content recommendation, computational advertising, and many more. In the multiarmed bandit framework an algorithm repeatedly takes actions (e.g., recommendation of content to a user) and observes outcomes of these actions (e.g., whether the user engaged with the content), whereas the outcome of alternative actions (e.g., alternative content that could have been recommended) remains unobserved. In many real-life situations the algorithm experience delay between execution of an action and observation of its outcome. Within the delay period the algorithm may be forced to make a series of other actions (e.g., interact with new users) before observing the outcomes of all the previous actions. This setup falls outside of the classical multiarmed bandit paradigm, where observations happen instantaneously after the actions, and motivates the study of bandit algorithms that are provably robust in the presence of delays.

We focus on the nonstochastic (a.k.a. oblivious adversarial) bandit setting, where the losses faced by the algorithm are generated by an unspecified deterministic mechanism. Though it might be of adversarial intent, the mechanism is oblivious to internal randomization of the algorithm. In the delayed version, the loss of an action executed at time t is observed at time $t + d_t$, where the delay d_t is also chosen deterministically and obliviously. Thus, at time step t the algorithm receives observations from time steps $s \leq t$ for which $s + d_s = t$. This delay is the independent of the action chosen. The algorithm's performance is evaluated by regret, which is the difference between the algorithm's cumulative loss and the cumulative loss of the best static action in hindsight. The regret definition is the same as in the ordinary setting without delays. When all the delays are constant ($d_t = d$ for all t), the optimal regret is known to scale as $\mathcal{O}\left(\sqrt{(K+d)T \ln K}\right)$, where T is the time horizon and K is the number of actions [Cesa-Bianchi et al., 2019]. Remarkably, this bound is achieved by "delayed" Exp3, which is a minor modification of the standard Exp3 algorithm performing updates as soon as the losses become available.

The case of variable delays has previously been studied in the full information setting by Joulani et al. [2016]. They prove a regret bound of order $\sqrt{(D+T)\ln K}$, where $D = \sum_{t=1}^{T} d_t$ is the total delay. Their proof is based on a generic reduction from delayed full information feedback to full information with no delay. The applicability of this technique to the bandit setting is unclear (see Appendix 3.8). Cesa-Bianchi et al. [2019] conjecture an upper bound of order $\sqrt{(KT+D)\ln K}$ for the bandit setting with variable delays. Note that this bound cannot be improved in the general case because of the lower bound $\Omega\left(\sqrt{(K+d)T}\right)$, which holds for any d. In a recent paper, Li et al. [2019] study a harder variant of bandits, where the delays d_t remain unknown. As a consequence, if an action is played at time s and then more times in between time steps s and $s + d_s$, the learner cannot tell which specific round the loss observed at time $s + d_s$ refers to. In this harder setting, for known T, D, and $d_{\max}$, Li et al. [2019] prove a regret bound of $\widetilde{\mathcal{O}}\left(\sqrt{d_{\max} K(T+D)}\right)$. Cesa-Bianchi et al. [2018] further study an even harder setting of bandits with anonymous composite feedback. In this setting at time step t the learner observes feedback, which is a composition of partial losses of the actions taken in the last $d_{\max}$ rounds. In this setting Cesa-Bianchi et al. [2018] obtain an $\mathcal{O}\left(\sqrt{d_{\max} KT \ln K}\right)$ regret bound (which is tight to within the $\ln K$ factor, and in fact tighter than the bound of Li et al. [2019] for an easier problem).

Our paper is structured in the following way. We start by investigating the regret of Exp3 in the variable delay setting. We prove that for known T, D, and $d_{\max}$, and assuming that $d_{\max}$ is at most of order $\sqrt{(KT+D)/(\ln K)}$, "delayed" Exp3 achieves the conjectured bound of $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$. In order to remove the restriction on $d_{\max}$ and eliminate the need of its knowledge we introduce a wrapper algorithm, Skipper. Skipper prevents the wrapped bandit algorithm from making updates using observations with delay exceeding a given threshold β. This threshold acts as a tunable upper bound on the delays observed by the underlying algorithm, so if T and D are known we can choose β that attains the desired $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$ regret bound with "delayed" Exp3 wrapped within Skipper.

To dispense of the need for knowing T and D, the first approach coming to mind is the doubling trick. However, applying the standard doubling to D is problematic, because the event that the actual total delay $d_1 + \cdots + d_t$ exceeds an estimate D is observed at time $t + d_t$ rather than at time t. In order to address this issue, we consider a setting in which the algorithm observes the delay d_t at time t rather than at time $t+d_t$. To distinguish between this setting and the previous one we say that "the delay is observed at action time" if it is observed at time t and "the delay is observed at observation time" if it is observed at time $t + d_t$. Observing the delay at action time is motivated by scenarios in which a learning agent depends on feedback from a third party, for instance a server or laboratory that processes the action in order to evaluate it. In such cases, the third party might partially control the delay, and provide the agent with a delay estimate based on contingent and possibly private information. In the server example the delay could depend on workload, while the laboratory might have processing times and an order backlog. Other examples include medical imaging where the availability of annotations depends on medical professionals work schedule. Common for these examples is that the third party knows the delay before the action is taken.

Within the "delay at action time" setting we achieve a much stronger regret bound. We show that Skipper wrapping delayed Exp3 and combined with a carefully designed doubling trick enjoys an implicit regret bound of order $\min_\beta \big(|S_\beta| + \beta \ln K + (KT + D_\beta)/\beta\big)$, where D_β is the total delay of observations with delay below β. This bound is attained without any assumptions on the sequence of delays d_t and with no need for prior knowledge of T and D. The implicit bound can be relaxed to an explicit bound of $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$, however if $D_\beta \ll D$ it can be much tighter. We provide an instance of such a problem in Example 14, where we get a polynomially tighter bound.

Table 3.1 summarizes the spectrum of delayed feedback models in the bandit case and places our results in the context of prior work.

3.2.1 Additional related work

Online learning with delays was pioneered by Mesterharm [2005] — see also [Mesterharm, 2007, Chapter 8]. More recent work in the full information setting include [Zinkevich et al., 2009, Quanrud and Khashabi, 2015, Ghosh and Ramchandran, 2018]. The theme of large or unbounded delays in the full information setting was also investigated by Mann et al. [2018] and Garrabrant et al. [2016]. Other related approaches are the works by Shamir and Szlak [2017], who use a semi-adversarial model, and Chapelle

Table 3.1: Spectrum of delayed feedback settings and the corresponding regret bounds, progressing from easier to harder settings. Results marked by (*) have matching lower bounds up to the $\sqrt{\ln K}$ factor. If all the delays are identical, then $D = dT$ and (**) has a lower bound following from Cesa-Bianchi et al. [2019] and matching up to the $\sqrt{\ln K}$ factor. However, for non-identical delays the regret can be much smaller, as we show in Example 14.

Setting and reference	Regret Bound	
Fixed delay [Cesa-Bianchi et al., 2019]	$\mathcal{O}\big(\sqrt{(K+d)T\ln K}\big)$	(*)
Delay at action time [This paper]	$\mathcal{O}\left(\min_\beta\left(\lvert S_\beta\rvert + \beta\ln K + \frac{KT+D_\beta}{\beta}\right)\right)$	
Delay at observation time with known T, D [This paper]	$\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$	(**)
Anonymous, composite with known $d_{\max}$ [Cesa-Bianchi et al., 2018]	$\mathcal{O}\big(\sqrt{d_{\max}KT\ln K}\big)$	(*)

[2014], who studies the role of delays in the context of onlne advertising. Chapelle and Li [2011] perform an empirical study of the impact of delay in bandit models. This is extended in [Mandel et al., 2015]. The analysis of Exp3 in a delayed setting was initiated by Neu et al. [2014]. In the stochastic case, bandit learning with delayed feedback was studied in [Dudík et al., 2011, Vernade et al., 2017]. The results were extended to the anonymous setting by Pike-Burke et al. [2018] and by Garg and Akash [2019], and to the contextual setting by Arya and Yang [2019].

3.3 Setting and notation

We consider an oblivious adversarial multiarmed bandit setting, where K sequences of losses are generated in an arbitrary way prior to the start of the game. The losses are denoted by ℓ_t^a, where t indexes the game rounds and $a \in \{1, \dots, K\}$ indexes the sequences. We assume that all losses are in the $[0, 1]$ interval. We use the notation $[K] = \{1, \dots, K\}$ for brevity. At each round of the game the learner picks an action A_t and suffers the loss of that action. The loss $\ell_t^{A_t}$ is observed by the learner after d_t rounds, where the sequence of delays $d_1, d_2, \dots$ is determined in an arbitrary way before the game starts. Thus, at round t the learner observes the losses of prior actions A_s for which $s + d_s = t$. We assume that the losses are observed "at the end of round t", *after* the action A_t has been selected. We consider two different settings for receiving information about the delays d_t:

Delay available at observation time The delay d_t is observed when the feedback $\ell_t^{A_t}$ arrives at the end of round $t + d_t$. This corresponds to the feedback being timestamped.

Delay available at action time The delay d_t is observed at the beginning of round t, prior to selecting the action A_t.

The following learning protocol provides a formal description of our setting.

Protocol for bandits with delayed feedback

For $t = 1, 2, \ldots$

1. If *delay is available at action time*, then $d_t \geq 0$ is revealed to the learner
2. The learner picks an action $A_t \in \{1, \ldots, K\}$ and suffers the loss $\ell_t^{A_t} \in [0, 1]$
3. Pairs $\left(s, \ell_s^{A_s}\right)$ for all $s \leq t$ such that $s + d_s = t$ are observed

We measure the performance of the learner by her *expected regret* $\bar{\mathcal{R}}_T$, which is defined as the difference between the expected cumulative loss of the learner and the loss of the best static strategy in hindsight:

$$\bar{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_a \sum_{t=1}^{T} \ell_t^a.$$

This regret definition is the same as the one used in the standard multiarmed bandit setting without delay.

3.4 Delay available at observation time: Algorithms and results

This section deals with the first of our two settings, namely when delays are observed together with the losses. We first introduce a modified version of "delayed" `Exp3`, which we name Delayed Exponential Weights (`DEW`) and which is capable of handling variable delays. We then introduce a wrapper algorithm, `Skipper`, which filters out excessively large delays. The two algorithms also serve as the basis for the next section, where we provide yet another wrapper for tuning the parameters of `Skipper`.

3.4.1 Delayed Exponential Weights (`DEW`)

`DEW` is an extension of the standard exponential weights approach to handle delayed feedback. The algorithm, laid out in Algorithm 2, performs an exponential update using every individual feedback as it arrives, which means that between each prediction either zero, one, or multiple updates might occur. The algorithm assumes that the delays are bounded and that an upper bound $d_{\max} \geq \max_t d_t$ on the delays is known.

The following theorem provides a regret bound for Algorithm 2. The bound is a generalization of a similar bound in Cesa-Bianchi et al. [2019].

```
Algorithm 2: Delayed exponential weights (DEW)
Input : Learning rate $\eta$; upper bound on the delays $d_{\max}$
Truncate the learning rate: $\eta' = \min\{\eta, (4ed_{\max})^{-1}\}$;
Initialize $w_0^a = 1$ for all $a \in [K]$;
for $t = 1, 2, \dots$ do
    Let $p_t^a = \frac{w_{t-1}^a}{\sum_b w_{t-1}^b}$ for $a \in [K]$;
    Draw an action $A_t \in [K]$ according to the distribution $\boldsymbol{p_t}$ and play it;
    Observe feedback $(s, \ell_s^{A_s})$ for all $\{s : s + d_s = t\}$ and construct estimators
      $\hat{\ell}_s^a = \frac{\ell_s^a \mathbb{1}(a=A_s)}{p_s^a}$;
    Update $w_t^a = w_{t-1}^a \exp\left(-\eta' \sum_{s:s+d_s=t} \hat{\ell}_s^a\right)$;
end
```

Theorem 7. *Under the assumption that an upper bound on the delays* $d_{\max}$ *is known, the regret of Algorithm 2 with a learning rate* η *against an oblivious adversary satisfies*

$$\bar{\mathcal{R}}_T \leq \max\left\{\frac{\ln K}{\eta}, 4ed_{\max} \ln K\right\} + \eta\left(\frac{KTe}{2} + D\right),$$

where $D = \sum_{t=1}^{T} d_t$*. In particular, if* T *and* D *are known and* $\eta = \sqrt{\frac{\ln K}{\frac{KTe}{2}+D}} \leq \frac{1}{4ed_{\max}}$*, we have*

$$\bar{\mathcal{R}}_T \leq 2\sqrt{\left(\frac{KTe}{2} + D\right) \ln K}. \tag{3.1}$$

The proof of Theorem 7 is based on proving the stability of the algorithm across rounds. The proof is sketched out in Section 3.6. As Theorem 7 shows, Algorithm 2 performs well if $d_{\max}$ is small and we also have preliminary knowledge of $d_{\max}$, T, and D. However, a single delay of order T increases $d_{\max}$ up to order T, which leads to a linear regret bound in Theorem 7. This is an undesired property, which we address with the skipping scheme presented next.

3.4.2 Skipping scheme

We introduce a wrapper for Algorithm 2, called `Skipper`, which disregards feedback from rounds with excessively large delays. The regret in the skipped rounds is trivially bounded by 1 (because the losses are assumed to be in $[0, 1]$) and the rounds are taken out of the analysis of the regret of `DEW`. `Skipper` operates with an externally provided threshold β and skips all rounds where $d_t \geq \beta$. The advantage of skipping is that it provides a natural upper bound on the delays for the subset of rounds processed by DEW, $d_{\max} = \beta$. Thus, we eliminate the need of knowledge of the maximal delay in the original problem. The cost of skipping is the number of skipped rounds, denoted by $|S_\beta|$, as captured in Lemma 8. Below we provide a regret bound for the combination of `Skipper` and DEW.

Algorithm 3: Skipper

```
Input : Threshold $\beta$; Algorithm $\mathcal{A}$.
for $t = 1, 2, \dots$ do
    Get prediction $A_t$ from $\mathcal{A}$ and play it;
    Observe feedback $(s, \ell_s^{A_s})$ for all $\{s : s + d_s = t\}$, and feed it to $\mathcal{A}$ for each
      $s$ with $d_s < \beta$;
end
```

Lemma 8. *The expected regret of* `Skipper` *with base algorithm $\mathcal{A}$ and threshold parameter β satisfies*

$$\bar{\mathcal{R}}_T \leq |S_\beta| + \bar{\mathcal{R}}_{T \setminus S_\beta}, \tag{3.2}$$

where $|S_\beta|$ is the number of skipped rounds (those for which $d_t \geq \beta$) and $\bar{\mathcal{R}}_{T \setminus S_\beta}$ is a regret bound for running $\mathcal{A}$ on the subset of rounds $[T] \setminus S_\beta$ (those, for which $d_t < \beta$).

A proof of the lemma is found in Appendix 3.10. When combined with the previous analysis for `DEW`, Lemma 8 gives us the following regret bound.

Theorem 9. *The expected regret of* $\texttt{Skipper}(\beta, \texttt{DEW}(\eta, \beta))$ *against an oblivious adversary satisfies*

$$\bar{\mathcal{R}}_T \leq |S_\beta| + \max\left\{\frac{\ln K}{\eta}, 4e\beta \ln K\right\} + \eta\left(\frac{KTe}{2} + D_\beta\right), \tag{3.3}$$

where $D_\beta = \sum_{t \notin S_\beta} d_t$ is the cumulative delay experienced by `DEW`*.*

Proof. Theorem 7 holds for parameters (η, β) for `DEW` run under `Skipper`. We then apply Lemma 8. □

Corollary 10. *Assume that T and D are known and take*

$$\eta = \frac{1}{4e\beta}, \quad \beta = \sqrt{\frac{\frac{eKT/2+D}{4e} + D}{4e \ln K}}.$$

Then the expected regret of $\texttt{Skipper}(\beta, \texttt{DEW}(\eta, \beta))$ *against an oblivious adversary satisfies*

$$\bar{\mathcal{R}}_T \leq 2\sqrt{\left(\frac{KTe}{2} + (1 + 4e)D\right) \ln K}.$$

Proof. Note that $D \geq \beta|S_\beta| \Rightarrow |S_\beta| \leq \frac{D}{\beta}$. By substituting this into (3.3), observing that $D_\beta \leq D$, and substituting the values of η and β we obtain the result. □

Note that Corollary 10 recovers the regret scaling in Theorem 7, equation (3.1) within constant factors in front of D without the need of knowledge of $d_{\max}$. Similar to Theorem 7, Corollary 10 is tight in the worst case. The tuning of β still requires the knowledge of T and D. In the next section we get rid of this requirement.

3.5 Delay available at action time: Oracle tuning and results

This section deals with the second setting, where the delays are observed before taking an action. The combined algorithm introduced in the previous section relies on prior knowledge of T and D for tuning the parameters. In this section we eliminate this requirement by leveraging the added information about the delays at the time of action. The information is used in an implicit doubling scheme for tuning `Skipper`'s threshold parameter β. Additionally, the new bound scales with the experienced delay D_β rather than the full delay D and is significantly tighter when $D_\beta \ll D$. This is achieved through direct optimization of the regret bound in terms of $|S_\beta|$ and D_β, as opposed to Corollary 10, which tunes β using the potentially loose inequality $|S_\beta| \leq D/\beta$.

3.5.1 Setup

Let m index the *epochs* of the doubling scheme. In each epoch we restart the algorithm with new parameters and continually monitor the termination condition in equation (3.6). The learning rate within epoch m is set to $\eta_m = \frac{1}{4e\beta_m}$, where β_m is the threshold parameter of the epoch. Theorem 9 provides a regret bound for epoch m denoted by

$$\mathrm{Bound}_m(\beta_m) := |S^m_{\beta_m}| + 4e\beta_m \ln K + \frac{\sigma(m)eK/2 + D^m_{\beta_m}}{4e\beta_m}, \tag{3.4}$$

where $\sigma(m)$ denotes the length of epoch m and $|S^m_{\beta_m}|$ and $D^m_{\beta_m}$ are, respectively, the number of skipped rounds and the experienced delay within epoch m.

Let $\omega_m = 2^m$. In epoch m we set

$$\beta_m = \frac{\sqrt{\omega_m}}{4e\ln K} \tag{3.5}$$

and we stay in epoch m as long as the following condition holds:

$$\max\left\{|S^m_{\beta_m}|^2, \left(\frac{eK\sigma(m)}{2} + D^m_{\beta_m}\right)\ln K\right\} \leq \omega_m. \tag{3.6}$$

Since d_t is observed at the beginning of round t, we are able to evaluate condition (3.6) and start a new epoch before making the selection of A_t. This provides the desired tuning of β_m for all rounds without the need of a separate treatment of epoch transition points.

While being more elaborate, this doubling scheme maintains the intuition of standard approaches. First of all, the condition for doubling (3.6) ensures that the regret bound in each period is optimized by explicitly balancing the contribution of each term in equation (3.4). Secondly, the geometric progression of the tuning (3.5) —and thus of the resulting regret bounds— means that the total regret bound summed over the epochs can be bounded in relation to the bound in the final completed epoch.

In the following we refer to the doubling scheme defined by (3.5) and (3.6) as `Doubling`.

3.5.2 Results

The following results show that the proposed doubling scheme works as well as oracle tuning of β when the learning rate is fixed at $\eta = 1/4e\beta$. We first compare our performance to the optimal tuning in a single epoch, where we let

$$\beta_m^* = \arg\min_{\beta_m} \mathrm{Bound}_m(\beta_m) \tag{3.7}$$

be the minimizer of (3.4).

Lemma 11. *The regret bound* (3.4) *for any non-final epoch m, with the epochs and β_m controlled by* `Doubling` *satisfies*

$$\mathrm{Bound}_m(\beta_m) \leq 3\sqrt{\omega_m} \leq 3\,\mathrm{Bound}_m(\beta_m^*) + 2e^2 K \ln K + 1. \tag{3.8}$$

The lemma is the main machinery of the analysis of `Doubling` and its proof is provided in Appendix 3.10. Applying it to `Skipper`(β, `DEW`(η,β)) leads to the following main result.

Theorem 12. *The expected regret of* `Skipper`$(\beta,$ `DEW`$(\eta, \beta))$ *tuned by* `Doubling` *satisfies for any T*

$$\bar{\mathcal{R}}_T \leq 15 \min_{\beta} \left\{ |S_\beta| + 4e\beta \ln K + \frac{KT + D_\beta}{4e\beta} \right\} + 10e^2 K \ln K + 5.$$

The proof of Theorem 12 is based on Lemma 11 and is provided in Appendix 3.10.

Corollary 13. *The expected regret of* `Skipper`$(\beta,$ `DEW`$(\eta, \beta))$ *tuned by* `Doubling` *can be relaxed for any T to*

$$\bar{\mathcal{R}}_T \leq 30\sqrt{\left(\frac{KTe}{2} + (1+4e)D\right)\ln K} + 10e^2 K \ln K + 5. \tag{3.9}$$

Proof. The first term in the bound of Theorem 12 can be directly bounded using Corollary 10. □

Note that both Theorem 12 and Corollary 13 require no knowledge of T and D.

3.5.3 Comparison of the oracle and explicit bounds

We finish the section with a comparison of the oracle bound in Theorem 12 and the explicit bound in Corollary 13. Ignoring the constant and additive terms, the bounds are

$$\begin{aligned} \text{explicit} \quad &: \quad \mathcal{O}\left(\sqrt{(KT+D)\ln K}\right), \\ \text{oracle} \quad &: \quad \mathcal{O}\left(\min_\beta \left\{ |S_\beta| + \beta \ln K + \frac{KT + D_\beta}{\beta} \right\}\right). \end{aligned}$$

Note that the oracle bound is always as strong as the explicit bound. There are, however, cases where it is much tighter. Consider the following example.

Example 14. *For $t < \sqrt{KT/\ln K}$ let $d_t = T - t$ and for $t \geq \sqrt{KT/\ln K}$ let $d_t = 0$. Take $\beta = \sqrt{KT/\ln K}$. Then $D = \Theta(T\sqrt{KT/\ln K})$, but $D_\beta = 0$ (assuming that $T \geq K \ln K$) and $|S_\beta| < \sqrt{KT/\ln K}$. The corresponding regret bounds are*

$$\begin{aligned} \text{explicit} \quad &: \quad \mathcal{O}\left(\sqrt{KT\ln K + T\sqrt{KT}}\right) = \mathcal{O}(T^{3/4}), \\ \text{oracle} \quad &: \quad \mathcal{O}\left(\sqrt{KT\ln K}\right) = \mathcal{O}(T^{1/2}). \end{aligned}$$

3.6 Analysis of Algorithm 2

This section contains the main points of the analysis of Algorithm 2 leading to the proof of Theorem 7 which were postponed from Section 3.4. Full proofs are found in Appendix 3.9.

The analysis is a generalization of the analysis of delayed Exp3 in Cesa-Bianchi et al. [2019], and consists of a general regret analysis and two stability lemmas.

3.6.1 Additional notation

We let $N_t = |\{s : s + d_s \in [t, t + d_t)\}|$ denote the *stability-span* of t, which is the amount of feedback that arrives between playing action A_t and observing its feedback. Note that letting $N = \max_t N_t$ we have $N \leq 2 \max_t d_t \leq 2d_{\max}$, since this may include feedback from up to $\max_s d_s$ rounds prior to round t and up to d_t rounds after round t.

We introduce $\mathcal{Z} = (z_1, ..., z_T)$ to be a permutation of $[T] = \{1, ..., T\}$ sorted in ascending order according to the value of $z + d_z$ with ties broken randomly, and let $\Psi_i = (z_1, ..., z_i)$ be its first i elements. Similarly, we also introduce $\mathcal{Z}'_t = (z'_1, ..., z'_{N_t})$ as an enumeration of $\{s : s + d_s \in [t, t + d_t)\}$.

For a subset the integers C, corresponding to timesteps, we also introduce

$$q^a(C) = \frac{\exp\left(-\eta' \sum_{s \in C} \hat{\ell}^a_s\right)}{\sum_b \exp\left(-\eta' \sum_{s \in C} \hat{\ell}^b_s\right)}. \tag{3.10}$$

The nominator and denominator in the above expression will also be denoted by $w^a(C)$ and $W(C)$ corresponding to the definition of p^a_t.

By finally letting $C_{t-1} = \{s : s + d_s < t\}$ we have $p^a_t = q^a(C_{t-1})$.

3.6.2 Analysis of delayed exponential weights

The starting point is the following modification of the basic lemma within the Exp3 analysis that takes care of delayed updates of the weights.

Lemma 15. *Algorithm 2 satisfies*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} p^a_{t+d_t} \hat{\ell}^a_t - \min_{a \in [K]} \sum_t \hat{\ell}^a_t \leq \frac{\ln K}{\eta'} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{a=1}^{K} p^a_{t+d_t} \left(\hat{\ell}^a_t\right)^2. \tag{3.11}$$

To make use of Lemma 15, we need to figure out the relationship between $p^a_{t+d_t}$ and p^a_t. This is achieved by the following two lemmas, which are generalizations and refinements of Lemmas 1 and 2 in Cesa-Bianchi et al. [2019].

Lemma 16. *When using Algorithm 2 the resulting probabilities fulfil for every t and a*

$$p^a_{t+d_t} - p^a_t \geq -\eta' \sum_{i=1}^{N_t} q^a\left(C_{t-1} \cup \{z'_j : j < i\}\right) \hat{\ell}^a_{z'_i}, \tag{3.12}$$

where z'_j is an enumeration of $\{s : s + d_s \in [t, t + d_t)\}$.

The above lemma allows us to bound $p^a_{t+d_t}$ from below in terms of p^a_t. We similarly need to be able to upper bound the probability, which is captured in the second probability drift lemma.

Lemma 17. *The probabilities defined by* (3.10) *satisfy for any* i

$$q^a(\Psi_i) \leq \left(1 + \frac{1}{2N-1}\right) q^a(\Psi_{i-1}). \tag{3.13}$$

3.6.3 Proof sketch of Theorem 7

By using Lemma 16 to bound the left hand side of (3.11) we have

$$\sum_t \sum_a p^a_t \hat{\ell}^a_t - \min_a \sum_t \hat{\ell}^a_t \leq \frac{\ln K}{\eta'} + \frac{\eta'}{2} \sum_{t=1}^{T} \sum_{a=1}^{K} p^a_{t+d_t} \left(\hat{\ell}^a_t\right)^2 + \eta' \sum_t \sum_a \hat{\ell}^a_t \sum_{i=1}^{N_t} q^a \left(C_{t-1} \cup \{z'_j : j < i\}\right) \hat{\ell}^a_{z'_i}.$$

Repeated use of Lemma 17 bounds the second term on the right hand side by $\eta' TKe/2$ in expectation. The third term on the right hand side can be bounded by D. Taking the maximum over the two possible values of the truncated learning rate finishes the proof. □

3.7 Discussion

We have presented an algorithm for multiarmed bandits with variably delayed feedback, which achieves the $\mathcal{O}\big(\sqrt{(KT+D)\ln K}\big)$ regret bound conjectured by Cesa-Bianchi et al. [2019]. The algorithm is based on a procedure for skipping rounds with excessively large delays and refined analysis of the exponential weights algorithm with delayed observations. At the moment the skipping procedure requires prior knowledge of T and D for tuning the skipping threshold. However, if the delay information is available "at action time", as in the examples described in the introduction, we provide a sophisticated doubling scheme for tuning the skipping threshold that requires no prior knowledge of T and D. Furthermore, the refined tuning also leads to a refined regret bound of order $\mathcal{O}\big(\min_\beta \left(|S_\beta| + \beta \ln K + \frac{KT+D_\beta}{\beta}\right)\big)$, which is polynomially tighter when $D_\beta \ll D$. We provide an example of such a problem in the paper.

Our work leads to a number of interesting research questions. The main one is whether the two regret bounds are achievable when the delays are available "at observation time" without prior knowledge of D and T. Alternatively, is it possible to derive lower bounds demonstrating the impossibility of further relaxation of the assumptions? More generally, it would be interesting to have refined lower bounds for problems with variably delayed feedback. Another interesting direction is a design of anytime algorithms, which do not rely on the doubling trick. Such algorithms can be used, for example, for achieving simultaneous optimality in stochastic and adversarial setups [Zimmert and Seldin, 2019a]. While a variety of anytime algorithms is available for non-delayed bandits, the extension to delayed feedback does not seem trivial. Some of these questions are addressed in a follow-up work by Zimmert and Seldin [2019b].

Supplementary Material for Chapter 3

3.8 Alternative approaches

This appendix is an addition to the discussion of relevant literature in the introduction.

The present paper follows an approach to delayed feedback based on explicitly analysing exponential weights with delays and considering the stability of this class of algorithms. An alternative approach in literature is instead to construct a reduction from the delayed case to the undelayed case and thus circumventing the need for direct analysis of the underlying algorithm, since this will usually be well established. Such a reduction is done in the full information case by Joulani et al. [2016] but with no mention of how it might apply to the bandit case. Below we briefly sketch their reduction in the case of OCO with linear loss functions, which specializes to bandits.

Let t index time and s index *virtual rounds*, meaning rounds where an update is made. In other words in every round t the algorithm makes a prediction, while in every round s the algorithm receives *one* point of feedback. Quantities indexed by virtual rounds are further denoted by a tilde. $\tilde{\tau}_s$ is the number of virtual rounds (equivalently updates) between when the action giving rise to loss $\tilde{\ell}_s$ is played and the loss is received. Let the function ρ map a virtual round s where $\tilde{\ell}_s$ is observed to the round $t = \rho(s)$ where it is played. As such $\ell_{\rho(s)} = \tilde{\ell}_s$, but $p_{\rho(s)} = \tilde{p}_{s-\tilde{\tau}_s}$.

Deterministic case: In the full information case for deterministic losses, the actions (probability distributions) played by the algorithm does not depend on any randomness, since the feedback is not dependent on the action played. The expected regret can thus be regarded as deterministic, and the following reduction can be carried out:

$$\begin{aligned}
\mathcal{R}_T &= \sum_t \left[\sum_a \ell_t^a p_t^a - \ell_t^\star \right] \\
&= \sum_s \left[\sum_a \ell_{\rho(s)}^a p_{\rho(s)}^a - \ell_{\rho(s)}^\star \right] \\
&= \sum_s \left[\sum_a \tilde{\ell}_s^a \tilde{p}_{s-\tilde{\tau}_s}^a - \tilde{\ell}_s^\star \right] \\
&= \sum_s \sum_a \tilde{\ell}_s^a \left(\tilde{p}_{s-\tilde{\tau}_s}^a - \tilde{p}_s^a \right) + \sum_s \left[\sum_a \tilde{\ell}_s^a \tilde{p}_s^a - \tilde{\ell}_s^\star \right],
\end{aligned}$$

where we let $\star$ denote the optimal action in hindsight. The point of this calculation is that the final term above is the regret of the undelayed base algorithm, while the first term is an additive drift term, similar to what we are considering in Lemma 16.

Conditional case: To extend this to bandits, we need to consider the case where the actions (probability distributions) of the algorithm depends on the internal randomness. The expected regret then requires taking expectation over this randomness:

$$\begin{aligned}
\bar{\mathcal{R}}_T &= \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_t \left[\sum_a \ell_t^a p_t^a - \ell_t^\star \right] \right] \\
&= \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_s \sum_a \tilde{\ell}_s^a \left(\tilde{p}_{s-\tilde{\tau}_s}^a - \tilde{p}_s^a \right) \right] + \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_s \left[\sum_a \tilde{\ell}_s^a \tilde{p}_s^a - \tilde{\ell}_s^\star \right] \right].
\end{aligned}$$

Now however the final term is no longer the expected regret of the underlying algorithm without delays, since the conditional expectations taken here are not the same as they would be for the undelayed algorithm. In particular the order of the conditional expectations might be different since the delays are not the same, so the reduction is not directly applicable.

3.9 Full proof of Theorem 7

This appendix contains the full analysis of Algorithm 2, i.e., proofs of the lemmas in Section 3.6 and the full proof of Theorem 7.

3.9.1 Proof of Lemma 15

We consider the quantity

$$
\begin{aligned}
\frac{W_t}{W_{t-1}} &= \frac{\sum_a w_{t-1}^a \prod_{s:s+d_s=t} \exp\left(-\eta' \hat{\ell}_s^a\right)}{W_{t-1}} \\
&= \sum_a p_t^a \prod_{s:s+d_s=t} \exp\left(-\eta' \hat{\ell}_s^a\right) \\
&\leq \sum_a p_t^a \sum_{s:s+d_s=t} \exp\left(-\eta' \hat{\ell}_s^a\right) \\
&\leq \sum_a p_t^a \sum_{s:s+d_s=t} \left(1 - \eta' \hat{\ell}_s^a + \frac{\eta'^2}{2}\left(\hat{\ell}_s^a\right)^2\right) \\
&= \sum_{s:s+d_s=t} \left(1 - \eta' \sum_a p_t^a \hat{\ell}_s^a + \frac{\eta'^2}{2} \sum_a p_t^a \left(\hat{\ell}_s^a\right)^2\right) \\
&= 1 + |\{s : s + d_s = t\}| - 1 - \eta' \sum_{s:s+d_s=t} \sum_a p_t^a \hat{\ell}_s^a + \frac{\eta'^2}{2} \sum_{s:s+d_s=t} \sum_a p_t^a \left(\hat{\ell}_s^a\right)^2 \\
&\leq \exp\left(|\{s : s + d_s = t\}| - 1 - \eta' \sum_{s:s+d_s=t} \sum_a p_t^a \hat{\ell}_s^a + \frac{\eta'^2}{2} \sum_{s:s+d_s=t} \sum_a p_t^a \left(\hat{\ell}_s^a\right)^2\right),
\end{aligned}
$$

where the first inequality uses that each $\exp\left(-\eta' \hat{\ell}_s^a\right)$ is in $(0, 1]$, the second inequality uses $e^x \leq 1 + x + x^2/2$ for $x \leq 0$, and the final inequality uses $e^x \geq 1 + x$ for all x.

By a telescoping sum and the above we get

$$
\frac{W_T}{W_0} \leq \exp\left(-\eta' \sum_t \sum_{s:s+d_s=t} \sum_a p_t^a \hat{\ell}_s^a + \frac{\eta'^2}{2} \sum_t \sum_{s:s+d_s=t} \sum_a p_t^a \left(\hat{\ell}_s^a\right)^2\right), \tag{3.14}
$$

using that $\sum_{t=1}^T |\{s : s + d_s = t\}| \leq T$. We also lower bound this fraction as

$$
\begin{aligned}
\frac{W_T}{W_0} &\geq \frac{\max_a \exp\left(-\eta' \sum_{s:s+d_s \leq T} \hat{\ell}_s^a\right)}{K} \\
&\geq \frac{\max_a \exp\left(-\eta' \sum_{s=1}^T \hat{\ell}_s^a\right)}{K} \\
&\geq \frac{\exp\left(-\eta' \min_a \sum_{s=1}^T \hat{\ell}_s^a\right)}{K}.
\end{aligned} \tag{3.15}
$$

The proof is completed by combining (3.14) and (3.15), taking logarithms and rearranging, and noting that the sums of the form $\sum_t \sum_{s:s+d_s=t}$ only include each value of s once, and thus are equivalent to summing over s and identifying $t = s + d_s$. □

3.9.2 Proof of Lemma 16

Note for any set of integers C containing a value x, we have

$$W(C) = \sum_a e^{-\eta' \hat{\ell}_x^a} e^{-\eta' \sum_{s \in C \backslash \{x\}} \hat{\ell}_s^a} \leq \sum_a e^{-\eta' \sum_{s \in C \backslash \{x\}} \hat{\ell}_s^a} = W(C \backslash \{x\}),$$

which means

$$q^a(C) = \frac{w^a(C)}{W(C)} \geq \frac{w^a(C)}{W(C \backslash \{x\})} = e^{-\eta' \hat{\ell}_x^a} \frac{w^a(C \backslash \{x\})}{W(C \backslash \{x\})} = e^{-\eta' \hat{\ell}_x^a} q^a(C \backslash \{x\}).$$

This in turn implies

$$q^a(C) - q^a(C \backslash \{x\}) \geq \left(e^{-\eta' \hat{\ell}_x^a} - 1 \right) q^a(C \backslash \{x\}) \geq -\eta' \hat{\ell}_x^a q^a(C \backslash \{x\}).$$

Telescoping this over the individual observations $z_1', ..., z_{N_t}'$ we get

$$\begin{aligned} p_{t+d_t}^a - p_t^a &= q^a(C_{t+d_t-1}) - q^a(C_{t-1}) \\ &= \sum_{i=1}^{N_t} q^a \left(C_{t-1} \cup \{z_j' : j \leq i\} \right) - q^a \left(C_{t-1} \cup \{z_j' : j < i\} \right) \\ &\geq -\eta' \sum_{i=1}^{N_t} \hat{\ell}_{z_i'}^a q^a \left(C_{t-1} \cup \{z_j' : j < i\} \right) \end{aligned}$$

3.9.3 Proof of Lemma 17

We prove the lemma by induction. For the base case, consider $i = 1$, where $\Psi_0 = \emptyset$, and thus $q^a(\Psi_{i-1}) = q^a(\Psi_0) = 1/K$. The maximal increase of q^a by making a single observation will be if another arm is chosen and receives a loss of 1, making the loss estimator equal to K. This means

$$q^a(\Psi_i) \leq \frac{1}{K-1 + e^{-\eta' K}} \leq \frac{1}{K - \eta' K} \leq \frac{1/K}{1 - \frac{1}{e2N}} \leq \frac{1/K}{1 - \frac{1}{2N}} = \frac{1}{K} \left(1 + \frac{1}{2N-1} \right),$$

where first use $e^x \geq 1 + x$ for all x and secondly use the upper bound on η'. since $1/K = q^a(\Psi_0)$ the base case is shown.

For the general case, assume that the lemma holds for $i - 1$. We first show that

$$\begin{aligned} q^a(\Psi_{i-1}) &\geq e^{-\eta' \hat{\ell}_{z_i}^a} q^a(\Psi_{i-1}) \\ &= \frac{w^a(\Psi_i)}{W(\Psi_{i-1})} \\ &= \frac{q^a(\Psi_i) W(\Psi_i)}{W(\Psi_{i-1})} \\ &= q^a(\Psi_i) \sum_b \frac{e^{-\eta' \hat{\ell}_{z_i}^b} w^b(\Psi_{i-1})}{W(\Psi_{i-1})} \\ &= q^a(\Psi_i) \sum_b e^{-\eta' \hat{\ell}_{z_i}^b} q^b(\Psi_{i-1}) \\ &\geq q^a(\Psi_i) \left(1 - \eta' \sum_b \hat{\ell}_{z_i}^b q^b(\Psi_{i-1}) \right). \end{aligned} \tag{3.16}$$

By expanding the loss estimator we get

$$\sum_b \hat{\ell}^b_{z_i} q^b(\Psi_{i-1}) = \hat{\ell}^{A_{z_i}}_{z_i} q^{A_{z_i}}(\Psi_{i-1}) \leq \frac{q^{A_{z_i}}(\Psi_{i-1})}{q^{A_{z_i}}(C_{z_i-1})} = \frac{q^{A_{z_i}}(\Psi_{i-1})}{q^{A_{z_i}}(\Psi_{l(i)})}, \tag{3.17}$$

by using $\ell^b_{z_i}\mathbb{1}(A_{z_i} = b) \leq 1$ for $b = A_{z_i}$ and the loss estimator is identically zero for all other b. We here define $l(i)$ as the index in $\mathcal{Z}$ of the last observation before round z_i. We now consider the difference in these indices, namely $(i-1) - l(i)$.

Note that the loss from z_i is observed at time $z_i + d_{z_i}$, but the losses from rounds $z_{i-1}, z_{i-2}, ...$ could potentially also be observed at this point. This means that all observations of losses from rounds $\Psi_{i-1} \backslash \Psi_{l(i)}$ are found in $[z_i, z_i + d_{z_i}]$. As maximally N observations can be made both in $[z_i, z_i + d_{z_i})$ and in $[z_i + d_{z_i}, z_i + d_{z_i}]$ by assumption, and these $2N$ observations must include the observation of the loss from round z_i, we have a bound of

$$(i-1) - l(i) \leq 2N - 1. \tag{3.18}$$

Telescoping the probability ratio and using the inductive assumption, we thus have

$$\begin{aligned} \frac{q^{A_{z_i}}(\Psi_{i-1})}{q^{A_{z_i}}(\Psi_{l(i)})} &= \prod_{j=l(i)+1}^{i-1} \frac{q^{A_{z_i}}(\Psi_j)}{q^{A_{z_i}}(\Psi_{j-1})} \\ &\leq \prod_{j=l(i)+1}^{i-1} \left(1 + \frac{1}{2N-1}\right) \\ &= \left(1 + \frac{1}{2N-1}\right)^{2N-1} \\ &\leq e. \end{aligned} \tag{3.19}$$

Inserting this into (3.16) and using the upper bound on the learning rate gives us

$$\begin{aligned} q^a(\Psi_{i-1}) &\geq q^a(\Psi_i)(1 - \eta' e) \\ &\geq q^a(\Psi_i)\left(1 - \frac{1}{2N}\right), \end{aligned}$$

which rearranges to the lemma statement. This concludes the inductive step □.

3.9.4 Full proof of Theorem 7

We start by combining Lemmas 15, 16 and 17 in the following way. By using Lemma 16 to bound the left hand side of (3.11) we have

$$\sum_t \sum_a p_{t+d_t} \hat{\ell}^a_t \geq \sum_t \sum_a p^a_t \hat{\ell}^a_t - \eta' \sum_t \sum_a \hat{\ell}^a_t \sum_{i=1}^{N_t} q^a\left(C_{t-1} \cup \{z'_j : j < i\}\right) \hat{\ell}^a_{z'_i}, \tag{3.20}$$

subtracting the final term gives us:

$$\sum_t \sum_a p_t^a \hat{\ell}_t^a - \min_a \sum_t \hat{\ell}_t^a \leq \frac{\ln K}{\eta'} + \frac{\eta'}{2} \sum_{t=1}^T \sum_{a=1}^K p_{t+d_t}^a \left(\hat{\ell}_t^a\right)^2 + \eta' \sum_t \sum_a \hat{\ell}_t^a \sum_{i=1}^{N_t} q^a \left(C_{t-1} \cup \{z_j' : j < i\}\right) \hat{\ell}_{z_i'}^a. \tag{3.21}$$

Where we note that the left hand side becomes the expected regret when taking expectations over the choice of A_t.

The second term on the right hand side of (3.21) can be bounded by repeated use of Lemma 17:

$$\begin{aligned}
\sum_t \sum_a p_{t+d_t}^a \left(\hat{\ell}_t^a\right)^2 &= \sum_t \sum_a p_t^a \frac{p_{t+d_t}^a}{p_t^a} \left(\hat{\ell}_t^a\right)^2 \\
&= \sum_t \sum_a p_t^a \left(\hat{\ell}_t^a\right)^2 \prod_{i=1}^{N_t} \frac{q^a(C_{t-1} \cup \{z_j' : j \leq i\})}{q^a(C_{t-1} \cup \{z_j' : j < i\})} \\
&\leq \sum_t \sum_a p_t^a \left(\hat{\ell}_t^a\right)^2 \left(1 + \frac{1}{2N-1}\right)^{N_t} \\
&\leq \sum_t \sum_a p_t^a \left(\hat{\ell}_t^a\right)^2 \left(1 + \frac{1}{2N-1}\right)^{2N-1} \\
&\leq \sum_t \sum_a p_t^a \left(\hat{\ell}_t^a\right)^2 e,
\end{aligned}$$

which in expectation is bounded by TKe.

The final term in (3.21) requires a bit more work. We first note that:

$$\mathbb{E}\left[\sum_t \sum_a \hat{\ell}_t^a \sum_{i=1}^{N_t} q^a \left(C_{t-1} \cup \{z_j' : j < i\}\right) \hat{\ell}_{z_i'}^a\right] \leq \sum_t N_t,$$

since t is not part of the enumeration z_j', so the two expectations are taken independently: $\mathbb{E}[\hat{\ell}_{z_j'}^a] \leq 1$ and $\mathbb{E}[\hat{\ell}_t^a] \leq 1$. Additionally we use that q^a is a distribution. We now note that summing over t or s is equivalent in the above, i.e.,

$$\sum_t N_t \leq \sum_t |\{s : s + d_s \in [t, t + d_t)\}| = \sum_s |\{t : s + d_s \in [t, t + d_t)\}|,$$

since counting in how many intervals every loss is observed in is the same as counting how many losses are observed in every interval. Note that we implicitly restrict both s and t to be in $[T]$.

We now split this

$$\sum_s |\{t : s + d_s \in [t, t + d_t)\}| = \sum_s |\{t > s : s + d_s \in [t, t + d_t)\}| + |\{t < s : s + d_s \in [t, t + d_t)\}|$$

and bound the first term as

$$\begin{aligned}|\{t > s : s + d_s \in [t, t + d_t)\}| &\leq |\{t > s : t \leq s + d_s\} \backslash \{t > s : t + d_t < s + d_s\}| \\ &\leq d_s - |\{t > s : t + d_t < s + d_s\}|, \end{aligned} \tag{3.22}$$

The second term is similarly bounded as

$$|\{t < s : s + d_s \in [t, t + d_t)\}| \leq |\{t < s : s + d_s < t + d_t\}|. \tag{3.23}$$

Finally we note that by the prior equivalency of summing over t or s, the negative term in (3.22) cancel with (3.23) once summed. This bounds the final term of (3.21) by D and results in

$$\bar{\mathcal{R}}_T \leq \frac{\ln K}{\eta'} + \eta' \left(\frac{eKT}{2} + D \right). \tag{3.24}$$

We now consider the truncation of the learning rate which is mandated by Lemma 17. If the input learning rate fulfils $\eta \leq (2eN)^{-1}$ then $\eta' = \eta$, and (3.24) simply becomes

$$\bar{\mathcal{R}}_T \leq \frac{\ln K}{\eta} + \eta \left(\frac{eKT}{2} + D \right),$$

where η is the input learning rate.

If instead the learning rate is truncated, meaning the input learning rate is larger than $(2eN)^{-1}$, the algorithm uses $\eta' = (2eN)^{-1}$, meaning (3.24) becomes

$$\bar{\mathcal{R}}_T \leq 2eN \ln K + \frac{\frac{eKT}{2} + D}{2eN} \leq 2eN \ln K + \eta \left(\frac{eKT}{2} + D \right).$$

Taking the maximum of these two regret bounds finalizes the proof for any input learning rate η. □

3.10 Additional proofs

3.10.1 Proof of Lemma 8

Consider first skipping just one round s. We then have

$$\begin{aligned}\bar{\mathcal{R}}_T &:= \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_t \sum_a p_t^a \ell_t^a \right] - \min_a \sum_t \ell_t^a \\ &\leq \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_a p_s^a \ell_s^a \right] - \min_a \ell_s^a + \mathop{\mathbb{E}}_{A_1,\ldots,A_T} \left[\sum_{t \neq s} \sum_a p_t^a \ell_t^a \right] - \min_a \sum_{t \neq s} \ell_t^a \\ &\leq 1 + \mathop{\mathbb{E}}_{A_1,\ldots,A_{s-1},A_s,\ldots,A_T} \left[\sum_{t \neq s} \sum_a p_t^a \ell_t^a \right] - \min_a \sum_{t \neq s} \ell_t^a \\ &= 1 + \bar{\mathcal{R}}_{T \backslash \{s\}},\end{aligned}$$

where the first inequality uses $\min_a [x_a + y_a] \geq \min_a x_a + \min_a y_a$ for any x, y and the second inequality uses $\ell_s^a \in [0, 1]$ for all a and p_s^a being a distribution. In this line we also use the fact that no p_t depend on A_s. The proof is then complete by iterating this argument over all $s \in C$.

3.10.2 Proof of Lemma 11

The first inequality follows directly from insertion of $\beta_m = \sqrt{\omega_m}/4e \ln K$ into (3.4) and using the doubling condition for staying in the epoch (3.6).

For the second condition, we consider several cases of the optimal value in epoch m:

Case 1 If $\beta_m^* \geq \beta_m$ we have

$$\mathrm{Bound}_m(\beta_m^*) \geq 4e\beta_m^* \ln K \geq 4e\beta_m \ln K = \sqrt{\omega_m} \geq \frac{\mathrm{Bound}_m(\beta_m)}{3}, \tag{3.25}$$

In all following cases we consider $\beta_m^* < \beta_m$.

Case 2 We now consider the case where $\beta_m^* < \beta_m$ and the doubling happened because the number of skipped rounds grew to large. This implies the following inequality

$$\left(|S_{\beta_m}^m| + 1\right)^2 \geq \omega_m,$$

leading to

$$\mathrm{Bound}_m(\beta_m^*) \geq |S_{\beta_m^*}^m| \geq |S_{\beta_m}^m| \geq \sqrt{\omega_m} - 1 \geq \frac{\mathrm{Bound}_m(\beta_m)}{3} - 1, \tag{3.26}$$

where the second inequality comes from the assumption that $\beta_m^* < \beta_m$, meaning at least as many delays are skipped using β_m^*, as this is a lower threshold for skipping.

Case 3 If $\beta_m^* < \beta_m$ and the doubling instead happened because the second term grew too large, we have the following inequality:

$$\left(Ke/2 \cdot (\sigma(m) + 1) + D_{\beta_m}^m + \beta_m\right) \ln K \geq \omega_m. \tag{3.27}$$

In this case we have

$$\begin{aligned}
\mathrm{Bound}_m(\beta_m^*) &\geq \frac{eK\sigma(m)/2 + D_{\beta_m^*}^m}{\beta_m^*} \\
&= \frac{\beta_m}{\beta_m^*}\left(\frac{eK\sigma(m)/2 + D_{\beta_m}^m}{\beta_m} + \frac{D_{\beta_m^*}^m - D_{\beta_m}^m}{\beta_m}\right) \\
&= \frac{\beta_m}{\beta_m^*}\left(\frac{eK\sigma(m)/2 + D_{\beta_m}^m}{\beta_m} - \Delta|S|\right) && (3.28) \\
&= \frac{\beta_m}{\beta_m^*}\left(4e\sqrt{\omega_m} - \frac{eK}{2\beta_m} - 1 - \Delta|S|\right) && (3.29) \\
&= \frac{\beta_m}{\beta_m^*}\left(4e\sqrt{\omega_m} - \frac{2e^2K\ln K}{\sqrt{\omega_m}} - 1 - \Delta|S|\right), && (3.30)
\end{aligned}$$

where (3.28) uses $D_{\beta_m}^m \leq D_{\beta_m^*}^m + \beta_m\Delta|S|$ for $\Delta|S| = |S_{\beta_m^*}^m| - |S_{\beta_m}^m|$. For (3.29) we use (3.27) with $\beta_m = \sqrt{\omega_m}/4e \ln K$.

Again we consider cases, this time of (3.30). Assume first

$$4e\sqrt{\omega_m} - \frac{2e^2 K \ln K}{\sqrt{\omega_m}} - 1 - \Delta|S| \geq 2e\sqrt{\omega_m}.$$

which means

$$\text{Bound}_m(\beta_m^*) \geq \frac{\beta_m}{\beta_m^*} 2e\sqrt{\omega_m} \geq \text{Bound}_m(\beta_m).$$

If we instead assume

$$4e\sqrt{\omega_m} - \frac{2e^2 K \ln K}{\sqrt{\omega_m}} - 1 - \Delta|S| \leq 2e\sqrt{\omega_m},$$

which implies

$$\Delta|S| \geq 2e\sqrt{\omega_m} - \frac{2e^2 K \ln K}{\sqrt{\omega_m}} - 1,$$

we directly have

$$\text{Bound}_m(\beta_m^*) \geq \Delta|S| \geq \text{Bound}_m(\beta_m) - 2e^2 K \ln K - 1, \tag{3.31}$$

where we have used $\sqrt{\omega_m} \geq 1$. Note that this final inequality is the worst case of case 3.

Finally we compare the cases: Noting that they are exhaustive and by comparing (3.25), (3.26) and (3.31) the lemma is proven. □

3.10.3 Proof of Theorem 12

The idea of the proof is to use the nature of the doubling schema in the usual way, combined with Lemma 11 for the second to last epoch.

Let M be the total number of epochs in the doubling schema:

$$\begin{aligned}
\bar{\mathcal{R}}_T &\leq \sum_{m=}^{M} \text{Bound}_m(\beta_m) \\
&\leq \sum_{m=1}^{M} 3\sqrt{\omega_m} \\
&= \sum_{m=1}^{M} 3\sqrt{2}^m \\
&= 3\frac{\sqrt{2}^{M+1} - 1}{\sqrt{2} - 1} \\
&\leq \frac{6}{\sqrt{2} - 1}\sqrt{2^{M-1}} \\
&= \frac{6}{\sqrt{2} - 1}\sqrt{\omega_{M-1}} \\
&\leq \frac{2}{\sqrt{2} - 1}\left(3\,\text{Bound}_{M-1}(\beta_{M-1}^*) + 2e^2 K \ln K + 1\right).
\end{aligned}$$

The proof is finalised by

$$\mathrm{Bound}_{M-1}(\beta^*_{M-1}) \leq \min_{\beta} \left\{ |S_\beta| + 4e\beta \ln K + \frac{KT + D_\beta}{4e\beta} \right\}. \quad \square$$

Chapter 4

Conclusion

In this thesis we have explored how the framework of multi-armed bandits can be used to understand two variations of online learning.

We have shown that one point of extra feedback is enough to circumvent a recently proven impossibility in multi-armed bandits of having a regret scaling linearly in the effective loss range. For two point feedback this desired scaling is achieved by an adaptive algorithm requiring no prior information about the learning setting. We have shown that this regret scaling is near-optimal and further that our algorithm achieves constant regret for stochastic losses while maintaining the scaling in the effective loss range. The algorithm requires no knowledge about the stochasticity in order to enjoy this regret bound, thereby allowing it to exploit two kinds of easiness at once while being robust to the worst case scenario.

We have further explored the setting of nonstochastic multi-armed bandits with arbitrary delays. Here we prove a recent conjecture in two steps. First we generalized the analytical approach of algorithmic stability from the fixed delays setting. This proves the conjectured regret bound for arbitrary, but bounded delays. Secondly we introduced a meta-algorithm that skips feedback with excessive delays. This alleviates the necessity of bounded delays for the combined algorithm, thereby proving the conjecture for any delay sequence. For the slightly easier learning setting where the learner has access to the present delay prior to making a prediction we designed a novel tuning scheme. With this our algorithm is able to achieve a much better regret bound for certain delay sequences, with a potential polynomial improvement. This is the case while maintaining the conjectured bound and without requiring knowledge of the time horizon or cumulative delay.

4.1 Future directions

Multi-armed bandits is a lively and active field and our knowledge keeps growing. As such this thesis presents just a few steps towards understanding online learning. For the specific models explored in this thesis, several open questions emerge:

In the case of the small effective loss range we do not know of a lower bound for the stochastic case. Such a lower bound would allow for an approach towards simultaneous optimality in addition to the current simultaneous adaptivity.

As Cesa-Bianchi and Shamir [2018] have shown that aggregate information about the loss vector also allows for adaptation to the effective loss range, it could be interesting to approach what a minimal, sufficient amount of additional information might be.

For the setting of delayed feedback, the question of the best possible regret scaling is left open. While the regret bound for the fixed delay setting of Cesa-Bianchi et al. [2019] is tight up to logarithmic factors, there exist no lower bound for the general case. Our improved regret bound for certain delay sequences suggest that the conjectured bound is not tight. The follow-up work of Zimmert and Seldin [2019b] discusses this however the question remains open.

Previously the case of arbitrary delays has been studied in the stochastic case [Joulani et al., 2016]. Achieving best-of-both-worlds results for arbitrary delays is an interesting open question.

Bibliography

Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

Sakshi Arya and Yuhong Yang. Randomized allocation with nonparametric estimation for contextual multi-armed bandits with delayed rewards. *arXiv preprint, arXiv:1902.00819*, 2019.

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

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

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

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

Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference on Learning Theory*, pages 508–528, 2019.

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

Nicolò Cesa-Bianchi and Ohad Shamir. Bandit regret scaling with the effective loss range. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2018.

Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51, 2005.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66, 2007.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Nonstochastic bandits with composite anonymous feedback. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2018.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *The Journal of Machine Learning Research*, 20(1):613–650, 2019.

Olivier Chapelle. Modeling delayed feedback in display advertising. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, 2014.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. *arXiv preprint arXiv:1703.03478*, 2017.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 459–467, 2014.

Miroslav Dudík, Daniel J. Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2011.

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

Siddhant Garg and Aditya Kumar Akash. Stochastic bandits with delayed composite anonymous feedback. *arXiv preprint, arXiv:1910.01161*, 2019.

Scott Garrabrant, Nate Soares, and Jessica Taylor. Asymptotic convergence in online learning with unbounded delays. *arXiv preprint, arXiv:1604.05280*, 2016.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Avishek Ghosh and Kannan Ramchandran. Online scoring with delayed information: A convex optimization viewpoint. In *the proceedings of the Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvári. Delay-tolerant online convex optimization: Unified analysis and adaptive-gradient algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Satyen Kale. Multiarmed bandits with limited expert advice. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2014.

Robert Kleinberg and Nicole Immorlica. Recharging bandits. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 309–319. IEEE, 2018.

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

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

Bingcong Li, Tianyi Chen, and Georgios B. Giannakis. Bandit online learning with unknown delays. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

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

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

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. The queue method: Handling delay, heuristics, prior data, and evaluation in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Timothy A Mann, Sven Gowal, Ray Jiang, Huiyi Hu, Balaji Lakshminarayanan, and Andras Gyorgy. Learning from delayed outcomes with intermediate observations. *arXiv preprint, arXiv:1807.09387*, 2018.

Chris Mesterharm. On-line learning with delayed label feedback. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2005.

Chris Mesterharm. *Improving Online Learning*. PhD thesis, Department of Computer Science, Rutgers University, 2007.

Gergely Neu, Andras Gyorgy, Csaba Szepesvari, and Andras Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59(3), 2014.

Ciara Pike-Burke, Shipra Agrawal, Csaba Szepesvari, and Steffen Grünewälder. Bandits with delayed, aggregated anonymous feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

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

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

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

Yevgeny Seldin, Koby Crammer, and Peter L. Bartlett. Open problem: Adversarial multiarmed bandits with limited advice. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2013.

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

Shai Shalev-Shwartz. Online learning and online convex optimization. *Machine Learning*, 4(2):107–194, 2011.

Ohad Shamir and Liran Szlak. Online learning with local permutations and delayed feedback. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Csaba Szepesvári. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

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

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. In *Advances in Neural Information Processing Systems*, pages 2909–2918, 2018a.

Tobias Sommer Thune and Yevgeny Seldin. Adaptation to easy data in prediction with limited advice. *arXiv preprint arXiv:1807.00636*, 2018b.

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, pages 6538–6547, 2019.

Claire Vernade, Olivier Cappé, and Vianney Perchet. Stochastic bandit models for delayed conversions. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2017.

Vladimir Vovk. Aggregating strategies. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 1990.

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

Olivier Wintenberger. Optimal learning with Bernstein online aggregation. *Machine Learning*, 106, 2017.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for stochastic and adversarial bandits. Technical report, https://arxiv.org/abs/1807.07623, 2018.

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

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. *arXiv preprint, arXiv:1910.06054*, 2019b.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692, 2019.

Martin Zinkevich, John Langford, and Alex J Smola. Slow learners are fast. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2009.