Adversarial Thresholding Semi-Bandits

Craig Steven Bower

A thesis submitted in partial fulfilment for the degree of Doctor of Philosophy

University of Derby

2020

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(Craig Steven Bower)

Acknowledgements

Whilst the process of undertaking my PhD has been thoroughly enjoyable, I was only aware of a fraction of the challenge it would be. The joy of piecing a proof together and finally getting a section of code to work, certainly outweighs all of the frustration that comes with getting there. Therefore, I am immensely grateful to everyone that has supported me along this journey.

Firstly, I would like to express sincere thanks to my Director of Studies Professor Ashiq Anjum for his invaluable insight, our varied conversations and his ability to always see the bigger picture. I cherish the friendship that we have built and look forward to collaborating together in the future. I would also like to thank my second supervisor, Dr. Ovidiu Bagdasar for his guidance, support and many cups of coffee all the way back to the beginning of my masters. I am also grateful for him spending time reading my mathematical proofs. I will definitely buy the next round. After seeing both Prof. Anjum and Dr. Bagdasar working up close, I have learned the dedication required to pursue a career in academic research and hope to model myself on their work-ethic at least. I would also like to thank Professor Yong Xue for not only willingly supporting my studies at such a later stage, but for his sage advice in the final push. To my physics collaborators Dr. Peter Chochula, Pete Bond from the ALICE DCS team and the Transition Radiation Detector expert Dr. Minjung Kweon for their wealth of knowledge and their ability to interpret my many lists of questions. I would also like to thank Dr. Lee Barnby for his support and friendship throughout my studies. I look forward to the detector starting up again in 2021 and working flawlessly. Finally to my family, for their unconditional love and my wife Jen, without whom I would not have the opportunity to do this. You are my inspiration and have given me the strength to see this through. I dedicate this thesis to you all.

Abstract

The classical multi-armed bandit is one of the most common examples of sequential decision-making, either by trading-off between exploiting and exploring arms to maximise some payoff or purely exploring arms until the optimal arm is identified. In particular, a bandit player wanting to only pull arms with stochastic feedback exceeding a given threshold, has been studied extensively in a pure exploration context. However, numerous applications fail to be expressed, where a player wishes to balance the need to observe regions of an uncertain environment that are currently interesting (exploit) and checking if neglected regions have become interesting since last observed (explore).

We introduce the *adversarial thresholding semi-bandit problem*: a non-stochastic bandit model, where a player wants to only pull (potentially several) arms with feedback meeting some threshold condition. Our main objective is to design algorithms that meet the requirements of the adversarial thresholding semi-bandit problem theoretically, empirically and algorithmically, for a given application. In other words, we want to develop a machine that learns to select options according to some threshold condition and adapts quickly if the feedback from selecting an option unexpectedly changes. This work has many real-world applications and is motivated by online detector control monitoring in high energy physics experiments, on the Large Hadron Collider.

We begin by describing the adversarial thresholding semi-bandit problem (ATSBP) in terms of a multi-armed bandit with multiple plays and extending the stochastic thresholding bandit problem to the adversarial setting. The adversarial thresholding exponentially-weighted exploration and exploitation with multiple plays algorithm (**T-Exp3.M**) and an algorithm combining label efficient prediction (**LET-Exp3.M**), are introduced that satisfy theoretical and computational Research specifications, but either perform poorly or fail completely under certain threshold conditions. To meet

empirical performance requirements, we propose the dynamic label efficient adversarial thresholding exponentially-weighted exploration and exploitation with multiple plays algorithm (**dLET-Exp3.M**). Whilst computational requirements match those for **T-Exp3.M**, theoretical upper bounds on performance are proven to be worse. We also introduce an ATSBP algorithm (**AliceBandit**) that decomposes the action of pulling an arm into selection and observation decisions. Computational complexity and empirical performance under two different threshold conditions are significantly improved, compared with exponentially-weighted adversarial thresholding semi-bandits. Theoretical upper bounds on performance are also significantly improved, for certain environments.

In the latter part of this thesis, we address the challenge of efficiently monitoring multiple condition parameters in high-energy experimental physics. Due to the extreme conditions experienced in heavy-ion particle colliders, the power supply to any device exceeding safe operating parameters is automatically shut down or *tripped* to preserve integrity and functionality of the device. Prior to recent upgrades, a device or *channel* trip would halt data-taking for the entire experiment. Post-trip recovery requires a costly procedure both in terms of expertise and data-taking time. After the completion of the current upgrading phase (scheduled for 2021), the detector will collect data continuously. In this new regime, a channel trip will result in only the affected components of the experiment being shut down. However, since the new upgraded experiment will enable data-taking to increase by a factor of 100, each trip will have a significant impact on the experiments ability to provide physicists with reliable data to analyse. We demonstrate that adversarial thresholding semi-bandits efficiently identify device channels either exceeding a fixed threshold or deviating by more than a prescribed range prior to a trip, extending the state-of-the-art in high-energy physics detector control.

Contents

Abstract viii

List of Figures xvi

List of Algorithms xvii

List of Tables xx

1 Introduction 1

1.1 High-energy experimental physics background . . . 2
1.2 Motivation . . . 6
1.3 Research specification . . . 8
1.4 Thesis outline . . . 9

2 Literature review 12

2.1 Multi-armed bandits . . . 12
2.1.1 Stochastic rewards . . . 13
2.1.2 Adversarial rewards . . . 15
2.2 Pure exploration problems . . . 17
2.3 Combinatorial bandits . . . 21
2.4 Label efficient prediction . . . 24
2.5 Adversarial partial monitoring . . . 25
2.6 Stochastic thresholding bandits . . . 28
2.7 Summary . . . 32

3 Adversarial thresholding semi-bandits 35

3.1 Formal description . . . 36
3.1.1 Bandit terminology . . . 37

3.1.2 ATSBP regret . . . 38
3.1.3 ATSBP outcomes . . . 41
3.2 Hardness of ATSBP . . . 44
3.3 Efficiency ratios . . . 46
3.4 Summary . . . 48

4 Adversarial thresholding bandits with multiple plays 49
4.1 Assumptions . . . 50
4.2 Multi-play arm pulling . . . 51
4.3 Multi-play Exp3 for thresholding bandits . . . 53
4.3.1 Multi-play Exp3 . . . 53
4.4 The T-Exp3.M algorithm . . . 55
4.5 Analysis of T-Exp3.M . . . 55
4.6 Summary . . . 61

5 Label efficient adversarial thresholding semi-bandits 63
5.1 Label efficient thresholding bandits . . . 64
5.1.1 The LET-Exp3.M algorithm . . . 65
5.2 Analysis of LET-Exp3.M . . . 66
5.3 Summary . . . 78

6 Dynamic label efficient adversarial thresholding semi-bandits 80
6.1 Dynamic label efficient prediction . . . 81
6.2 Adaptive k and threshold lists . . . 82
6.3 The dLET-Exp3.M algorithm . . . 83
6.4 Analysis of dLET-Exp3.M . . . 84
6.5 Summary . . . 96

7 AliceBandit 99
7.1 Randomised round robin selection . . . 99
7.1.1 Round robin sub-sequences . . . 100
7.2 The AliceBandit algorithm . . . 104
7.3 Analysis of AliceBandit . . . 106
7.4 Summary . . . 112

8 Online monitoring of ALICE detector controls at the LHC 114

8.1 Transition Radiation Detector . . . 115

8.2 TRD Detector Controls . . . 117

8.3 DCS preparations for RUN3 . . . 118

8.4 Experimental environment . . . 119

8.4.1 TRD conditions and RUN3 simulation data . . . 120

8.4.2 Channels of interest . . . 122

8.5 ATSBP for ALICE . . . 123

8.6 Empirical analysis . . . 124

8.6.1 Fixed threshold results . . . 126

8.6.2 Fixed threshold summary . . . 138

8.6.3 Threshold interval results . . . 139

8.6.4 Threshold interval summary . . . 152

8.7 Summary . . . 154

9 Conclusions 156

9.1 Summary of results . . . 156

9.2 Future directions . . . 161

Bibliography 162

List of Figures

1.1 A Large Ion Collider Experiment. Photo by Antonio Saba - http://cds.cern.ch/record/1436153?ln=it, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31414778. 4

1.2 The adversarial thresholding semi-bandit storyline, evolving from combining elements of non-stochastic bandits with multiple plays, label efficient prediction and extending the stochastic thresholding bandit problem. 10

2.1 Taxonomy of how adversarial thresholding semi-bandits extend the state-of-the-art in multi-armed bandit problems. Dashed lines indicate perceived gaps in knowledge. 34

4.1 A schematic diagram of the thresholding exponentially weighted exploration and exploitation model with multiple plays. 57

5.1 A schematic diagram combining label efficient prediction with adversarial thresholding semi-bandits. 68

6.1 A schematic diagram of the dynamic exponentially weighted exploration and exploitation with multiple plays algorithm. 84

7.1 A schematic diagram of the AliceBandit algorithm. 106

8.1 Schematic cross-section of a TRD chamber, illustrating a pion and electron local track segment. Large energy deposition due to TR photon absorption (indicated by the large red circle in the drift region) is used to distinguish electron tracks. The solid drift lines are calculated algorithmically and relate to the drift voltages set for nominal chamber operation, [4]. 116

8.2 A cross-section schematic of ALICE, perpendicular to the LHC beam. The central barrel detectors are located within a solenoid magnet and provide a magnetic flux density of $B = 0.5\,T$, parallel to the beam direction, [4]. 116

8.3 Dataset of multiple ($K = 44$) TRD Anode Current channels between 12:00:00, 9^{th} June 2018 and 00:00:00, 11^{th} June 2018. Data is binned into two second intervals, resulting in $T = 64800$ rounds. Arm 3 (red line) indicates the tripped channel, with ITRIPLimit (black line). 120

8.4 Simulated dataset of ($K = 10$) Anode Current channels (blue, purple, orange, green, red and black lines) from the ALICE Cosmic Ray detector, for $T = 50000$ rounds. 121

8.5 Average cumulative regret of label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, in the fixed threshold setting, on the ALICE TRD conditions dataset. 126

8.6 Average cumulative pull frequency for **LET-Exp3.M** algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top to bottom). (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 12000, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines. 127

8.7 Average cumulative regret of dynamic label efficient adversarial thresholding semi-bandit algorithms, with subsets of size $k \in \{5, 10, 15, 20\}$, selected each round and maximum observation probability, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right). Experiments on the ALICE TRD conditions dataset, in the fixed threshold setting, were repeated ten times. 130

8.8 Average cumulative pull frequency for **dLET-Exp3.M** algorithms with subsets of $k \in \{5, 10, 15, 20\}$ arms selected each round (top to bottom) and $\epsilon = 1$. (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 1, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines. 131

8.9 Average cumulative regret of **AliceBandit**, with $t_{max} = \{1, 10, 30, 60\}$ minutes. Experiments on the ALICE TRD conditions dataset, in the fixed threshold setting were repeated ten times. 134

8.10 Average cumulative pull frequency for **AliceBandit** with $t_{max} = \{1, 10, 30, 60\}$ minutes (top to bottom). (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 12000, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines. 135

8.11 Average cumulative regret of selected adversarial thresholding semi-bandit algorithms, on the ALICE TRD conditions dataset, in the fixed threshold setting, where experiments were repeated ten times. 138

8.12 Average cumulative reward of label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, in the threshold interval setting, on the ALICE TRD conditions dataset. 140

8.13 Average cumulative pull frequency for label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right), in the threshold interval setting, on the ALICE TRD conditions dataset. Arms $\{3, 4, 27\}$ are highlight by red lines, with remaining arms given by blue lines. 141

8.14 Average cumulative reward of **dLET-Exp3.M**, with subsets of size $k \in \{5, 10, 15, 20\}$, selected each round and maximum observation probability, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right). Experiments on the ALICE TRD conditions dataset, in the threshold interval setting, were repeated ten times. 143

8.15 Average cumulative pull frequency for **dLET-Exp3.M** with $k \in \{5, 10, 15, 20\}$ (top left to bottom right) and $\epsilon = 1$, in the threshold interval setting, on the ALICE TRD conditions dataset. Arms $\{3, 4, 27\}$ are highlight by red lines, with remaining arms given by blue lines. Black lines indicate trip event, recovery and switch off times. 144

8.16 Average cumulative reward of **AliceBandit**, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes. Experiments on the ALICE TRD conditions dataset, in the threshold interval setting were repeated ten times. 146

8.17 Average cumulative pull frequency for **AliceBandit** on the ALICE TRD conditions dataset, with $t_{max} = \{1, 5, 10, 15, 30, 60\}$ minutes (top to bottom) with trip, recovery and switch off indicated (black lines). Arms $\{0, 1, 2, 3, 4\}$ shown by red lines, remaining arms given by blue lines. . . 147

8.18 Average cumulative reward of **AliceBandit**, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes. Experiments on the RUN3 simulated dataset, in the threshold interval setting were repeated ten times. 149

8.19 Average cumulative pull frequency for **AliceBandit** on the RUN3 simulated dataset, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (top left to bottom right). Arms $\{0, 1, 2, 3, 4\}$ shown by red, green, magenta, orange and purple lines, remaining arms given by light blue lines. 150

List of Algorithms

1 UCB1 . . . 15
2 Exp3 . . . 16
3 Successive Rejects . . . 18
4 Successive Halving . . . 20
5 Prediction with partial monitoring . . . 25
6 Anytime parameter-free thresholding bandit . . . 29
7 ATSBP . . . 37
8 DepRound . . . 52
9 T-Exp3.M . . . 56
10 LET-Exp3.M . . . 67
11 dLET-Exp3.M algorithm . . . 85
12 $S3_m$, m^{th} Sub-sequence sizes algorithm . . . 102
13 AliceBandit($\mathcal{I}_{m,\tau(t)}$), Anytime AliceBandit framework . . . 103
14 FTAS, Finite time arm selection algorithm . . . 103
15 AliceBandit . . . 107

List of Tables

8.1 Comparison of average false positive (α), negative (β) and discovery (FDR) rates of label efficient adversarial thresholding semi-bandit algorithms, in the fixed threshold setting, on the ALICE TRD conditions dataset. 129

8.2 Comparison of average false positive (α), negative (β) and discovery (FDR) rates of **dLET-Exp3.M** with maximum observation probabilities $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ and subsets selection sizes $k \in \{5, 10, 15, 20\}$, in the fixed threshold setting, on the ALICE TRD conditions dataset. . . . 133

8.3 Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 10, 30, 60\}$ minutes, in the fixed threshold setting, on the ALICE TRD conditions dataset. 137

8.4 Comparison of average false positive (α), negative (β) and discovery (FDR) rates of label efficient adversarial thresholding semi-bandit algorithms, in the threshold interval setting, on the ALICE TRD conditions dataset. 141

8.5 Comparison of average false positive (α), negative (β) and discovery (FDR) rates of **dLET-Exp3.M** with maximum observation probabilities $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ and subsets selection sizes $k \in \{5, 10, 15, 20\}$, in the threshold interval setting, on the ALICE TRD conditions dataset. . 145

8.6 Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 10, 30, 60\}$ minutes, in the threshold interval setting, on the ALICE TRD conditions dataset. 148

8.7 Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 5, 10, 15, 30, 60\}$ minutes, in the threshold interval setting, on the ALICE RUN3 simulated dataset. 151

9.1 Summary of adversarial thresholding semi-bandit algorithms satisfying Research specifications. The symbols $\star$ and ★ denote partial and full satisfaction of Research specifications 1 and 3. 161

Chapter 1

Introduction

In any given moment, we potentially receive millions of bits of information via our senses, yet our brains are only capable of consciously processing tens of bits of information per second. Situations where we miss an opportunity to influence our survival prospects may have significant implications. Furthermore, in today's data-driven world, we are bombarded with a flood of information that is unpractical to observe in its entirety, if possible at all. When we only want to observe those sources that are of interest, we need to be looking in the right place to make the observation when it happens. We need to shift our focus when the environment changes and not waste effort looking when the environment behaves as expected. The challenge is to know where to look whilst the action of interest is taking place.

Our objective under such circumstances is to observe data streams exclusively for those moments where a certain threshold is exceeded, for all sources that exceed threshold during the period of receiving information. This challenge is made even more difficult if we have no knowledge of when or how often the threshold is breached. Since we only have access to a data stream for the moments we have decided to observe it, we can only base our decisions on the history of its observations. If a data stream has been observed behaving appropriately in the recent past, we can *exploit* this information and choose to look elsewhere or choose to observe a data stream more frequently if it is doing something interesting. If we have not observed a data stream for some time, we can *explore* it to gain knowledge of its current behaviour. In a potentially *adversarial* environment, where no statistical assumptions can be placed on the environmental feed-

back, too much *exploitation* can be taken advantage of by sudden and large deviations about the threshold. Too much *exploration* leads to extended periods of intermittent observation, since we cannot or it too costly to look at everything all the time.

1.1 High-energy experimental physics background

In this section, we outline the origins of the requirements for this research and where upon investigating how to approach this problem, current gaps in knowledge were identified. The reader is introduced to details on the high-energy experimental physics setting that motivates this research and the purpose, characteristics and challenges involved in controlling the experiment in question.

The Large Hadron Colider (LHC) is currently the world's largest and most powerful high-energy physics particle accelerator. The machine was built by the European Organisation for Nuclear Research (CERN) near Geneva, Switzerland. Construction was completed in 2008, in collaboration with thousands of scientists and hundreds of universities world-wide. The term *hadron* refers to subatomic particles composed of several *quarks*, held together by the strong nuclear force, where quarks are one of the fundamental constituents of matter and the strong force is one of the fundamental interactions confining matter together.

The LHC consists of a 27 km ring of superconducting magnets, on average 100 metres below the France-Switzerland border. The LHC is an electromagnetic particle accelerator, where two separate beams of nuclei are accelerated to almost the speed of light in opposite directions. Although travelling at 99.9999991 percent of light speed, the beam does not constitute a continuous stream of particles.

Content removed due to copyright reasons.

A full heavy-ion beam consists of almost 600 bunches, comprised of approximately 7×10^7 lead ions. In proton experiments, almost 3000 bunches can circulate the LHC with 1.2×10^{11} protons in each bunch per second. The terms *collision*, *interaction* and *event* all refer to the crossing of a bunch from each beam.

The LHC is designed to bring both beams together by manipulating the surrounding magnetic field at four separate locations, known as nominal interaction points. The main experiments conducted at the four interaction point are called ATLAS (A Toroidal LHC ApparatuS), ALICE (A Large Ion Collider Experiment), LHCb (Large Hadron Collider beauty) and CMS (Compact Muon Solenoid), with each experiment having different physics objectives. Most of the data collected at the LHC involves the by-products from colliding protons together (p-p collisions). However, one month each year is allocated for the study of heavy ion collisions, involving Lead nuclei (p-Pb and Pb-Pb collisions). Due to the relative size of lead ions, Pb-Pb collisions are 100 times more energetic than those colliding protons, generating temperatures up to $100,000$ times hotter than the temperature in the core of the Sun.

Trigger systems are commonly used in high-energy particle accelerators to decide which events are of physics interest and worth instructing the Data Acquisition (DAQ) system to collect data from the experiment and send to storage for further analysis. To date, there have been two extended periods of data-taking at the LHC, referred to as RUN1 (2009-2013) and RUN2 (2015-2018), respectively. In both runs, the physics data is collected after a triggering event has been detected. The data is then combined with the experiment condition parameters for validated and further analysis.

A Large Ion Collider Experiment (ALICE) [3] is specifically designed to conduct heavy-ion experiments at the LHC (see Figure 1.1). In high energy nucleus-nucleus collisions, a high-density de-confined state of strongly interacting matter, called quark-gluon plasma (QGP) is thought to be created. ALICE's objective is to study the physics of QGP. The statistical theory of thermodynamics for matter at extreme energy densities is called Quantum Chromodynamics (QCD) and predicts that at sufficiently high densities, hadronic matter will transition to QGP. In principle, the reverse of this transition can model the early universe, some 10^{-5} s after the Big Bang and is also thought to describe the core of collapsing neutron stars.

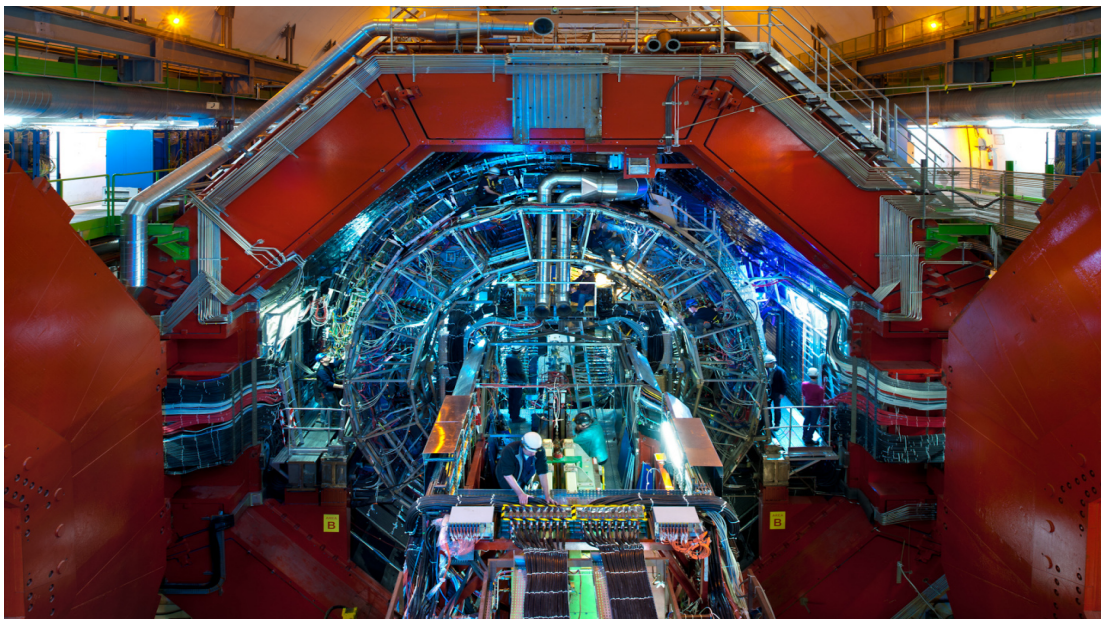


Figure 1.1: A Large Ion Collider Experiment. Photo by Antonio Saba - http://cds.cern.ch/record/1436153?ln=it, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31414778.

Several layers of detectors surrounding the central barrel of the two particle beams are contained within a solenoid magnet, which generates a magnetic field along the direction of the particle beam. It is worth noting here that while ALICE is referred to by the term *detector*, the experiment is a collection of many detectors. The Inner Tracking System (ITS), found closest to the nominal interaction point is used for tracking particle with low momentum, particle identification (PID) and locating the primary and any secondary vertices. The Time Projection Chamber (TPC) surrounds the ITS and is used for further tracking and PID. The Transition Radiation Detector (see Section 8.1) surrounds the TPC, which itself is surrounded by the Time-Of-Flight detector (TOF). The TOF is used for identifying hadrons. The ElectroMagnetic Calorimeter (EMCal), PHOton Spectromenter (PHOS) and the High Momentum Particle Identification Detector (HMPID) are used for identifying electrons, jets (beams of collimated particles) and hadrons.

ALICE is currently undergoing significant upgrades and will start collecting data again in 2021, following *Long Shutdown* 2 (LS2, 2018-2021). The upgraded detector will be required to read out and inspect Pb-Pb collisions at a rate of 50 kHz and sampling p-p and p-Pb collisions at up to 200 kHz. The ALICE Online-Offline (O^2) computing project will enable data-taking at up to 1.1 TB s^{-1} and storage of approximately 60 PB

per year. For the first time, the detector will need the continuous flow of approximately $100,000$ condition parameters to validate collision event reconstruction. Due to the higher collision rates, the detector data cannot be stored on disk and then validated at the end of data-taking run. The new continuous data-taking regime hence requires a continuous stream of condition parameters for validation at approximately 1 Hz.

Since it is computationally costly to monitor every condition parameter continuously over the entire period that the detector is collecting data, CERN requires a way to improve monitoring efficiency that can learn to prioritise the observing of parameters that are not behaving as expected in the event that intervening actions need to be taken, learn to disregard parameters that are behaving as expected and not forget about any particular parameter completely.

The safety of detector components is ensured by continuously maintaining nominal operating conditions using overlapping layers of alerts and interlocks. Independent mechanisms, particularly *trip* protocols are used to preserve the integrity, reliability and functionality of a device. In the event that the operating parameters of such a device exceed a given threshold, a tiered alarm system is tripped, advising a detector operator (*shifter*) to potentially shut that device down. The following examples highlight how such circumstances may arise:

- During critical beam injection (where particles are pushed into the 27 km LHC ring) or adjusting beam optics poses a potential hazard to any gas-filled detector. High voltage power supply *channels* are tuned to the LHC beam status. Upon injection, the Anode voltages are *ramped-down* to reduce gain in the gas chamber of whichever device is gas-filled. Voltages are only *ramped-up* to nominal operating levels after stable beam conditions have been declared by LHC operators.

- If the Anode current builds up without supervision and control, detector components can be severely damaged. To prevent this, the channel is tripped at a set threshold and automatically switched off.

Recovery of a tripped channel may take several hours of maintenance. In the event of a trip, procedure dictates that the detector expert be contacted, and depending on the component affected, this may require the expert to by physically on-site. For an international organisation such as CERN, this poses its own challenges. Since any single channel is connected to many other components due to the distribution of power supplies, large parts of the detector may need to be excluded from the experimental controls. A tripped power supply potentially requires the reset of an entire module of the experiment, which renders it inoperable during data-taking.

During RUN3 (2021-2024) and beyond, the experiment will collect data continuously. If a detector experiences a channel trip, all other detectors will continue taking data. The exclusion of a detector due to a tripped channel will have a significant impact on data quality, particularly if the trip concerns a critical region of the experiment. In RUN3, upgraded components will result in an increase in data-taking by a factor of 100. The severity of each trip will thus be 100 times more serious than before LS2. Predictive control such as power-cycling or ramping-down of voltage channels could be implemented if pre-trip behaviour can be detected in the stream of DCS channels monitored.

1.2 Motivation

This work originates from being presented with the problem of continuously monitoring multiple streams of detector conditions parameters, because of the paradigm shift in how the LHC will collect data after the commencement of RUN3.

One of the main challenges posed by the ALICE DCS team is predicting if a device parameter will exceed a given threshold (upon which the device channel will potentially trip). Since there is an inevitable cost associated with observing every channel (physically by the shifter or computationally), the challenge is made more complicated by the desire to observe channels as efficiently as possible. A shifter only wants to observe a channel when it is exceeding threshold and ignore it for the remainder. However, due to the non-stochastic nature of the experimental conditions, no channel can be ignored completely. This is an example of the *exploration-exploitation dilemma*, [9].

The most common expression of sequential decision-making for balancing exploration and exploitation is the *multi-armed bandit* (MAB), [65], [49] and [9]. The classical MAB is analogous to a casino gambler, playing a set of K one-armed slot machines (colloquially known as *bandits*). The game proceeds in *rounds* and the player *pulls* one arm (*bandit setting*) each round and receives a *reward* associated with the payoff probability of the arm pulled. The game is in the *full information* setting if the player pulls arms and then receives information on the payoff of every arm each round. The *semi-bandit* setting applies to games where any number of arms can be pulled each round and the player only observes feedback from those arms pulled. The objective of the player is to maximise their cumulative reward over the number of rounds they play the game.

Since Locatelli *et al.* [55], MAB variants have existed in the literature designed for identifying which bandit *arms* are expected to exceed a known threshold, after a given number of rounds played. However to the best of our knowledge, these *thresholding bandits* only apply to stochastic feedback, obeying the condition that the realisation of each arm is identically and independently distributed (iid).

Given a fixed number of channels, feedback from each channel arrives sequentially. In each time-step, feedback associated with all channels is generated. A player in this game, must decide in every time-step which of the channels will return a value exceeding some threshold and which will not. The player wants to exploit channels that have recently exceeded threshold but also explore channels that may not have been inspected for some time and could be unexpectedly exceeding threshold.

To satisfy the requirements of the ALICE DCS team and address an apparent gap in knowledge, this work investigates the thresholding bandit problem in an adversarial (non-stochastic) setting, where feedback cannot be assumed to obey the iid condition.

As mentioned previously, critical to the application is balancing the desire for observing feedback that has been observed exceeding threshold in recent history with the need to observe feedback from a channel that has not been inspected for a while and may have begun exceeding threshold without warning. Current thresholding bandit algorithms follow a pure exploration regime (see Chapter 2), where a determination of which arms exceed a threshold in expectation, after a give number of observations. In the

adversarial thresholding problem, this determination needs to be made every time arms are pulled, without relying on the iid condition. The objective of a player in this game is to observe all of the arms exceeding threshold in every round of the game, which could potentially be none, some or all of the channels. This has led to the development of algorithms in this thesis for *adversarial thresholding semi-bandits problems* (ATSBP).

1.3 Research specification

The objective of this thesis is to design algorithms that sequentially decide which arms to pull, efficiently pulling arms only when their associated feedback exceeds a given threshold and minimise their cumulative regret compared to some optimal pulling policy when playing against an oblivious adversary. The specific requirements for such algorithms incorporate both *empirical* and *theoretical* performance. Empirical considerations mainly focus on the effectiveness of an algorithm to solve a specific real-world problem and are usually measured by averaging the outcomes from repeated experiments, particularly when the algorithm possesses inherent randomness. Theoretical performance guarantees require that algorithms have regret that is provably sub-linear in the total number of rounds, T. That is, the longer the game is played, the more the player converges towards playing optimally.

In many applications, the total number of arms K, and or total number of rounds played may be large. As the size of the problem increases then so will the computation required. In terms of application of the problem, algorithms with low computational cost are always essential and an additional request of the ALICE DCS team was to detect deviations in channel parameters for a tunable deviation threshold.

- **Theoretical guarantees (specification 1):** As the number of rounds that the game is played for increases, the algorithms developed are required to make decisions that converge towards the optimal solution, pulling arms when their feedback is likely to exceed a threshold, learn to ignore arms when their feedback is within threshold without forgetting about them completely due to the potentially dynamic nature of the environment. This implies pulling policies with sub-linear regret.

- **Empirical performance (specification 2):** Real-world problems are typically motivated by the quality of decisions made on average and thus require low regret on as many rounds as possible. Reliably making high-quality decisions is thus essential for optimal performance.

- **Computational complexity (specification 3):** As well as empirical performance, real-world applications require that decisions are made in a timely manner and the computational resources available are usually constrained. Consequently, algorithms developed for such problems must have low computational complexity.

- **Deviation detection (specification 4):** Fixed thresholds have been used throughout the lifetime of the LHC, for preventing channels outside of nominal operating parameters becoming damaged and inoperable. Algorithms that can detect significant deviations in operating parameters, both efficiently and promptly, will provide valuable information on the behaviour and interaction of prototype components.

Industrial applications would typically give more weighting towards Research specifications 2 and 3 over Research specification 1. However, the real-world application introduced in Section 1.1, values Research specification 1 almost as highly as Research specification 2 or 3, with Research specification 4 providing an additional source of valuable research. In fact, the transparency of methods used for sequential decision-making problems is currently a contentious topic within the both the computer science community and society as a whole. Several recent works by Joseph *et al.* [42], Lepri *et al.* [53] and Claure *et al.* [30] highlight the significant role that multi-armed bandits are playing within this debate. Ultimately, the aim in this thesis is to develop algorithms that optimise all of the specifications set out in this section.

1.4 Thesis outline

In this section, we set out the structure of the remainder of this thesis. We also describe the journey our research took to reach this point. In Figure 1.2, we illustrate how combining elements of adversarial multi-armed bandits with multiple plays and label efficient prediction (discussed in Section 2.4), lead us to the development of label efficient adversarial thresholding bandits with multiple plays and introduces threshold-

ing bandit to problems with non-stochastic reward sequences. Limitations inherent in adversarial bandit with multiple plays then lead to the development of dynamic label efficient thresholding bandits, where we do not need to know in advance how many players are playing. Elements of dynamic label efficient thresholding bandits were further developed to improve algorithmic complexity.

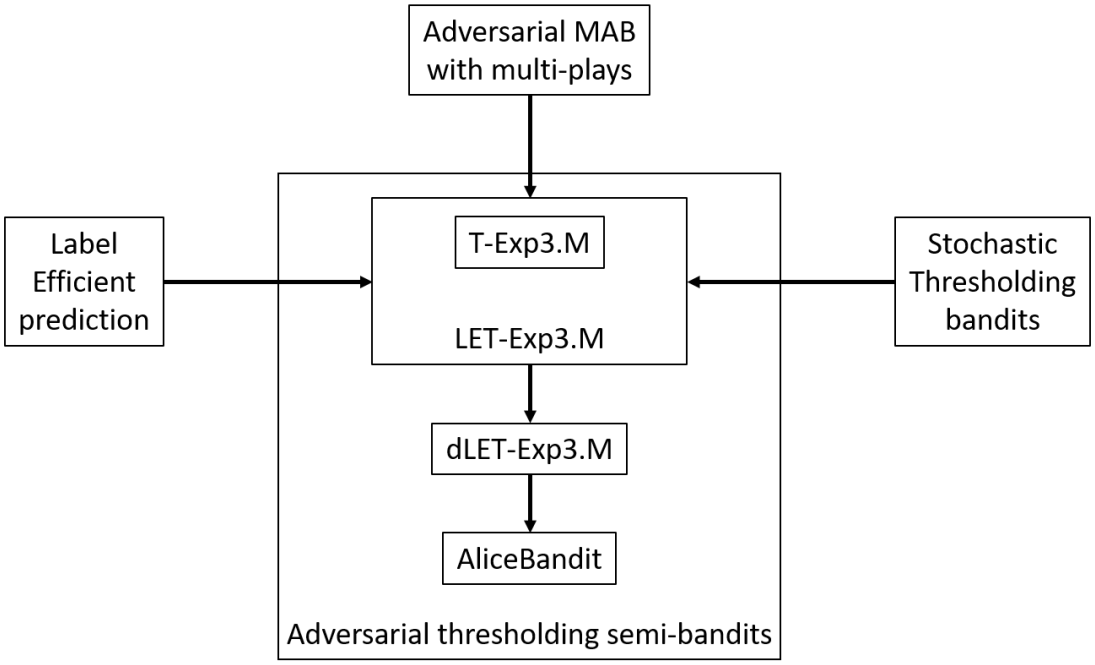


Figure 1.2: The adversarial thresholding semi-bandit storyline, evolving from combining elements of non-stochastic bandits with multiple plays, label efficient prediction and extending the stochastic thresholding bandit problem.

In Chapter 2, we review the state-of-the-art in multi-armed bandits. We discuss aspects of stochastic and adversarial bandits, regret minimisation and pure exploration bandits. In particular, combinatorial pure exploration bandits are identified to introduce the current thinking on thresholding bandits. Generalisations of the bandit problem are discussed and we focus on several variants of stochastic thresholding bandit algorithms, specifically to highlight the novelty of adversarial thresholding semi-bandits. Chapter 3 formally introduces adversarial thresholding semi-bandits and complexity of the problem. Chapter 4 introduces the first algorithm for ATSBP. A regret analysis is performed under certain assumptions and we prove an upper bound on theoretical performance. In Chapter 5, we extend ATSBP to scenarios where the total number of observations allowed over the course of the game is controlled. Theoretical performance guarantees are proven in terms of the number of allowed observations with high prob-

ability. In Chapter 6, we present an algorithm that dynamically adapts the number of allowed observations, depending on the feedback it receives. Slightly worse performance guarantees are provided with high probability. Our final algorithm is presented in Chapter 7. Performance guarantees are given in terms of the expected number of mis-classifications of each arm, over the entire game. Computational complexity is also discussed in all four chapters. In Chapter 8, we perform an empirical analysis on each algorithm presented in Chapters 4, 5, 6 and 7. We describe the origins of the dataset and relevant pre-processing required. We discuss the results of our analysis and how they relate to our Research specification. Finally, Chapter 9 concludes our work and we discuss potential future directions for the adversarial thresholding semi-bandit problem.

Chapter 2

Literature review

In this chapter, we give an overview of the current literature relevant to our research. The reader is briefly introduced to some of the significant works on the MAB problem in Section 2.1, with an emphasis on balancing exploration and exploitation. We review previous work on classical stochastic bandits in Section 2.1.1 and in Section 2.1.2 we discuss non-stochastic versions, also known as adversarial bandits. Section 2.2 provides a discussion of existing work on bandits that focus only on exploration, mainly pertaining to stochastic rewards with non-stationary means and those with adversarial rewards. This leads into work on problems with a combinatorial structure to arms and the semi-bandit setting in Section 2.3. Section 2.4 discusses some relevant literature on the problem of *label efficient prediction*. Section 2.5 gives an overview of *adversarial partial monitoring* problems and how they relate to the adversarial thresholding semi-bandit problem. In Section 2.6, we present an overview of the state-of-the-art in thresholding bandit problems. We conclude by giving a summary of how our work extends the field in terms of the details given in this chapter.

2.1 Multi-armed bandits

Multi-armed bandit problems were conceived almost one hundred years ago by Thompson in his seminal work [69], whilst investigating how to improve medical trials where a general practitioner can choose from a range of treatments and patients arrive sequentially. Thompson's research studied how to make trials more adaptable as new information was made available. Although medical research has been slow to apply MABs to

their studies, many other applications have been discovered where multi-armed bandits are a vital tool. Within this section, we discuss the problem of reward maximisation (regret minimisation) by balancing the exploration-exploitation dilemma. In Section 2.1.1, we review some of the influential work, fundamental to multi-armed bandits with stochastic rewards. In Section 2.1.2, we discuss the non-stochastic MAB version and highlight the work from which all later non-stochastic bandits can be derived.

There are three branches of study in multi-armed bandits, with each paradigm based on the constraints applied to how rewards are generated. *Stochastic* MABs have rewards that are generated by independently sampling from a probability distribution associated with each arm. *Adversarial* bandits are concerned with problems where no statistical assumptions can be made and rewards can be generated by some potentially competitive environment. An *oblivious* adversary in this non-stochastic variant can choose a sequence of rewards regardless of which arms a player decides to pull. An *adaptive* adversary can give out rewards based on which arms the player has previously pulled. The third branch of bandit problems are related to Markovian processes, where each arm is associated with a distinct state space. In this thesis, we are not concerned with Markovian bandits. However, the interested reader is directed towards [34] in particular and [50] for a more general exposition. Our interest in stochastic bandits is solely to arm the reader with sufficient knowledge to introduce the concept of thresholding bandit problems.

2.1.1 Stochastic rewards

In this section, we present the idea of multi-armed bandits with stochastic rewards. This forms the foundation of the subject and we reference the concepts required to detail a specific algorithm.

Although the design of online experiments had been previously studied, the notion of balancing exploration and exploitation was introduced by Robbins, [65]. Lai and Robbins [49] were the first to show that the regret of multi-armed bandits must asymptotically grow at least logarithmically with the number of rounds played. Although the idea was introduced in [49], the **UCB** index-based policy was independently developed by Katehakis and Robbins in [44] and Agrawal in [1]. The mathematical model for the stochastic MAB over a finite number of rounds was established by Auer *et al.* [9]

and the entire body of literature since has elaborated on their framework. For the remainder of this section, we will briefly review this model.

The classical MAB problem is defined over a set of K arms, $\{i \in [K] \,|\, [K] = 1, \ldots, K\}$ and a sequence of rounds $t \in [T]$. When a player pulls arm i in round t, they receive a reward $X_{t,i}$ which is a random variable associated with an unknown distribution for arm i. The rewards for each arm are assumed to be independent and identically distributed, where ν_i represents the distribution of arm i. The mean reward for arm i is given by $\mu_i = \mathbb{E}_{i\sim\nu_i}\big[X_{t,i}\big]$, where the expectation is taken over the randomness in the sequence of previous draws.

The mechanism (or algorithm) that a player uses to decide which arm to pull next is based on the rewards the player has received in previous rounds. The cumulative regret a player expects to experiences after playing T rounds is given by the cumulative difference between the playing the reward expected to be optimal every round and the expected reward for the arms pulled during those T rounds, multiplied by the number of times those arms are pulled,

$$\mathbb{E}\big[\mathcal{R}_T\big] = \mu^* T - \sum_{i=1}^{K} \mu_i \mathbb{E}\big[N_{T,i}\big],$$

where $\mu^* = \max_{i\in[K]} \mu_i$, $N_{t,i} = \sum_{s=1}^{t} \mathbb{I}\{i = I_s\}$ and I_s denotes the realisation of arm i in round s. Without loss of generality, the unknown mean rewards for each arm can be arranged such that, $\mu_1 \leq, \ldots, \leq \mu_K$, where $\mu^* = \mu_1$. This allows for the definition of the sub-optimality gap, $\Delta_i := \mu^* - \mu_i$.

The main contribution from [9] was the introduction of a deterministic algorithm that could operate in finite time, known as **UCB1** and relies on the principle of *optimistic uncertainty*. This concept is based on two possible outcomes each round, either the best arm was pulled or not. Intuitively, the player wants to pull the arm that maximises the estimated mean reward as much as possible. The more an arm is pulled, the more confidence the player has in the accuracy of their reward estimate. The player optimistically assumes that the true mean reward for each arm is as large as possible based on the evidence they have. An upper confidence bound is maintained for each arm and every time arm i is pulled, the player gains more confidence on their empirical estimate of the mean for arm i. The arm which has the largest upper confidence

Algorithm 1 UCB1

Input: number of arms K, number of rounds T, distributions $\nu_1, \ldots, \nu_K$, arm pull counter $N_{0,i} = 0$.

1: **for** $t = 1, \ldots, K$ **do**
2: pull arm $i = t$
3: receive reward $x_{t,i} \sim \nu_i$
4: let $\hat{x}_{t,i} = x_{t,i}$
5: let $N_{t,i} = 1$
6: **for** $t = K + 1, \ldots, T$ **do**
7: pull arm I_t such that,

$$I_t = \text{argmax}_{i \in [K]} \, \hat{x}_{t,i} + \sqrt{\frac{2 \log(T)}{N_{t,i}}}$$

8: update $\hat{x}_{t,I_t}$
9: let $N_{t,I_t} = N_{t,I_t} + 1$

bound is pulled each round (Algorithm 1). Auer *et al.* [9] elegantly prove that **UCB1** is optimal when compared with the lower bound stated in [49] and robust to mild assumptions on the rewards. We omit a discussion on tuning the confidence level δ and how it is approximated in [9] due to its irrelevance to the theme of this work. However, further details can found in [50] and [19]. The success and power of this approach is testament to the number of variants that bare the UCB label. A quick search for 'UCB multi-armed bandit' on Google scholar returned almost eight hundred results for 2019 alone.

2.1.2 Adversarial rewards

In this section we dispense with the iid assumption of stochastic bandits as it is violated in almost all real-world applications. This setting takes its name from the potentially worst-case scenario where rewards are generated by some adversary that may have knowledge of the pulling policy used by the player. The adversary can be *oblivious* to the actions of the player and the sequence of rewards can be considered to have been chosen before the player make their first move. The player is thus searching for a pulling policy that coincides with the optimal sequence. In the event that the adversary adapts to the actions of the player, an *adaptive* adversary can react to whatever move the player makes. This time the player is searching for a dynamic policy that counteract the move of a non-oblivious environment.

Algorithm 2 Exp3

Input: $K, T \geq 2$, $\eta > 0$, $\gamma \in [0,1]$, $w_{1,i} = 1$ for $i = 1, \ldots, K$.

1: **for** $t = 1, \ldots, T$ **do**

2: Set,

$$p_{t,i} = (1-\gamma)\frac{w_{t,i}}{\sum_{j=1}^{K} w_{t,j}} + \frac{\gamma}{K} \quad \forall\, i = 1, \ldots, K$$

3: draw I_t randomly according to $p_{t,1}, \ldots, p_{t,K}$

4: receive reward $x_{t,i} \in [0,1]$

5: **for** $j = 1, \ldots, K$ **do**

$$\tilde{x}_{t,j} = \begin{cases} \frac{x_{t,j}}{p_{t,j}} & j = I_t \\ 0 & \text{otherwise} \end{cases}$$

6: Update weights,

$$w_{t+1,j} = w_{t,j} \exp\Big(\frac{\gamma}{K}\tilde{x}_{t,j}\Big)$$

Auer *et al.* [8] were the first to give a lower bound of $\Omega(\sqrt{KT})$ for the adversarial bandit problem. It was pointed out in [8] that the only way for a player to compete with an adversary on this problem is play randomly and proposed the *Exponential weight algorithm for Exploration and Exploitation*, **Exp3** (Algorithm 2). The algorithm uses a weighted probability distribution over the set of arms. An *egalitarian* parameter γ, is used to vary the degree of randomness in the choice of arm each round. The weight of each arm is exponentially updated based on an unbiased estimate of the observed reward, from dividing the actual reward observed by the probability of choosing that arm. The estimate compensates for the rewards of arms that are not likely to be pulled. An upper bound on the expected regret of **Exp3** that matches the lower bound was proved in [8], with the structure of the proof becoming the foundation for almost all future theoretical guarantees that rely on exponential weights.

The main purpose of this section was to introduce the general concept of the multi-armed bandit and highlight the difference between the two main approaches to **UCB**-type algorithms for arms with stochastic rewards and **Exp3**-type algorithms with non-stochastic (adversarial) rewards. Since our research is aimed towards an application involving the measurement of physical phenomenon, several of the algorithms developed in this thesis are extensions of **Exp3**.

2.2 Pure exploration problems

The previous section primarily addresses the objective of reward maximisation, which is equivalent to regret minimisation. In this section, we review the literature where the player incurs no exploration cost and simply wants to identify the optimal arm under some criteria. This naturally leads to two options: (a) *fixed confidence*, where the player wants to have a given level of confidence δ, that their choice of arm is optimal by pulling as few arms as possible or (b) *fixed budget*, where the player has a fixed number of arm pulls available and wants to make the best choice possible after all pulls have been exhausted.

One of the first works on pure exploration bandits in the machine learning community was given by Even-Dar *et al.* [31], in 2002. The classical MAB problem was analysed in the fixed budget setting using the *Probably-Approximately Correct* (PAC) framework, introduced by Valiant in [73]. In 2006, Even-Dar *et al.* [32], propose that the goal of the player is to output an arm that is within ϵ of the optimal arm with probability of at least $1 - \delta$, referred to as (ϵ, δ)-PAC. The algorithms proposed in [32] rely on the statistical characteristics of the distributions associated with each arm to stop the exploration of arms found to have expected rewards significantly smaller than all other remaining arms. This gave rise to a family of *elimination* algorithms, where arms are sequentially eliminated from the exploration process until only one arm remains or all pulls have been completed. In the latter case, the player recommends the arm that maximises their empirical mean. In this setting, *sample complexity* is used to measure performance, which represents the number of times the player needs to pull arms before achieving the (ϵ, δ)-PAC criterion. A lower bound on the expected sample complexity was given by Mannor and Tsitsiklis, [56] in 2004 and the *worst trajectory* upper bound algorithms given in [31] and [32] match the lower bound. The measure of *simple regret* was introduced by Bubeck *et al.* [17] and [18],

$$r_T := \mu^* - \mathbb{E}\big[\mu_{J_T}\big],$$

where J_T is the recommended arm by the player after T rounds. It can be seen that the simple regret represents the expected error of the player after T rounds. Bubeck *et al.* prove in [17] that minimsing simple regret and minimising pseudo regret can require conflicting approaches and improving one worsens the other, in most cases.

However, they also note in discussion that certain examples allow for the minimisation of both simultaneously. When playing against an oblivious adversary, pseudo-regret and expected regret and equivalent. As such, we will simply use the term *regret* for the sequel.

Motivated by the channel allocation for mobile phone communications, Audibert *et al.* [7] study the problem of *best arm identification* (BAI). An algorithm with a highly exploring policy is presented that is based on Algorithm 1 from [9], called *Upper Confidence Bound Exploration* (**UCB-E**). They prove that once appropriately tuned, **UCB-E** has an upper bound on the probability of error after T rounds,

$$\mathbb{P}\big[J_T \neq i^*\big] = \mathcal{O}\left(\exp\left(-\frac{T}{H_1}\right)\right), \tag{2.1}$$

where i^* denotes the optimal arm and,

$$H_1 = \sum_{i=1}^{K} \Delta_i^{-2}, \tag{2.2}$$

characterises the hardness of the problem.

Since **UCB-E** requires the knowledge of H_1 to achieve the bound in (2.1), Audibert *et al.* propose the *Successive Rejects* (**SR**) algorithm that proceeds in phases and relates to the action elimination approaches from [32] by rejecting the arm at the end of each phase with the smallest empirical mean reward (see Algorithm 3).

The notation used in Algorithm 3 is taken directly from [7] and does not reflect the notation used throughout this thesis. The **SR** algorithm is parameter-free and has a

Algorithm 3 Successive Rejects

Input: Let $A_1 = \{1, \dots, K\}$, $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$, $n_0 = 0$, for $k \in \{1, \dots, K-1\}$,

$$n_k = \left\lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{K+1-k} \right\rceil.$$

1: **for** $k = 1, \dots, K-1$ **do**
2: For each $i \in A_k$, select arm i for $n_k - n_{k-1}$ rounds
3: Let $A_{k+1} = A_k \setminus \operatorname{argmin}_{i \in A_k} \hat{X}_{i,n_k}$ (we only remove one element from A_k, if there is a tie, select randomly the arm to dismiss among the worst arms).

Output: Let J_n be the unique element of A_K.

probability of error such that,

$$\mathbb{P}\big[J_T \neq i^*\big] = \mathcal{O}\left(\exp\left(-\frac{T}{\log(2K)H_2}\right)\right),$$

where,

$$H_2 = \max_{i\in[K]} i\Delta_i^{-2}. \tag{2.3}$$

Since $H_2 \leq H_1 \leq H_2 \log(2K)$, SR performs as well as **UCB-E** up to the logarithmic term in K without needing to know H_1.

In 2012, Bubeck *et al.* [20] studied the problem of identifying the top-m arms in a multi-armed bandit that extends the **SR** algorithm, called *Successive Accepts and Rejects* (**SAR**). The problem of generalising the BAI problem to identifying the top-m distributions was discussed as an open problem in [7]. The same arguments were applied to the related problem of multi-bandit best arm identification. Empirical evidence in [20] shows that Successive Rejects performs poorly when extended to $m > 1$ compared with **SAR**, suggesting that there are fundamental differences in how to solve the top-m and single BAI problem. Due to these subtle differences, previous definitions of the sub-optimality gap and associated complexity measures required modification,

$$\Delta_i^{\langle m\rangle} = \begin{cases} \mu_i - \mu_{m+1} & \text{if } i \leq m \\ \mu_m - \mu_i & \text{if } i > m \end{cases} \tag{2.4}$$

$$H_1^{\langle m\rangle} = \sum_{i=1}^{K} \frac{1}{\big(\Delta_i^{\langle m\rangle}\big)^2}, \tag{2.5}$$

$$H_2^{\langle m\rangle} = \max_{i\in[K]} i\big(\Delta_{(i)}^{\langle m\rangle}\big)^{-2}, \tag{2.6}$$

where $(i) \in \{1,\ldots,K\}$ is defined such that $\Delta_{(1)}^{\langle m\rangle} \leq \ldots \leq \Delta_{(K)}^{\langle m\rangle}$. The probability of error in the top-m problem is referred to as the probability of misidentification and given by,

$$\mathbb{P}\big[J_1,\ldots,J_m \neq \{1,\ldots,m\}\big] = \mathcal{O}\left(\exp\left(-\frac{T}{\overline{\log}(K)H_2^{\langle m\rangle}}\right)\right),$$

where $J_1,\ldots,J_m$ are the top-m recommended arms and $\{1,\ldots,m\}$ denotes the m arms with the largest means. The sub-optimality gap in (2.4) defines how far the mean of

arm i is from the best arm not in the top m, if arm i is also not in the top m. If arm i is in the top m, then (2.4) defines the distance between the mean of arm i and the mean of the worst arm in the top m. As such, the gap is smaller for arms that are closer to the boundary between those in the top m and those that are not. The hardness measures (2.5) and (2.6) merely reflect TBP complexity and (2.3) from the single BAI problem.

So far, the approaches in this section relate to stochastic rewards. In 2015 Jamieson and Talwalkar [41], presented an algorithm called *Successive Halving* (**SH**) for non-stochastic BAI. Motivated by the problem of hyperparameter optimisation in machine learning, no assumptions were placed on the sequence of arm losses. Each arm corresponded to a different hyperparameter configuration of a machine learning model and the objective is to assign more compute resource to more promising configurations. Under the assumption that the losses for every arm converge, Jamieson and Talwalkar made two observations. (i) No arm can be rejected as optimal with certainty and (ii) no arm can be verified as optimal or even within ϵ of the optimal arm. Hence their approach is for fixed budget best arm identification. The **SH** algorithm proceeds by pulling every arm for a number of times, depending the number of arms available and a budget B. The losses from each arm are then ranked and the lowest half are rejected. The algorithm terminates when a single arm remains (see Algorithm 4).

Algorithm 4 Successive Halving

Input: Budget $B > 0$, K arms, k^{th} loss $\ell_{i,k}$, for arms $i = 1, \dots, K$ and $S_0 = \{1, \dots, K\}$

1: **for** $k = 0, \dots, \lceil \log_2(K) \rceil - 1$ **do**

2: Pull all arms $i \in S_k$ for,

$$r_k = \left\lfloor \frac{B}{|S_k| \lceil \log_2(K) \rceil} \right\rfloor,$$

times. Set $R_k = \sum_{j=0}^{k} r_j$.

3: Let σ_k be a bijection on S_k such that,

$$\ell_{\sigma_k(1),R_k} \leq \cdots \leq \ell_{\sigma_k(|S_k|),R_k}.$$

4: Set,

$$S_{k+1} = \left\{ i \in S_k \,|\, \ell_{\sigma_k(i),R_k} \leq \ell_{\sigma_k(\lfloor |S_k|/2 \rfloor),R_k} \right\}.$$

Output: Singleton element, $S_{\lceil \log_2(K) \rceil}$.

Theoretical guarantees on the performance of **SH** are given in terms of the number of rounds and the inverse envelope function for each arm, defined as $\gamma^{-1}(t)$ for time-step t (see [41] for further details).

The work and corresponding algorithms reviewed in this section are all concerned with finding the best arm (or arms) and almost exclusively for the stochastic reward case. In each of the algorithms described, every arm is pulled for a set number of times each round and arms are either accepted or rejected depending on the empirical mean of the reward from each arm. However, the player makes no decision on which arms to pull, pulling every arm that has not been elimination yet.

2.3 Combinatorial bandits

In this section, we review the literature on bandits with a combinatorial element to arms. The body of work mostly relates to stochastic pure exploration bandits, but we do identify work relating to regret minimisation with adversarial rewards that forms the basis of work in future chapters of this thesis.

In 2012, Cesa-Bianchi and Lugosi [25] introduce the sequential prediction problem where a finite class of experts has a certain combinatorial structure and demonstrate the existence of non-trivial algorithms that are computationally efficient. Sequential prediction is a variant of online linear optimisation, where a player plays a repeated game against an adversarial opponent. The opponent generates a loss that is a linear function of the experts but only reveals the loss of the chosen expert to the player each round. The **ComBand** algorithm proposed in [25] is based on maintaining exponential weights for each expert and is proven to be optimal. Combinatorial bandits are a special case of online linear optimisation where the combinatorial set of arms (experts) is a subset of the binary hypercube $\{0,1\}^K$. Several applications are suggested, such as the sequential generation of random spanning trees for network communications.

In 2013, Chen *et al.* [26] define a framework for the stochastic *combinatorial* MAB problem. Motivated by applications where there is a combinatorial structure between subsets of arms, an extension of **UCB1** is presented, called *Combinatorial Upper Confidence Bound* (**CUCB**). Optimal regret bounds are proven that coincide with bounds for the classical MAB problem up to constant factors. The player proceeds by observ-

ing a collection of arms each round, known as a *super arm* and has a direct association with problems in the semi-bandit setting. Also in 2013, Neu and Bartòk [61] consider online combinatorial optimisation in the non-oblivious adversarial semi-bandit setting. The goal of the player is minimise their cumulative loss after choosing up to $m < K$ arms from a combinatorial decision set. An efficient learning algorithm is proposed that comprises the classical *Follow the Perturbed Leader* (**FPL**) algorithm from Hannan [37] and a novel loss estimation procedure called *Geometric Resampling* (**GR**). The loss estimate is unbiased in the same respect as in [9] but is also dependent on a geometrically distributed random variable after repeatedly sampling from the probability of drawing the subset of arms. The total expected regret of their **FPL** with **GR** algorithm is of order $m\sqrt{KT \log K}$.

In 2014, Chen *et al.* [27] introduce *combinatorial pure exploration* (CPE) to multi-armed bandits. The *Combinatorial Lower-Upper Confidence Bound* (**CLUCB**) algorithm is presented in the fixed confidence regime. A sample complexity lower bound for the CPE problem is given and upper bounds are derived for *Combinatorial Successive Accepts and Rejects* (**CSAR**), a parameter-free algorithm in the fixed budget setting. **CSAR** generalises [20] where the combinatorial set is the set of arms. In 2015, Neu [62] studies problems where the player chooses from a combinatorial decision set of and the players' objective is to minimise their overall loss. First-order bounds are presented that depend on the quadratic variation in adversarial loss sequences. Their approach combines the **FPL** algorithm from [37] and a loss estimate using an implicit exploration parameter. More recently, Wang and Chen [75] in 2018 use Thompson Sampling [69] for the stochastic combinatorial semi-bandit problem. They derive distribution-dependent regret bounds of order $\mathcal{O}(K\Delta_{min}^{-1} \log(T))$ where Δ_{min} is the smallest non-zero sub-optimality gap.

MABs with *multiple plays* are equivalent to semi-bandit problems when m distinct arms are pulled by the player each round. In 2010, Uchiya, Nakamura and Kudo [72] extend the work of Auer, Cesa-Bianchi, Freund and Schapire [10] on non-stochastic multi-armed bandit problems to the $k \geq 1$ multi-play setting. The multiple play version of **Exp3** from [10] was proposed in [72] called **Exp3.M**. As inferred by the name, the basic concept of **Exp3.M** uses an exponentially weighted distribution to balance exploration and exploitation. The problem of choosing k distinct arms according to

this distribution is satisfied by using the **DepRound** function [33], which uniformly selects k arms with probability $p_{t,i}$ provided that $\sum_{i\in[K]} p_{t,i} = k$. Lower bounds for the multiple play bandit problem and matching upper bounds for **Exp3.M** are given. Details of a thresholding version of **Exp3.M** and **DepRound** are discussed in Chapter 4.

In 2017, Zhou and Tomlin [80] investigate a budget-constrained variant of the multi-play bandit problem for both stochastic and non-stochastic rewards. A tuple of arm rewards and associated cost for pulling each arm is learned sequentially until the cumulative cost incurred by the player exceeds some a prior defined budget B. Their approach for the stochastic setting uses UCB1 to achieve an upper bound on regret of order logarithmic in B but exponential in the total number of arms K. Their non-stochastic version assumes the game is played against an oblivious adversary and achieves $\mathcal{O}\big(\sqrt{KB\log(K/k)}\big)$, using **Exp3** as a framework.

More recently in 2019, Alatur, Levy and Krause [2] propose multi-player bandit algorithms for applications based in cognitive radio networks, firstly where players can communicate with each other to avoid colliding with other players (choosing the same arm) and then with a no-communication constraint. The **k-Metaplayer** algorithm chooses a combintorial set of k arms from $\binom{K}{k}$, known as a *meta-arm* and their objective is to minimise the cumulative losses of meta-arms over T rounds. Alatur, Levy and Krause utilise the fact that different meta-arms are dependent upon each other and thus observing the losses of each arm within a meta-arm gives the player information about other meta-arms, reducing the size of the combinatorial search space. The **k-Metaplayer** algorithm achieves an upper bound of $\mathcal{O}\big(k\sqrt{KT\log(K)}\big)$.

The concepts of combinatorial bandits are introduced to the reader by reviewing literature on pure exploration bandit algorithms with stochastic rewards, with the aim of then extending the discussion to problems where they player wants to partition arms depending on the expected mean reward and a given partition parameter. The work of Uchiya *et al.* [72], on multi-player bandits for regret minimisation with adversarial rewards will become a fundamental part of fulfilling specification 1 in our research.

2.4 Label efficient prediction

We depart from our discussion solely from multi-armed bandits and present work on generalisations of the MAB problem. Efficiently querying the feedback from an environment is an important aspect of our Research specification 2. In terms of our application, not only does observing arms when needed reduces the implicit cost of observation, it can also be used as a model of arms with unreliable feedback. Arms that intermittently fail to send feedback is a realistic concern in most applications. A key aspect of our work extends that of György and Ottucsák [35], combining the theory of repeated games with multi-armed bandits.

Prediction with expert advice originates from the pioneering work of Blackwell [16] and Hannan [37] in the 1950s to describe problems involving repeated games. *Label efficient prediction* [39] is a variant of sequential decision-making problems with expert advice, where the player predicts the next value in a sequence generated by some environment and decides whether to observe feedback from their prediction. However, the player has a fixed budget on the number of queries the player can make on the environment. Cesa-Bianchi, Lugosi and Stoltz [23] studied Label efficient prediction in 2004. A lower bound for the problem is given such that for a bound on the total number of experts queried over T rounds, a randomised player cannot achieve better cumulative loss than of order $T\sqrt{\log(K)/m}$, which is sub-linear in T when $m \leq 2\epsilon T$, where $\epsilon > 0$ is the probability of querying an arm.

Also in 2006, György and Ottucsák [35] consider the problem of adaptive routing in the maintenance of packet-switched communication networks. The state of such networks require constant monitoring to control routing protocols using sequential prediction algorithms. In this setting, the performance of a decision maker is measured against a set of experts and the goal is to asymptotically converge to the same average loss as the best expert. The player potentially has access to the advice given by any expert but has no knowledge of how accurate their advice is.

Several algorithms are proposed for tracking the best expert using exponential weights in the full and partial information settings. The label efficient prediction player has access to m experts on average over the time horizon T and randomly makes queries according to a Bernoulli distribution with fixed probability. The choice of expert each

round is based on exponentially weighted unbiased estimates of the experts' advice in line with adversarial multi-armed bandits. A combination of **Exp3** and label efficient prediction is given where the player only learns the loss of her choice if she decides to query it.

In this section, we discussed several works from game theory, specifically related to combining label efficient prediction and multi-armed bandits. The main aim was to review pertinent work for applying MABs with a focus on pulling as few arms as possible without having too much of a detrimental affect on performance. This is a particular requirement of specification 2. However, the review in this section highlights the lack of study on problems involving the semi-bandit setting or the combinatorial partitioning of arms into subsets.

2.5 Adversarial partial monitoring

The *adversarial partial monitoring* problem is a generalisation of multi-armed bandits and label efficient prediction, where the connection between what feedback the adversary reveals and what the player gets to observe is relaxed. Following the discussion in Section 2.4, the work reviewed in this section relates to applications where the player does not want to see or does not have access to the feedback from certain arms.

In 2006, Cesa-Bianchi, Lugosi and Stoltz [24] study repeated games where the player observed a combination of feedback from both the player and their opponent, as an example of adversarial partial monitoring. Several motivating examples are presented before a detailed description of a repeated prediction game is given to define partial monitoring. Algorithm 5 can be seen to work as follows. In each round $t = 1, \ldots, T$, the player chooses an arm I_t randomly according to a probability distribution over

Algorithm 5 Prediction with partial monitoring

Input: number of arms K, number of outcomes M, loss function l, feedback function h.

1: **for** $t = 1, \ldots, T$ **do**
2: Adversary chooses outcome $y_t \in \{1, \ldots, M\}$ without revealing it
3: player chooses probability distribution $\mathbf{p}_t$ over the set of K arms
4: Player draws arm $I_t \sim \mathbf{p}_t$
5: Every arm incurs loss $l(I_t, y_t)$ but not revealed to player
6: Feedback $h(I_t, y_t)$ is revealed to the player

the set of K arms. At the same time, the adversary chooses an outcome y_t from the set of M outcomes but does not reveal it to the player. The loss function for each arm and outcome is given by the loss matrix $\mathbf{L} = [l(i,j)]_{K\times M}$ for $i = 1,\dots,K$ and $j = 1,\dots,M$, which is assumed to be known by the player. The only feedback that the player observed is given by the feedback function $h(I_t, y_y)$ and summarised by the feedback matrix $\mathbf{H} = [h(i,j)]_{K\times M}$. The performance measures considered in [24] is the average regret,

$$\hat{\mathcal{R}}_T = \frac{1}{T}\sum_{t=1}^{T} l(I_t, y_t) - \min_{i\in[K]} \frac{1}{T}\sum_{t=1}^{T} l(i, y_t),$$

and quantifies the difference between average loss of the player and the average loss of the best arm per round. A player that can almost certainly guarantee that their average loss will converge to the average loss of the best arm regardless of the adversaries strategy is known as *Hannan consistent*, after Hannan [37] proved the existence of such a policy in the full information setting, i.e. where $h(i,j) = y_t$. Partial monitoring resolves to the multi-armed bandit when $\mathbf{H} = \mathbf{L}$ and the feedback observed by the player is the loss of the arm pulled. Label efficient prediction (see Section 2.4) is a special case of partial monitoring, where the player decides to query y_t after she has chosen an arm to pull and has to moderate the number of queries they make over the time horizon T. A lower bound on the per round regret for the partial monitoring problem of order $\mathcal{O}(T^{-1/3})$ is given and a Hannan-consistent strategy based on exponential weights that matches the lower bound is also presented.

Many interesting applications exist of repeated games under partial feedback, including prediction with expert advice [16], label efficient prediction [35], multi-armed bandits [10] and the *dynamic pricing* problem, described in terms of online posted-price auctions by Kleinberg and Leighton in 2004, [46]. In learning theory, dynamic pricing is described as a sequential contest between a vendor attempting to sell a product to a stream of customers. The vendor sets the price of the product for each particular customer without knowing how much each customer is willing to spend. The goal of the vendor is to maximise their total revenue or minimise their total lost income. More formally, let $t \in \{1,\dots,T\}$, $x_t \in [0,1]$ denote the price of some product offered by the vendor in round t and $y_t \in [0,1]$ be the most the t^{th} customer is willing to spend then $y_t - x_t$ represents the lost income opportunity from the t^{th} customer if the customer decides to

buy the product. In [46], the vendors' lost income if the customer decides not to buy, i.e. when $x_t > y_t$, is the entire amount the customer would have spent, y_t. In [24], the cost to the vendor when the customer does not buy the product is set at a constant $c > 0$ and can be considered the cost of buying that item of stock and related storage cost. The lost revenue from the t^{th} customer is thus given by,

$$\ell(x_t, y_t) = (y_t - x_t)\mathbb{I}\{x_t \leq y_t\} + c\mathbb{I}\{x_t > y_t\},$$

where $c = y_y$ in [46]. The challenge with such problems is that the vendor only receives information regarding customer spending behaviour, when they actually spend. In [46], the vendor only receives information on whether a sale takes place or not. In a blind auction, neither y_t or $\ell(x_t, y_t)$ are available to the vendor.

In 2013, Bartòk [12] proposed and analysed a novel algorithm for *locally observable* partial monitoring games, where a pair of arms are locally observable if the difference between the losses associated with each arm satisfy certain criteria. A game is locally observable if and only if all neighbouring arms are locally observable, where arms are referred if to as neighbours if the intersection between the cell decomposition $\mathcal{C}_i$ for each arm is empty such that, $\mathcal{C}_i = \{q \in \Delta_M \,|\, \ell_i^\top q = \min_j \ell_j^\top\}$, where $\Delta_M = \{q \in \mathbb{R}^M \,|\, \|q\|_1 = 1, q \geq 0\}$ denotes the $M-1$-dimensional probability simplex, q is a vector consisting of relative frequencies with which each feedback belongs to Δ_M and ℓ_i is the associated loss of arm i.

More recently in 2018, Lattimore and Szepsvàri [51] provide a complete classification of finite partial monitoring games and propose a novel algorithm that extends the work of Bartòk [12] highlighting the connection between the theoretical performance of such algorithms and the inherent structure of the game.

The work reviewed in this section and that of Section 2.4, demonstrates a rich area of research in sequential decision-making problems. In fact, all the algorithms discussed so far in this chapter are at least almost optimal and fulfil Research specification 1. However, none of the work we have discussed so far operates under the requirement to identify arms with reward exceeding a given value, thus not satisfying specification 2.

2.6 Stochastic thresholding bandits

In terms of extending the state-of-the-art in sequential decision-making problems, the work in this thesis is motivated to close several gaps in the current understanding on the *thresholding bandit problem* (TBP). In this section, we introduce the origins of the problem and how TBP research has developed to date.

In 2006, Streeter and Smith [68] consider the fixed budget thresholding multi-armed bandit problem. The goal of the player is to maximise the total number of times an arm returns a value exceeding some threshold after being pulled and over a total of T pulls. The cumulative regret of a player attempting to achieve such a goal is of order $\mathcal{O}(\sqrt{p^* KT \log(T)})$, where p_i denotes the probability that arm i exceeds the threshold and $p^* = \max_{i \in [K]} p_i$.

Introduced as a particular case of a combinatorial pure exploration bandit problem, Locatelli *et al.* [55] study a stochastic bandit setting where the player can sequentially sample from K arms a total of T times. The goal of the player is to identify as efficiently as possible all the arms with expected rewards $\mathbb{E}_{X \sim \nu_i}[X_{t,i}]$, that are larger than a fixed threshold $\theta \in \mathbb{R}$, where $\mu_i = \mathbb{E}[X_{t,i}]$, arm i is associated with unknown reward distribution ν_i. A lower bound for the thresholding bandit problem is given in terms of the probability of mis-classifying an arm as either below or above the threshold of order at least $\mathcal{O}(\exp(-3T/H - 4\log(12\log(T) + 1)K))$, where H defines the hardness or complexity of the problem. Several motivating examples are provided, including active classification problems and anomaly detection.

The *Anytime Parameter-free Thresholding* (**APT**) algorithm is based on the assumption that an almost optimal strategy when the arm distributions are static will want to follow a policy where the number of pulls for arm i, multiplied by the square of the sub-optimality gap for $N_{t,i}\Delta_i^2$, remains constant for arm i throughout the time horizon T. Intuitively, this led Locatelli *et al.* to develop a pulling policy where in round t, the arm that minimises the estimate $N_{t,i}\hat{\Delta}_i^2$ is pulled. The **APT** algorithm is given in Algorithm 6. It is proved that **APT** is comparable with the lower bound and improves on that of **CSAR** and **SR**. TBP is also modified for the problem of reward maximisation, where the goal of the player is to maximise their total cumulative reward and the threshold is set equal to expected value of the optimal arm. The upper bound on

Algorithm 6 Anytime parameter-free thresholding bandit

Input: number of arms $K \geq 2$, number of rounds $T \geq 2K$, distributions $\nu_1, \ldots, \nu_K$, threshold $\theta \in \mathbb{R}$, precision $\epsilon \in \mathbb{R}$.

Pull each arm once

1: **for** $t = K+1, \ldots, T$ **do**

2: Pull arm $I_t = \text{argmin}_{i \in [K]} \sqrt{N_{t,i}} \hat{\Delta}_i$

3: Observe reward $X_{t,I_t} \sim \nu_{I_t}$

Output: $\hat{S}_\theta = \{i \in [K] \,|\, \hat{\mu}_{T,i} \geq \theta\}$

regret given for the problem-dependent case is of order $\mathcal{O}(\Delta_{min}^{-1} \log(T))$ and of order $\mathcal{O}(\sqrt{KT \log(T)})$ for the problem-independent case. It is noted that these performance bounds are threshold-independent, because the threshold is set to equal to the expected value of the optimal arm, μ^*. In the event that μ^* is known (but not which arm it is associated with), then **APT** satisfied both the reward maximisation problem and best arm identification simultaneously. The complexity of TBP is defined as,

$$H := H_{\theta,\varepsilon} = \sum_{i=1}^{K} \frac{1}{\Delta_i},$$

where $\Delta_i = |\mu_i - \theta| + \varepsilon$ for precision $\varepsilon \geq 0$.

Zhong *et al.* [79], propose an asynchronous parallel, empirical variance-guided algorithm. The empirical variance aspect for their *Empirical Variance Thresholding* (**EVT**) algorithm utilises,

$$\hat{\sigma}_{N_{t,i},i} = \sqrt{\frac{1}{N_{t,i}} \sum_{j=1}^{N_{t,i}} \left(X_{t,j} - \hat{\mu}_{t,j}\right)^2},$$

where $N_{t,i}$ denotes the number of times arm i has been pulled after t rounds and $\hat{\mu}_{t,i}$ is the empirical mean of arm i up to round t. The pulling condition differs from **APT** by replacing $\sqrt{N_{t,i}}$ with,

$$\left(\frac{a}{N_{t,i}} + \sqrt{\frac{a}{N_{t,i}}} \hat{\sigma}_{N_{t,i},i}\right)^{-1}, \tag{2.7}$$

where a is a pre-defined parameter.

Theoretical guarantees given in [79] define a in terms of T and their algorithm-specific hardness measure,

$$H_{EVT} = \sum_{i=1}^{K} (\sigma_i^2 \Delta_i^{-2} + \Delta_i^{-1}).$$

The framework for **EVT** also differs from **APT** by initially pulling all arms twice. For the asynchronous parallel aspect of **EVT**, the condition is relaxed on the decision to pull the next arm after the reward from the current pull has been observed. Where most multi-armed bandit algorithms assume that a reward is made available immediately after an arm is pulled, *Asynchronous Parallel*-**EVT** (**AP-EVT**) monitors both the number of times a reward has been observed $N_{t,i}$ and the number of times a reward has been assigned but not yet observed $n_{t,i}$ for each arm up to round t. **AP-EVT** replaces (2.7) with,

$$\left(\frac{a}{N_{t,i} + \delta n_{t,i}} + \sqrt{\frac{a}{N_{t,i} + \delta n_{t,i}}} \hat{\sigma}_{N_{t,i},i} \right)^{-1},$$

where $\delta \in [0, 1]$ determines the weight of assigned but not observed rewards. The hardness measure for AP-EVT H_{AP-EVT} is,

$$H_{AP-EVT} = (1 + \delta\rho)^2 H_{EVT},$$

where it assumed that $n_{t,i} \leq \rho N_{t,i}$ in the theoretical guarantees given. It is also noted that ρ is proportional to the number of players in the asynchronous multi-player bandit setting. Parameter-free versions of both EVT and AP-EVT are given with similar theoretical guarantees.

Mukherjee *et al.* [60] propose a variance-based combinatorial pure exploration bandit, specifically designed for TBP. The *Augmented-UCB* (**AugUCB**) algorithm combines **APT** from [55] with a *round-based* **UCB**-type algorithm from [11] called *UCB-improved*, where all arms are pulled equally each round and eliminates those identified as sub-optimal. Upper regret bounds are proved that improve on those for **CSAR** and **APT**, in terms of a variance-based complexity measure.

Zhuang *et al.* [81] consider an adaptive version of TBP in the fixed-confidence setting, where the goal of the player is to identify outlier arms. The detection of an outlier arm is based on evaluating an outlier threshold that depends on the mean reward of all arms. Confidence bounds are maintained for each arm and that of the outlier threshold. As soon as there is no overlap between the bounds for all arms with the outlier threshold, the algorithm terminates with a high probability of correctness. Theoretical guarantees are given as bounds on the number of rounds needed for the algorithm to terminate and in terms of a round robin version of the problem complexity,

$$T \leq 8R^2 H_{RR}\left[\log\left(\frac{2R^2\pi^2(K+1)H_{RR}}{3\delta}+1\right)\right]+4K,$$

where $X_{t,i} \in [a,b]$, $R = b-a$, $H_{RR} = (1+\sqrt{l(i)})^2 H_2$ and,

$$l(i) = \left[\sqrt{\frac{(1+i\sqrt{K-1})^2}{K}}+\sqrt{\frac{i^2}{2\log(\pi^2 K^3/6\delta)}}\right]^2.$$

Improved guarantees are also given for a weighted round robin version such that after all arms have been pulled once, the player keeps pulling the same arm until the number of pulls exceeds some multiple of a parameter ρ or there is no overlap between the confidence bounds on the arm and outlier threshold. The player then moves on to pulling the next arm.

Katz-Samuels and Scott [45], study feasible arm identification, where the goal is to find which arms belong to some polyhedral set, which generalises the stochastic thresholding bandit problem. Several algorithms are proposed that are multi-dimensional extensions to **SAR**, **UCB** and **APT**. Lower bounds are given for the problem and near-optimal upper bounds are given.

Kano *et al.* [43], introduce the problem of *Good Arm Identification* (GAI) and define a problem-dependent exploration-exploitation dilemma, called a *dilemma of confidence*, to achieve more confidence in the goodness of arms. The *Hybrid algorithm for the Dilemma of Confidence* (**HDoC**) is based on confidence bounds for cumulative regret minimisation and best arm identification. A *good* arm is defined as having expected reward exceeding a given threshold and it is assumed that there exists at least one arm with expected reward greater than threshold. The player pulls arms each round and receives an iid reward from the associated arm distribution. As soon as the player can

confidently identify an arm as good, she outputs the arm, eliminates it from the search and then repeats the process until no good arms remain. Lower and upper bounds are presented for GAI, in terms of a problem-specific probably-approximately correct framework.

All of the works presented in this section assume that rewards are generated by some stochastic process associated with each arm. In particular, it is assumed that the sequence of rewards returned by some arm are independent samples of that process. When the feedback from an arm is associated with some physical process, this assumption does not hold. However, since there is dependence, we can use previous observations to support our decision to observe the arm in future.

2.7 Summary

In this section we evaluate the similarities and differences that we identify between the literature highlighted in this chapter and that of the adversarial thresholding semi-bandit problem.

The reader is introduced to the classical multi-armed bandit problem, where the sequence of rewards are generated both stochastically and adversarially. The objective of these types of games is to maximise the cumulative sum of rewards after pulling arms over a sequence of T rounds, whilst balancing the exploration-exploitation dilemma. Alternatively, this is studied in terms of minimising the expected regret between the player and playing optimally in hindsight. This gives the reader an understanding of the theoretical framework for the problem and highlights algorithms that have driven many developments in the literature. The adversarial setting in particular, provides the general concepts for the adversarial thresholding semi-bandit problem.

We present some of the literature on bandits where a combinatorial structure is placed on the set of arms. This includes works on the semi-bandit problem, where at least one arm is pulled each round and problems where the player can have multiple plays or there are multiple players. The semi-bandit problem has been studied for both regret minimisation and pure exploration. However, the majority of work in the multi-play setting faces towards regret minimisation against an adversarial opponent.

The adversarial partial monitoring problem generalises multi-armed bandits, where there is a distinction between the feedback generated by an adversary for each arm every round and the reward the player actually observes. The connection between adversarial partial monitoring and our problem is that in playing the ATSBP game, the player is interested in which arms exceed a given threshold and by how much each round, and the reward received by the player is a function of whether the feedback from the environment for observed arms exceeds threshold or not.

Label efficient prediction is a variant of sequential decision-making and problems involving prediction with expert advice. The general idea is that the player only observes feedback when they specifically ask for it and this is constrained by a fixed budget of queries. The adversarial thresholding semi-bandit problem can relate to this idea in two potential ways. Firstly, by wanting to observe arms when the player believes their feedback will exceed threshold but still playing randomly to remain competitive in an adversarial environment. Secondly, The label-efficient semi-bandit can model applications where the receipt of information from some sensor may not be reliable.

The thresholding bandit originates as a pure exploration problem, where exploiting arms is redundant and the objective of the game is to identify which arm satisfies some condition either after a fixed number of rounds have been played or a pre-defined level of accuracy has been achieved to a given level of confidence. Various approaches are discussed for attacking the best arm identification problem, the most common type of pure exploration multi-armed bandit. We dedicate a section to work studying stochastic thresholding bandits, specifically to highlight the gap in the literature for adversarial thresholding bandits.

In Figure 2.1, we illustrate a taxonomy of selected multi-armed bandit and semi-bandit problems in particular, to explain our motivation for the literature reviewed in this chapter and is in no way intended to be comprehensive. The paradigm of regret minimisation relates to the balancing of the exploration-exploitation dilemma, with works discussed in Section 2.1. In Section 2.2, we reviewed the literature on multi-armed bandits designed purely to identify the *best* arm. This field of research is mainly related to elimination-type algorithms and aim to minimise the sample complexity of finding the best arm. In Section 2.3, we discuss a variant of pure exploration bandits, where several arms can be pulled together and the combinatorial structure of the

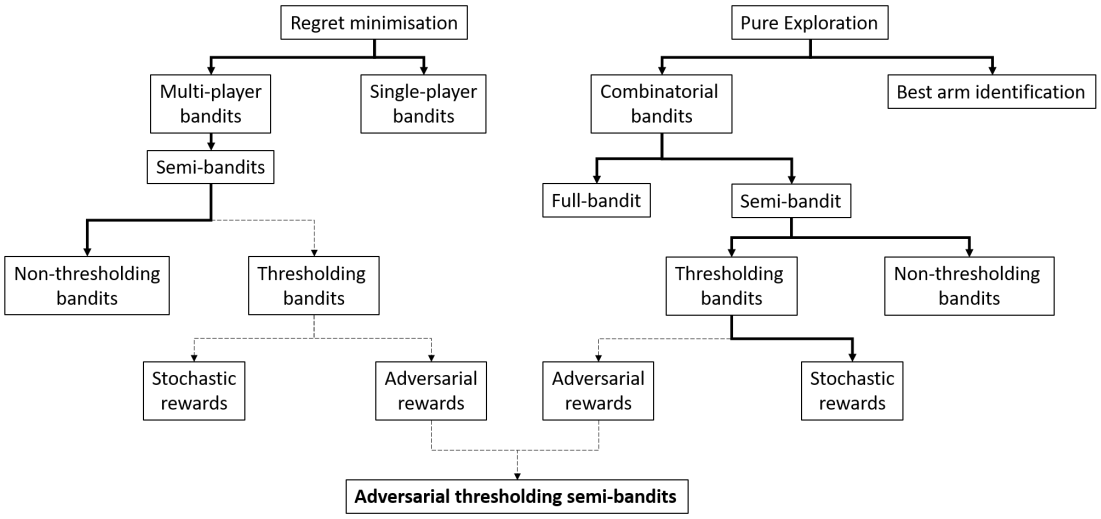


Figure 2.1: Taxonomy of how adversarial thresholding semi-bandits extend the state-of-the-art in multi-armed bandit problems. Dashed lines indicate perceived gaps in knowledge.

arms is taken advantage of. We identify a connection between regret minimisation and pure exploration. In the former, the semi-bandit setting is equivalent to having multiple players all co-operating and communicating, with the aim of pulling the best arm possible as often as possible for each player. In the latter, a combinatorial semi-bandit player has access to the rewards from all the arms they pull and want to identify the top k arms. A constrained variant of the combinatorial semi-bandit problem is the thresholding bandit, where the objective is to find all arms with mean reward exceeding a given threshold (and thus all remaining arms with mean reward within threshold). Within the pure exploration field, thresholding bandits has a growing body of work (see Section 2.6) but only for the stochastic case. For regret minimisation, no threshold constraint has yet been applied to non-stochastic multi-armed bandits. To the best of our knowledge, adversarial thresholding semi-bandits is a novel bandit model, extending the state-of-the-art in thresholding bandits for adversarial rewards in the regret minimisation regime.

Many sequential decision-making problems do not benefit from rewards generated by some stochastic process, especially when we consider sequential feedback from monitoring natural phenomenon. The ability to identify where and when responses deviate from typical behaviour is crucial in such scenarios. The work reviewed in this chapter also demonstrates a gap for thresholding bandits in the semi-bandit or multi-play setting and in label efficient prediction.

Chapter 3

Adversarial thresholding semi-bandits

In this chapter, we introduce a new abstraction of the MAB model, called the *adversarial thresholding semi-bandit problem* (ATSBP). A player selects a subset of arms each round and commits to observing an arm depending on the history of previous observations for that arm. The objective of the player is to observe only those arms that exceed a threshold in every round. This includes observing no arms if none exceed threshold and observing every arm if they all do.

Current approaches to stochastic thresholding bandit problems (TBP) bound their performance guarantees in terms of the probability of mistakenly classifying the expected reward from each arm as either exceeding a threshold or not. This is quantified by taking the expectation of the number of mistakes made by the algorithm, over T rounds. Since we are considering TBP related to balancing exploration and exploitation under adversarial feedback, we measure algorithmic performance in terms of expected regret.

In Section 3.1, the adversarial thresholding semi-bandit problem (ATSBP) and associated terminology is defined. In Theorem 1, the existence of a unique subset of arms is shown, which maximises feedback available to the player. A notion of expected regret, specific to ATSBP is defined and finally, we present the ATSBP algorithm (see Algorithm 7). In Section 3.2, we define the *hardness* or complexity of the adversarial thresholding semi-bandit problem in terms of the frequency with which an arm crosses

the threshold boundary and state a prove a bound on ATSBP hardness in terms of the number of arms exceeding threshold in consecutive rounds. In Section 3.3, we define a measure of arm-pulling efficiency for the adversarial thresholding semi-bandit problem.

3.1 Formal description

In this section, we introduce the mathematical framework used for describing the adversarial thresholding semi-bandit problem. We define the required terminology related to the problem and prevalent throughout MAB literature in Section 3.1.1. We identify the requirements for measuring the regret of adversarial thresholding semi-bandit algorithms in Section 3.1.2.

Following the introduction to multi-armed bandits given in Section 1.2, consider a set of K arms, where a player decides on a policy of pulling arms over a sequence of $T > 0$ rounds. In each round $t = 1, \ldots, T$, the player selects a subset of arms and intends to observe only those arms that are exceeding a pre-defined threshold value.

Let $K > 0$ be an integer and $[K] = \{1, \ldots, K\}$ be the set of arms. The player selects a set of arms I_t such that the number of arms selected in round t, is denoted by $1 \leq k_t \leq K$. The player decides to observe the feedback from none, some or all of the arms, $i \in I_t$, based on their policy for observing arms. The set of selected arms I_t belongs to the space of possible subsets of $[K]$, with exactly k elements, denoted $\mathbb{C}([K], k) = \{I \subseteq [K] \,|\, |I| = k\}$, where $|I|$ denotes the cardinality of set I.

In the adversarial thresholding semi-bandit problem at least one arm is selected each round, $I_t \notin \emptyset$ for $t = 1, \ldots, T$, but the player can decide not to query any arms, $i \in I_t$. The combinatorial game is played in the full-bandit setting where the player only receives the combined feedback from arms in the subset and not the feedback from individual arms. The semi-bandit resolves to the classical MAB when $k = 1$.

The general framework for the adversarial thresholding semi-bandit problem is described in the following algorithm (see Algorithm 7). A player has to decide how she selects arms to ensure that every arm is inspected sufficiently to predict in which rounds it will exceed the threshold, when the player only has capacity to observe a subset of the information or wants to minimise the total number of observations. The player needs to do this whilst competing against an adversary.

Algorithm 7 ATSBP

Input: Number of arms K, number of rounds T, number of selections $1 \leq k_t \leq K$ for $t \in \{1, \ldots, T\}$ and threshold $0 < \theta < 1$.

1: **for** $t = 1, \ldots, T$ **do**
2: adversary has knowledge of a vector of feedback $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,K})$ without revealing to player.
3: player decides how to select k_t arms $I_t \subseteq [K]$.
4: player decides whether to observe $y_{t,i}$ or not for $i \in I_t$ depending on history of observations exceeding threshold, θ.
5: player receives reward $x_{t,i}$, depending on reward criteria.

We have now set out a formal description for multi-armed bandits and semi-bandits in particular. A general framework for the adversarial thresholding semi-bandit problem has also been introduced. Before proceeding any further, we address the language of multi-armed bandits.

3.1.1 Bandit terminology

The terminology in this work remains closely related to that in literature. In this section, we define terms that are used frequently throughout this work, whilst describing the mechanics of ATSBP. We also use this opportunity to decompose the decision process inherent in all bandit problems.

We denote independent and identically distributed (iid) samples from a probability distribution in upper case, with subscripts denoting the round and arm, where context appropriate. In this context, an *arm* is defined as a generator of information or *feedback* and a *round* is a single instance of the game. We will denote feedback from an adversarial (non-stochastic) environment with lower case.

A *player* or *algorithm* plays a sequential game in rounds, against an *adversary*, *opponent* or *environment*. Classically, decisions made by the player each round corresponds to them *pulling* one or some of K possible arms. The player receives feedback (*reward* or *loss*) only for the arm or arms pulled (in this semi-bandit setting). In this work, the decision process is decomposed further. Before pulling an arm, the player makes the decision to pull that arm such that the arm is *selected* or *chosen*. The player then makes the decision to *query*, *observe* or *inspect* selected arms.

The adversary generates or has knowledge of feedback from the environment for each arm in every round. The player is provided with the feedback for selecting and observing an arm. In this context, *pulling* an arm corresponds to selection and observation of that arm.

Before the game has begun, an oblivious adversary chooses a set of KT observations over all arms and every round. Without loss of generality, we assume that feedback is bound by the unit interval, $(y_{t,i}) \in [0,1]^{T\times K}$. Each round, the adversary retrieves a vector of feedback $\mathbf{y}_t \in [0,1]^K$ for every arm. Arm selected by the player in round t, $i \in I_t$, are observed dependent on their history of previous observations, leading to none, some or all selected arms potentially being observed. We denote the observation from arm i in rounds t as $x_{t,i}$, when feedback $y_{t,i}$ is selected and observed.

The player has knowledge of a threshold value $\theta \in (0,1)$. Where $\theta = 1$, the game is considered to be *degenerate*. The players' objective is to observe arms only for the rounds that their feedback exceeds the threshold. The challenge is to achieve this objective with the player having no prior knowledge of the adversary's choice of feedback. The player is rewarded with the value of the feedback when it exceeds threshold, $x_{t,i}\mathbb{I}\{x_{t,i} > \theta\}$, where $\mathbb{I}\{\cdot\} \in \{0,1\}$ is one when the condition $\{\cdot\}$ is satisfied and zero otherwise. We introduce the problem theoretically, in terms of a static threshold $\theta_t = \theta$, in every round. However, we discuss how an adaptive threshold can be implemented (see Chapter 8). In the more fundamental context, *exceeding threshold* refers to the feedback received for pulling arm i in round t being more than the threshold value, $x_{t,i} > \theta$. Alternatively, feedback exceeds threshold if it belongs to the half-open interval, $x_{t,i} \in (\theta, 1]$ and within threshold if $x_{t,i} \in [0, \theta]$.

The general properties of adversarial thresholding semi-bandits have been defined, with the most commonly used terms clearly defined. Several key aspects of ATSBP have been discussed, including decomposition of the arm-pulling process and the dual-aspect objective. We now need to address a performance metric, suitable for ATSBP.

3.1.2 ATSBP regret

Since the objective of ATSBP is twofold: (i) pull only arms with feedback exceeding threshold and (ii) accumulate as much combined feedback as possible (also known as *reward maximisation*, it is conceivable that a player could pull a collection of arms with

combined feedback greater than all arms exceeding threshold, but none of the collection have feedback exceeding threshold. Consequently, this section presents a novel notion of regret that provably ensures the (optimal) set of arms, exclusively exceeding threshold, maximises the combined feedback from any combination of k arms. In Theorem 1, we prove the unique existence of an optimal subset of arms that maximises a quantity combining the feedback from all pulled arms.

Optimality in terms of ATSBP is thus, deciding to pull arms exclusively in rounds where their feedback exceeds threshold, minimising the regret incurred from pulling arms when their feedback is within threshold and not pulling arms when their feedback exceeds threshold. An optimal policy when the objective is concerned only with pulling all arms exceeding a threshold $\theta \in (0, 1)$, is to pull the *top* q_t arms, when exactly q_t are exceeding threshold each round,

$$q_t = \sum_{i=1}^{K} \mathbb{I}\{x_{t,i} > \theta\}, \tag{3.1}$$

where $q_t = |I_t^*|$, denotes the number of arms pulled by the optimal policy in round t for $t = 1, \ldots, T$. For reward maximisation to be achievable, there must exist a unique subset $I_t^* \subseteq [K]$. The reward accrued each round by playing the adversarial thresholding semi-bandit problem is thus given by,

$$\frac{1}{|k_t - q_t| + 1} \sum_{i=1}^{K} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\}. \tag{3.2}$$

The form of (3.2) is justified by the following theorem.

Theorem 1. *For integers $K > 0$, $T > 0$, $t = 1, \ldots, T$ and $0 \leq q_t \leq K$, there exists a subset of arms $I_t^* \subseteq [K]$ where $[K] = \{1, \ldots, K\}$, $q_t = |I_t^*|$ and $q_t \neq \emptyset$ such that,*

$$\sum_{i \in I_t^*} x_{t,i} \geq \frac{1}{|k_t - q_t| + 1} \sum_{i \in I_t} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\},$$

where $k_t = |I_t|$, $x_{t,i} \in [0, 1]$ and $\theta \in (0, 1)$. Equality holds if and only if $I_t = I_t^*$.

Proof of Theorem 1. We prove the existence of I_t^* by considering the relationship between (3.2) and $\sum_{i\in I_t^*} x_{t,i}$.

Assuming that $k_t < q_t$, where the player pulls fewer arms than are optimal, then $\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} < \sum_{i\in I_t^*} x_{t,i}$ and since $1/(|k_t - q_t| + 1) < 1$ we have,

$$\frac{1}{|k_t - q_t| + 1} \sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} < \sum_{i\in I_t^*} x_{t,i}.$$

Assume $k_t > q_t$ and $\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} = \sum_{i\in I_t^*} x_{t,i}$, where the player pulls all optimal arms and $k_t - q_t$ sub-optimal arms. Then,

$$\frac{1}{|k_t - q_t| + 1} \sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} < \sum_{i\in I_t^*} x_{t,i}.$$

Assume that $k_t = q_t$ and $\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} < \sum_{i\in I_t^*} x_{t,i}$, where the player pulls the same number of arms as those that are optimal but some or all are sub-optimal. Then we have,

$$\frac{1}{|k_t - q_t| + 1} \sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} < \sum_{i\in I_t^*} x_{t,i}.$$

The only event where (3.2) is equal to the sum of only the optimal arms in round t is when $\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} = \sum_{i\in I_t^*} x_{t,i}$ and $k_t = q_t$. That is, when the sum of rewards for arms pulled by the player is the sum of rewards for the optimal arms and only those arms. This completes the proof. □

Corollary 1.1 follows by considering the event where every arm generates feedback within threshold.

Corollary 1.1. *In the event that none of the K arms exceed threshold in round t, where $I_t^* = \emptyset$ and $q_t = 0$ then,*

$$\sum_{i\in I_t^*} x_{t,i} = 0 \quad \textit{and} \quad \frac{1}{|k_t - q_t| + 1} \sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} = 0.$$

Corollary 1.1 states that when the player has no possibility of finding any arms exceeding threshold, they receives no reward. However, since $q_t = 0$ then the optimal policy also returns no reward. The player learns no new information about arms exceeding threshold, but the decisions made incur no regret. We refer to these rounds as *free-play*, noting that the player has no prior knowledge of when a free-play round takes place.

As introduced in Section 2.1.1, the classical notion of regret is the difference between the reward from pulling the best arm or set of arms in hindsight and the reward accumulated by the arm or arms pulled by the player. The regret in the adversarial thresholding semi-bandit game can thus be considered as the difference between the combined feedback from pulling only those arms exceeding threshold and of those pulled by the player.

Definition 1. *Adversarial thresholding semi-bandit regret is defined as,*

$$\overline{\mathcal{R}}_T := \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in I_t^*} x_{t,i} - \sum_{t=1}^{T}\frac{1}{\Delta_t}\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\}\,\middle|\, I_1,\ldots,I_T\right],$$

where $\Delta_t = |k_t - q_t| + 1$ *and expectation is taken in terms of the randomness in selecting subsets of arms, over the course of the game.*

In this section, we proved the existence of an optimal subset of arms that maximises the combined feedback from pulled arms, when those arms exceed threshold. From this, we defined a notion of regret for the adversarial thresholding semi-bandit problem suitable for the development of algorithms that perform reward maximisation.

In a bandit setting, optimality is only associated with the pulling of arms when their feedback exceeds threshold. The optimal policy maximises the total feedback revealed to the player. If the player pulls an arm with feedback within threshold, they are not rewarded with that feedback.The player receives no information regarding feedback from arms not pulled.

3.1.3 ATSBP outcomes

In this section, we discuss why ATSBP is a multi-armed bandit variant and how the problem could also be considered as a more general sequential decision-making problem (as discussed in Section 2.5).

When playing the adversarial thresholding semi-bandit game, a player must decide to select arm i in round t and then decide to observe it, depending on whether she believes the feedback will exceed threshold or not. A *correct* outcome is thus satisfied by one of three conditions: (i) select and observe arm i in round t such that $x_{t,i} = y_{t,i}$ when $y_{t,i} > \theta$, (ii) select and not observe arm i when $y_{t,i} \leq \theta$ or (iii) not select arm i when $y_{t,i} \leq \theta$.

The metric given in Definition 1, accounts for correct outcome (i) but not for outcomes (ii) and (iii), as the player does not have access to feedback from arms not pulled. By considering all of the potential outcomes, it is possible for the player to gain appropriate feedback depending on whether their decision was correct or not as follows,

(i) Arms selected, queried and with feedback exceeding threshold. The feedback received by the player is given by,

$$y_{t,i}\mathbb{I}\{i \in I_t\}\mathbb{I}\{i \in \Lambda_t\}\mathbb{I}\{y_{t,i} > \theta\},$$

(ii) Arms selected but not queried, with feedback within threshold,

$$(1 - y_{t,i})\mathbb{I}\{i \in I_t\}\mathbb{I}\{i \in \overline{\Lambda}_t\}\mathbb{I}\{y_{t,i} \leq \theta\},$$

(iii) Arms not selected (thus also not queried), with feedback within threshold,

$$(1 - y_{t,i})\mathbb{I}\{i \notin I_t\}\mathbb{I}\{y_{t,i} \leq \theta\},$$

(iv) Arms selected and queried, but with feedback within threshold,

$$y_{t,i}\mathbb{I}\{i \in I_t\}\mathbb{I}\{i \in \Lambda_t\}\mathbb{I}\{y_{t,i} \leq \theta\},$$

(v) Arms selected but not queried, with feedback exceeding threshold,

$$(1 - y_{t,i})\mathbb{I}\{i \in I_t\}\mathbb{I}\{i \in \overline{\Lambda}_t\}\mathbb{I}\{y_{t,i} > \theta\},$$

(vi) Arms not selected, with feedback exceeding threshold,

$$(1 - y_{t,i})\mathbb{I}\{i \notin I_t\}\mathbb{I}\{y_{t,i} > \theta\},$$

where $\Lambda_t = \{i \in I_t \,|\, B_{t,i} = 1\}$ and $\overline{\Lambda}_t = \{i \in I_t \,|\, B_{t,i} = 0\}$ denote the sets of selected arms observed and not observed by the player, respectively for Bernoulli parameter $B_{t,i} \in \{0, 1\}$, (see [35] and [23] for further details on making label efficient decisions, using sequences of Bernoulli random variables).

Subsequently, the feedback gained (but not necessarily completely revealed to the player) is,

$$x_{t,i} = \begin{cases} y_{t,i} & \text{if } i \in I_t \cap \Lambda_t, y_{t,i} > \theta \\ 1 - y_{t,i} & \text{otherwise,} \end{cases}$$

where $i \in I_t \cap \Lambda_t$, denotes the arms selected and observed by the player in round t. The only feedback revealed to the player is associated with outcomes (i) and (iv). In this respect, the adversarial thresholding semi-bandit player does not have access to all the information with which their performance is being judged. Under these circumstances ATSBP corresponds with a repeated game under partial monitoring (see Chapter 2 for further details) and should be referred to as the *adversarial thresholding partial monitoring problem*.

However in bandit games, the ATSBP feedback observed is the reward received,

$$x_{t,i} = \begin{cases} y_{t,i} & \text{if } i \in I_t \cap \Lambda_t, y_{t,i} > \theta \\ 0 & \text{otherwise.} \end{cases}$$

The feedback from arms not pulled is never revealed to the player in ATSBP. In fact, the player is only rewarded with feedback when they pull an arm exceeding threshold. From this, we can compute an un-biased estimator of the combined feedback from pulled arms, crucial for satisfying specification 1 with adversarial multi-armed bandits (see Section 4.3.1).

The argument presented in this section justifies the use of multi-armed bandit theory on this problem. This is particularly appropriate for applications where information is only revealed when an observation is made.

3.2 Hardness of ATSBP

In this section we discuss the complexity or *hardness* H_θ, of the adversarial thresholding semi-bandit problem. Intuitively, the dynamics of this measure depend on the relationships between the threshold θ and the number of arms pulled k_t with the total number of arms K and the time horizon T.

The complexity of the stochastic thresholding bandit problem H, is measured in terms of the sub-optimality gap between the expected value of each arm and the threshold, plus some precision (see Section 2.6). However, H requires that $\mu_i = \mathbb{E}[X_{t,i}]$, which is meaningless in an adversarial setting. As such, ATSBP requires its own measure on the complexity of the problem. Since we make no assumptions on the sequence of observations we receive, we cannot plan ahead based on any statistical confidence. The problem itself dictates that our primary concern is correctly classifying whether an arm is either above or below the threshold value, hence we need to pay particular attention to frequency with which an arm crosses the threshold boundary. Our justification being that the more times an arm vacillates about the threshold, the more opportunity the player has to make a mistake.

Since we are making no statistical assumptions on how the adversary generates the feedback observed by the player, the complexity of the problem cannot be defined in the same way it is for the stochastic thresholding bandit problem, given in [55]. As introduced in (2.2), H defines the *hardness* of the stochastic problem in terms of the sum of squared sub-optimality gaps between the threshold and the expected mean of each arm. However, we define the complexity of the problem in terms of the threshold, the frequency with which an arm's feedback crosses the threshold boundary and the size of that crossing.

Definition 2. *The hardness (complexity) of the adversarial thresholding semi-bandit problem is defined as,*

$$H_\theta := \frac{2}{1-\theta}\sum_{t=1}^{T-1}\sum_{i=1}^{K}\frac{1}{d_{t,i}}\mathbb{I}\{(y_{t,i}-\theta)(y_{t+1,i}-\theta)\leq 0\},$$

where $d_{t,i} = |y_{t,i} - y_{t+1,i}| + 1$.

The more an arm's feedback crosses the threshold boundary, the more difficult ATSBP becomes. ATSBP also increases in challenge, as likelihood of detecting an with feedback exceeding threshold decreases.

Since the number of times the feedback from an arm crosses the threshold boundary is intimately linked to the number of times feedback for an arm exceeds threshold in successive rounds (and thus the number of times feedback in successive rounds is within threshold) and the number of arms with feedback exceeding threshold in successive rounds, we can state the following.

Theorem 2. *Given K sequences of feedback $y_{1,i},\ldots,y_{T,i}$, where $y_{t,i}\in[0,1]$ for rounds $t = 1,\ldots,T$ and arms $i = 1,\ldots,K$, let H_θ, denote the hardness measure for the adversarial thresholding semi-bandit problem with threshold $\theta\in(0,1)$. Then,*

$$(1-\theta)H_\theta \geq \sum_{t=1}^{T-1} q_t + q_{t+1} - 2Q_t,$$

where $q_t = \sum_{i=1}^{K}\mathbb{I}\{y_{t,i} > \theta\}$ is the number of arms with feedback exceeding threshold in round t and $Q_t = \sum_{i=1}^{K}\mathbb{I}\{y_{t,i} > \theta\}\mathbb{I}\{y_{t+1,i} > \theta\}$ denotes the number of arms with feedback exceeding threshold in successive rounds.

Proof. From Definition 2 we have,

$$H_\theta \geq \frac{1}{1-\theta}\sum_{t=1}^{T-1}\sum_{i=1}^{K}\mathbb{I}\{(y_{t,i}-\theta)(y_{t+1,i}-\theta)\leq 0\}, \tag{3.3}$$

since $y_{t,i}\in[0,1]$ for $t = 1,\ldots,T$ and $i = 1,\ldots,K$, then $0 \leq |y_{t,i} - y_{t+1,i}| \leq 1$ and $d_{t,i}\leq 2$.

Let $h_{t,i}$ denote a binary variable indicating whether arm i crosses the threshold boundary between rounds t and $t+1$,

$$\begin{aligned} h_{t,i} &= \mathbb{I}\{(y_{t,i}-\theta)(y_{t+1,i}-\theta) \leq 0\} \\ &= \mathbb{I}\{y_{t,i} > \theta\}\mathbb{I}\{y_{t+1,i} \leq \theta\} + \mathbb{I}\{y_{t,i} \leq \theta\}\mathbb{I}\{y_{t+1,i} > \theta\} \\ &= \mathbb{I}\{y_{t,i} > \theta\}(1 - \mathbb{I}\{y_{t+1,i} > \theta\}) + (1 - \mathbb{I}\{y_{t,i} > \theta\})\mathbb{I}\{y_{t+1,i} > \theta\} \\ &= \mathbb{I}\{y_{t,i} > \theta\} + \mathbb{I}\{y_{t+1,i} > \theta\} - 2\mathbb{I}\{y_{t,i} > \theta\}\mathbb{I}\{y_{t+1,i} > \theta\}. \end{aligned} \tag{3.4}$$

where the final term in (3.4) represents the binary indicator for feedback exceeding threshold in consecutive rounds. Summing over rounds $t = 1, \ldots, T-1$, all arms $i = 1 \ldots, K$ and combining with (3.3) gives,

$$H_\theta \geq \frac{1}{1-\theta}\sum_{t=1}^{T-1} h_{t,i} \geq \frac{1}{1-\theta}\sum_{t=1}^{T-1} q_t + q_{t+1} - 2Q_t, \tag{3.5}$$

where $q_t = \sum_{i=1}^{K}\mathbb{I}\{y_{t,i} > \theta\}$ and $Q_t = \sum_{i=1}^{K}\mathbb{I}\{y_{t,i} > \theta\}\mathbb{I}\{y_{t+1,i} > \theta\}$ in (3.5). This completes the proof. □

Not only does this section provide a theoretical measure for the hardness of the adversarial thresholding semi-bandit problem, but it also elucidates the differences in challenge between that of its stochastic counterpart. We also prove that the hardness of the adversarial thresholding semi-bandit problem is intimately linked with the number of arms exceeding threshold in consecutive rounds.

3.3 Efficiency ratios

In this section, we define a measure of efficiency for the adversarial thresholding semi-bandit problem, using statistical inference theory. Integral to ATSBP is the decision that the player must make each round, relating to two competing propositions. That is, the feedback from observed arms is either exceeding threshold or not. Recalling that the objective of the player is to only pull arms when their corresponding feedback is exceeding threshold, the player makes an erroneous decision for outcomes (iv), (v) and (vi) from Section 3.1.

After the results from T rounds of the game, theoretically, the adversary can define the efficiency of an ATSBP algorithm as (α,β)-ATSBP, where α and β are defined as follows:

- **False positive rate, α:** the number of arms pulled with feedback within threshold out of the total number of arms with feedback within threshold, summed over every round,

$$\alpha = \sum_{t=1}^{T} \sum_{i=1}^{K} \frac{\mathbb{I}\{i \in I_t \cap \Lambda_t \,|\, y_{t,i} \leq \theta\}}{\mathbb{I}\{y_{t,i} \leq \theta\}}. \tag{3.6}$$

- **False negative rate, β:** the number of arms not pulled with feedback exceeding threshold out of the total number of arms with feedback exceeding threshold, summed over every round,

$$\beta = \sum_{t=1}^{T} \sum_{i=1}^{K} \frac{\left(\mathbb{I}\{i \in I_t \cap \overline{\Lambda}_t \,|\, y_{t,i} > \theta\} + \mathbb{I}\{i \notin I_t \,|\, y_{t,i} > \theta\}\right)}{\mathbb{I}\{y_{t,i} > \theta\}}. \tag{3.7}$$

We note that if no arms are within threshold, then no arms can also be observed within threshold. We also note that if no arms exceed threshold, then no arms can not be observed exceeding threshold. However, the player does not have access to feedback when arms are selected but not queried ($i \in I_t \cap \overline{\Lambda}_t$) or not selected at all ($i \notin I_t$). So in this respect, the only measure of efficiency available to the player is referred to in statistical inference as,

- **False discovery rate, FDR:** the number of arms pulled with feedback within threshold out of the total number of arms pulled, summed over every round,

$$\text{FDR} = \sum_{t=1}^{T} \sum_{i=1}^{K} \frac{\mathbb{I}\{i \in I_t \cap \Lambda_t \,|\, y_{t,i} \leq \theta\}}{\mathbb{I}\{i \in I_t \cap \Lambda_t\}}. \tag{3.8}$$

For the player, we define the efficiency of an adversarial thresholding semi-bandit algorithm as FDR-ATSBP. Since in terms of pulling arms, an efficient algorithm wants to minimise the number of pulls when the feedback is within threshold. Hence, an optimal adversarial thresholding semi-bandit algorithm is 0-ASTBP for the player. In each of the quantities, α, β and FDR, it is assumed that there is at least one arm

within threshold, at least one exceeding threshold and at least one arm is queried, respectively in at least one of the T rounds and a corresponding assumption is also made in statistical inference. If these assumptions do not hold, then ATSBP is considered trivial. Unfortunately, the player has no idea of the number of false negatives. In critical situations such as monitoring current flow in electrical devices on the LHC, the only way to minimise false negatives (disregard a channel exhibiting overcurrent) is to pull arms more frequently.

3.4 Summary

In this chapter, we formally presented the adversarial thresholding semi-bandit problem. In Section 3.1, we defined a novel bandit model for pulling arms with associated non-stochastic feedback that exceeds a given threshold. The pulling process itself has been decomposed into a selection phase and a query phase, allowing for the development of innovative algorithms that optimise both phases.

The existence of an optimal subset of arms is proven that maximises the combined feedback from pulled arms and used to define a problem-specific notion of regret in Section 3.1.2. We discuss how feedback can be interpreted to fit a generalisation of the bandit problem and why it is appropriate to call ATSBP a bandit in Section 3.1.3.

In Section 3.2, we highlighted the need for a different notion of problem complexity compared with the original stochastic TBP and defined the hardness of the adversarial thresholding semi-bandit problem, in terms of the total number of arms exceeding threshold in consecutive rounds. We also prove a bound on H_θ linked with the number of arms with feedback exceeding threshold each round and the total length of sequences of consecutive rounds exceeding threshold.

In Section 3.3, we defined efficiency ratios for the adversary and the player since each has access to different information. The adversary and player efficiency ratios (3.8), will be used for empirical analysis of the algorithms developed and aid in the evaluation of specification 2.

Chapter 4

Adversarial thresholding bandits with multiple plays

In 2010, the adversarial multi-armed bandit problem was extended to the multiple player setting by Uchiya *et al.* [72], where $k \leq K$ players aim to maximise their combined cumulative reward compared with pulling the optimal fixed subset of k arms. In this chapter, Chapters 5 and 6, we use elements of the work set out in [72] as a framework for developing algorithms that can extend the adversarial multi-player bandit to the adversarial thresholding semi-bandit problem.

In Section 4.1, we state a co-operative assumption, allowing for equivalence in the multi-play and semi-bandit settings. As a consequence of the multi-player framework, we also make a simplifying assumption in terms of the number of arms exceeding threshold each round. In Section 4.2, we briefly describe the k-combination selection procedure used in [72], to select a set of k distinct arms such that each arm is selected according to a given probability and discuss the key components of how the adversarial multi-player bandit is extended to the adversarial thresholding semi-bandit model. We introduce the *thresholding exponentially weighted exploration and exploitation with multiple plays* (**T-Exp3.M**) algorithm to bridge the gap between thresholding bandits, adversarial environments and regret minimisation in Section 4.4. Finally in Section 4.5, we analyse the theoretical performance, discuss the computational complexity of the proposed approach and evaluate both in terms of our Research specifications.

4.1 Assumptions

To begin our study of adversarial thresholding semi-bandit problems, we modify the approach as set out by Uchiya *et al.*, [72] for adversarial multi-armed bandit problems with multiple plays to thresholding bandits. Whilst the case where the k players do not communicate is also considered, we assume that the k players do co-operate with each other and no two players choose the same arm in any round. This allows the multi-play and semi-bandit setting to be interchangeable and leads to the following assumption,

Assumption 1. *The adversarial thresholding semi-bandit player selects* $1 \leq k_t \leq K$ *distinct arms for rounds* $t = 1, \ldots, T$.

We can assume that the player has the capacity to only observe k out of a possible K arms every round. The player competes against an oblivious adversary that has knowledge of a $K \times T$ feedback matrix $\mathcal{F} = (x_{t,i})$ for $t = 1, \ldots, T$ and $i = 1, \ldots, K$. In the event that the number of arms exceeding threshold each round is known, the player is making the following assumption,

Assumption 2. *The adversary generates feedback exceeding threshold for exactly* $1 \leq k \leq K$ *arms each round.*

Assumption 2 enables the player to guarantee the existence of a unique set of k arms $I_t \in \mathbb{C}([K], k)$, such that the cumulative sum of the feedback (gain) returned by the adversary is maximised. This leads to the following definition.

Definition 3. *The cumulative gain from playing the adversarial thresholding bandit game with multiple plays over a sequence of* T *rounds and for a set of* k *arms* I_t *is defined as,*

$$G := \sum_{t=1}^{T} \frac{1}{\Delta_t} \sum_{i \in I_t} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\},$$

for the player randomly choosing a sequence of arms $I_1, \ldots, I_T$.

Under Assumption 2, we have that $q_t = k_t = k$, recalling that q_t denotes the number of arms exceeding threshold in round t, giving $\Delta_t = 1$. It can be seen that the adversarial thresholding bandit problem with multiple plays is a generalisation of semi-bandits

or multi-play bandit problems when $\theta = 0$. Performance in classical semi-bandits and multi-play bandit problems is measured against the best fixed policy in hindsight. Hence the expected regret for the adversarial thresholding bandit problem with multiple plays is,

$$\begin{aligned}\overline{\mathcal{R}}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \max_{I'_t \in \mathbb{C}([K],k)} \sum_{i \in I'_t} x_{t,i} - G \,\middle|\, I_1, \ldots, I_T\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{i \in I^*_t} x_{t,i} - G \,\middle|\, I_1, \ldots, I_T\right]. \end{aligned} \tag{4.1}$$

where I^*_t denotes the optimal set of k arms, G is taken from Definition 3 and the expectation is taken over the randomness in selecting the sequence of subsets $I_1, \ldots, I_T$. Since the player has to decide which of the k out of $[K]$ arms to choose each round, the policy space is of size $\binom{K}{k}^T$. If we let $k_t = q_t = k$, then from Theorem 2 we have,

$$H_\theta \geq \frac{2}{1-\theta}\left[k(T-1) - \sum_{t=1}^{T-1} Q_t\right], \tag{4.2}$$

where $Q_t = \sum_{i \in [K]} \mathbb{I}\{y_{t,i} > \theta\}\mathbb{I}\{y_{t+1,i} > \theta\}$. The hardness of the adversarial thresholding semi-bandit problem when Assumption2 holds in (4.2), reflects the frequency of threshold boundary crossings by balancing the total number of feedback exceeding threshold with the total number of consecutive rounds with feedback exceeding threshold from each arm.

This section introduces two assumptions for developing algorithms for the adversarial thresholding semi-bandit problem, with the purpose of fulfilling Research specifications 1 and 3. Assumption 1 is critical for the equivalence of multi-player bandits and the semi-bandit setting. Under Assumption 2, the adversarial thresholding semi-bandit player can concentrate on finding which of k arms exceed threshold each round. The more frequently those k arms cross the threshold boundary, the more difficult the task is for the player.

4.2 Multi-play arm pulling

We now discuss the key procedure used in [72] to select k arms from K, according to a probability distribution over all arms. The **DepRound** algorithm was developed

for combinatorial optimisation problems, where weights associated with edges on a bipartite graph are rounded to one or zero. The general framework for the algorithm is described in this section.

Recall from Chapter 2 that fundamental to the **Exp3.M** algorithm is the need to select a subset of k arms uniformly from a distribution over all K arms. This is done by the **DepRound** algorithm [33], given in Algorithm 8, The **DepRound** algorithm was

Algorithm 8 DepRound

Input: Natural number K, subset size $k \leq K$, $(p_{t,1}, \ldots, p_{t,K})$ such that $\sum_{i=1}^{K} p_{t,i} = k$.

1: Let $p'_{t,i} = p_{t,i}$ for $i \in [K]$
2: **while** $\exists\, i \,|\, 0 < p'_{t,i} < 1$ **do**
3: Select distinct i, j such that

$$0 < p'_{t,i} < 1$$
$$0 < p'_{t,j} < 1$$

4: Let $\alpha = \min\{1 - p'_{t,i}, p'_{t,j}\}$ and $\beta = \min\{p'_{t,i}, 1 - p'_{t,j}\}$.
5: Let

$$(p'_{t,i}, p'_{t,j}) = \begin{cases} (p'_{t,i} + \alpha, p'_{t,j} - \alpha) & \text{with probability } \frac{\beta}{\alpha+\beta}, \\ (p'_{t,i} - \beta, p'_{t,j} + \beta) & \text{with probability } \frac{\alpha}{\alpha+\beta}. \end{cases}$$

Output: $I_t = \{i \,|\, p'_{t,i} = 1, i = 1, \ldots, K\}$.

developed as a randomised rounding method for fractional vectors defined on the edge-sets of bipartite graphs. For a bipartite graph $G(A, B, E)$, where A and B represent the partitioned sets and E is the edge set. For a pair of vertices $a \in A$ and $b \in B$, an edge $(a, b) \in E$ has an associated weight $p_{t,(a,b)} \in [0, 1]$. The procedure iterates through at most $|E|$ steps, modifying proxy-weights $p'_{t,(a,b)}$ until all weights have been rounded to either one or zero. Since the sum of weights $\sum_{(a,b)\in E} p_{t,(a,b)} = k$, the algorithm rounds exactly k edges to one and $|E| - k$ edges to zero.

DepRound begins by setting all proxy-weights equal to weights, $p'_{t,(a,b)} = p_{t,(a,b)}$ for each edge and satisfies the following two statements: (i) Every edge $(a, b) \in E$ has proxy-weight $p'_{t,(a,b)} \in [0, 1]$ and (ii) an edge is *rounded* if the proxy-weight equals one or zero, $p'_{t,(a,b)} \in \{0, 1\}$ and *floating* otherwise, $p'_{t,(a,b)} \in (0, 1)$. Once an edge is labelled as rounded, it is always rounded. While the floating set of edges is non-empty, two variables α and β are defined for two distinct edges, with the objective of performing a depth-first search to find the maximal path in the sub-graph of floating

edges, (see Gandhi *et al.* [33], for technical details). The values of $p'_{t,(a,b)}$ are then probabilistically incremented towards either one or zero by either α or β each step. In their paper, Gandhi *et al.* present a randomised rounding algorithm, proving that $\mathbb{P}[p'_{t,(a,b)} = 1] = p_{t,(a,b)}$ and $\sum_{(a,b)\in E} p'_{t,(a,b)} = k$.

4.3 Multi-play Exp3 for thresholding bandits

In this section we introduce some of the salient properties of **Exp3.M** proposed by Uchiya *et al.* [72], which extends adversarial multi-armed bandits to multiple play problems. We then present the **T-Exp3.M** algorithm, which extends adversarial multi-armed bandits with multiple plays to thresholding bandit problems.

4.3.1 Multi-play Exp3

As with the classical **Exp3** [10], the pulling policy for **Exp3.M** is governed by uniformly sampling from a probability distribution $p_t = (p_{t,1}, \ldots, p_{t,K})$ and unbiased estimates $\hat{x}_{t,i}$ of rewards are calculated, where $\hat{x}_{t,i} = x_{t,i}/p_{t,i}$ if arm i is chosen and observed exceeding threshold, otherwise $\hat{x}_{t,i} = 0$. *Dependent Rounding* [33] (see [72] for details) enables k distinct arms to be uniformly drawn, each with probability $p_{t,i}$ with the condition that $\sum_{i=1}^{K} p_{t,i} = k$. Uchiya *et al.* ensure that the probability associated with each arm remains valid by truncating weights that become too large, according to a critical value α_t. If all K weight are,

$$w_{t,i} < \left(\frac{K - \gamma k}{kK(1-\gamma)}\right) \sum_{i=1}^{K} w_{t,i} \quad \text{for } i = 1, \ldots, K, \tag{4.3}$$

then no weight requires modification and a truncation set $S_t = \emptyset$. If (4.3) does not hold then α_t is determined such that,

$$\frac{\alpha_t}{\sum_{w_{t,i}\geq\alpha_t} \alpha_t + \sum_{w_{t,i}<\alpha_t} w_{t,i}} = \frac{K - \gamma k}{kK(1-\gamma)}. \tag{4.4}$$

A truncation set S_t, is populated with arms of weights at least as large as α_t, where $S_t = \{i \in [K] \,|\, w_{t,i} \geq \alpha_t\}$ and modified weights $w'_{t,i}$ are set to equal α_t for $i \in S_t$.

The remaining arms retain their current weight, $w'_{t,i} = w_{t,i}$ for $i \in [K]\backslash S_t$, where $[A]\backslash B$ denotes the set of arms in A and not in B. The probability of all arms is calculated,

$$p_{t,i} = k\left((1-\gamma)\frac{w'_{t,i}}{\sum_{j=1}^{K} w'_{t,j}} + \frac{\gamma}{K}\right) \quad \text{for } i = 1, \ldots, K. \tag{4.5}$$

The **DepRound** algorithm proposed in [33] is then provided with input parameters $p_{t,1}, \ldots, p_{t,K}$ and k and then outputs a set of k distinct arms I_t.

The reward $x_{t,i}$ for $i \in I_t$ is revealed to the player and she computes an unbiased estimator $\hat{x}_{t,i}$ if arm i is observed exceeding threshold. If arm i is observed within threshold then no reward is received.

The weights of arms in the set $\{i \in [K] \,|\, x_{t,i} > \theta\} \cap I_t \backslash S_t$ are increased by a factor of $\exp(\gamma k \hat{x}_{t,i}/K)$. The weights of arms belonging to the set $i \in \{i \in [K] \,|\, x_{t,i} \leq \theta\} \cap I_t \backslash S_t$ are updated by a factor of $\theta \log(\beta)$, where β is a parameter that will be discussed during analysis of the algorithm.

In the first round, all weights are initialised equal and the probability of arm i being selected is $\mathbb{P}[i \in I_t] = k/K$, under the assumption that arms are independent. If we let the optimal subset of k arms be denoted by I_t^*, where $x_{t,i} > \theta$ for every $i \in I_t^*$ then the probability that a randomly selected subset not containing any arms exceeding threshold is given by,

$$\mathbb{P}\big[I_t \neq I_t^*\big] = \frac{\binom{[K-k]}{k}}{\binom{K}{k}}, \tag{4.6}$$

under Assumption 2. Hence the probability of n out of the k chosen arms exceeding threshold when choosing k arms at random from K is given by,

$$\mathbb{P}\big[|A_t| = n\big] = \frac{\binom{[K-k]}{k-n}\binom{[k]}{n}}{\binom{K}{k}}, \tag{4.7}$$

where $|A_t| = |I_t \cap I_t^*|$ denotes the number of arms belonging to the subset of selected arms that are exceeding threshold in round t. The derivation of (4.7) only requires standard techniques in combinatorics. Any reasonable player wants to improve their chances over playing randomly.

It is noted that since we define $\theta \in (0,1)$ and $x_{t,i} \in [0,1]$ for $t = 1,\ldots,T$ and $i = 1,\ldots,K$ then we have,

$$(1-\theta)kT \leq \sum_{t=1}^{T}\sum_{i\in I_t^*} x_{t,i} \leq kT,$$

under Assumption 2. Hence, the regret of an algorithm when Assumption 2 holds is,

$$\overline{\mathcal{R}}_T \leq kT - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} \,\middle|\, I_1,\ldots,I_T\right].$$

4.4 The T-Exp3.M algorithm

We propose an extension of the adversarial multi-armed bandit with multiple plays for the thresholding bandit problem, called *thresholding exponentially weighted exploration and exploitation with multiple plays* (**T-Exp3.M**). We ensure that the subset of k distinct arms chosen each round is sampled uniformly from a weighted distribution over all arms by constraining weights that become too large as in [72] and using the approach set out in [33]. The significant difference between **T-Exp3.M** and **Exp3.M** is obviously the thresholding element. The objective for **Exp3.M** is to pull a selection of k arms each round that will ultimately achieve maximum cumulative reward over T rounds. Under Assumption 2, this amounts to the same problem, but in **T-Exp3.M** the player receives no reward for pulling arms within threshold and consequently cannot infer any information about arms that are exceeding threshold. Indeed, the only course of action the player can take is to penalise the weights of those arms by a fixed amount. The pseudo-code for **T-Exp3.M** is given in Algorithm 9 and a schematic diagram of the thresholding exponentially weighted exploration and exploitation model with multiple plays is illustrated in Figure 4.1.

4.5 Analysis of T-Exp3.M

It can be seen that theoretically the expected cumulative regret of **T-Exp3.M** is at least $\mathcal{O}\left(\sqrt{kKT\log\left(\frac{K}{k}\right)}\right)$ since the problem is more challenging than MABs with multiple plays. The proof follows similar lines as that given in Theorem 2 and Corollary 1 of [72] and extends Theorem 3.1 from [10].

Algorithm 9 T-Exp3.M

Input: Number of arms K, subset size $k \leq \exp(-1)K$, $w_{1,i} = 1$ for $i = 1, \ldots, K$, threshold $\theta \in (0, 1)$, $\gamma \in (0, 1]$, $0 < \beta < \exp(-\theta)$.

for $t = 1, \ldots, T$ **do**
  if $\operatorname{argmax}_{j \in K} w_{t,j} \geq \frac{K - \gamma k}{kK(1-\gamma)} \sum_{i=1}^{K} w_{t,i}$ **then**
    Determine α_t such that,

$$\frac{\alpha_t}{\sum_{w_{t,i} \geq \alpha_t} \alpha_t + \sum_{w_{t,i} < \alpha_t} w_{t,i}} = \frac{K - \gamma k}{kK(1-\gamma)}.$$

    Let $S_t = \{i \in [K] \,|\, w_{t,i} \geq \alpha_t\}$.
    Let $w'_{t,i} = \alpha_t$ for $i \in S_t$.
  else
    Let $S_t = \emptyset$.
  Let $w'_{t,i} = w_{t,i}$ for $i \in [K] \backslash S_t$.
  Let

$$p_{t,i} = k\left((1-\gamma)\frac{w'_{t,i}}{\sum_{j=1}^{K} w'_{t,j}} + \frac{\gamma}{K}\right).$$

  Let $I_t = \textbf{DepRound}(k, (p_{t,1}, \ldots, p_{t,K}))$.
  Adversary generates feedback $x_{t,i}$ for $i \in [K]$.
  For $i = 1, \ldots, K$

$$\hat{x}_{t,i} = \begin{cases} \frac{x_{t,i}}{p_{t,i}} & \text{if } i \in \{i \in [K] \,|\, x_{t,i} > \theta\} \cap I_t, \\ 0 & \text{otherwise.} \end{cases}$$

$$w_{t+1,i} = \begin{cases} w_{t,i} \exp\left(\frac{\gamma k}{K} \hat{x}_{t,i}\right) & \text{if } i \in \{i \in [K] \,|\, x_{t,i} > \theta\} \cap I_t \backslash S_t, \\ w_{t,i} \exp\left(\theta \log(\beta)\right) & \text{if } i \in \{i \in [K] \,|\, x_{t,i} \leq \theta\} \cap I_t \backslash S_t, \\ w_{t,i} & \text{otherwise.} \end{cases}$$

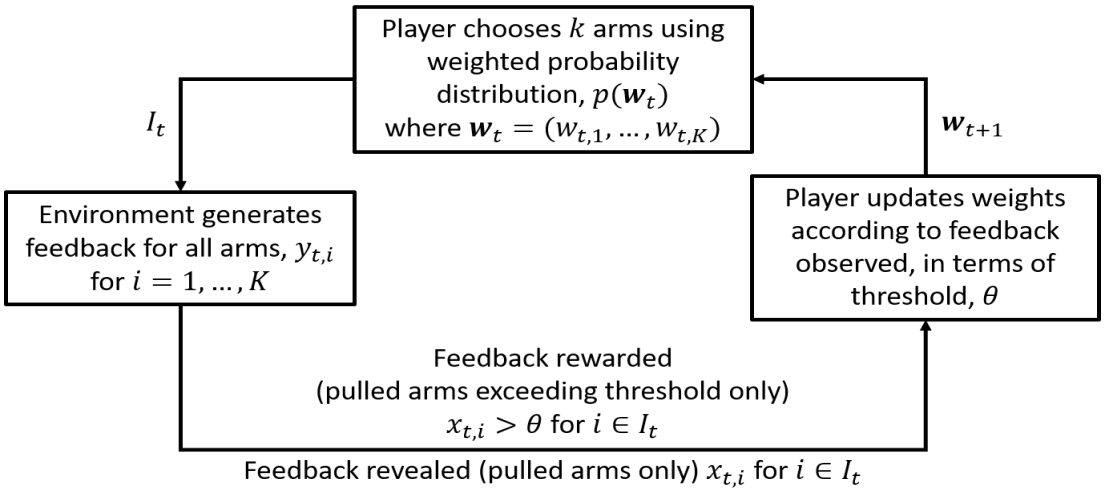


Figure 4.1: A schematic diagram of the thresholding exponentially weighted exploration and exploitation model with multiple plays.

The analysis draws attention to the relationships between k, K, T and the threshold θ by considering how the cumulative sum of weights over all arms evolves as t approaches the time horizon T.

Theorem 3. *For any integer $K \geq k \geq 2$, real numbers $\theta \in (0,1)$, $\beta \in (0,1]$ and $\gamma \in [0,1]$,*

then,

$$\overline{\mathcal{R}}_T \leq (e-1)\gamma G_{max-k,\theta} + \frac{K(1+\beta T)}{\gamma(1-\theta)} \log\left(\frac{K}{k}\right)$$

holds for any assignment of rewards and for any $T > 0$, $G_{max-k,\theta} = \mathbb{E}\Big[\sum_{t\in[T]} \sum_{i\in I_t^*} \hat{x}_{t,i}\Big]$.

Proof of Theorem 3. From the proof of Theorem 2 in [72] we denote W_t and W_t' as $\sum_{i=1}^K w_{t,i}$ and $\sum_{i=1}^K w'_{t,i}$, respectively. For ease of notation, we denote the following,

$$A = \{i \in [K] \,|\, x_{t,i} > \theta\} \cap I_t \backslash S_t, \tag{4.8}$$

$$B = \{i \in [K] \,|\, x_{t,i} \leq \theta\} \cap I_t \backslash S_t, \tag{4.9}$$

$$C = I_t^c \cup I_t \cap S_t, \tag{4.10}$$

where the sets A, B and C represent the set of arms selected by **DepRound** with non-truncated weights $(I_t \backslash S_t)$ that have feedback exceeding threshold, set of arms selected by **DepRound** with non-truncated weights, having feedback within threshold and

finally, the set of arms not selected or arms selected by **DepRound** but with truncated weights, respectively. Then, since $w_{T+1,i}$ can be computed for any arm in round T, we have for any round $t = 1, \ldots, T$,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i \in A} \frac{w_{t+1,i}}{W_t} + \sum_{i \in B} \frac{w_{t+1,i}}{W_t} + \sum_{i \in C} \frac{w_{t+1,i}}{W_t} \\
&= \sum_{i \in A} \frac{w_{t,i}}{W_t} \exp\left(\frac{\gamma k}{K} \hat{x}_{t,i}\right) + \sum_{i \in B} \frac{w_{t,i}}{W_t} \exp(\theta \log(\beta)) + \sum_{i \in C} \frac{w_{t+1,i}}{W_t} \\
&\leq 1 + \beta \exp(\theta) + \frac{W'_t}{W_t} \sum_{i \in A} \frac{w'_{t,i}}{W'_t} \left[1 + \frac{\gamma k}{K} \hat{x}_{t,i} + \frac{(e-2)\gamma^2 k^2}{K^2} \hat{x}^2_{t,i}\right] \qquad (4.11) \\
&\leq 1 + \beta \log(\theta) + \sum_{i \in A} \frac{\frac{p_{t,i}}{k} - \frac{\gamma}{K}}{1 - \gamma} \left[\frac{\gamma k}{K} \hat{x}_{t,i} + \frac{(e-2)\gamma^2 k^2}{K^2} \hat{x}^2_{t,i}\right] \qquad (4.12) \\
&\leq 1 + \beta \exp(\theta) + \frac{\gamma}{K(1-\gamma)} \sum_{i \in A} x_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)} \sum_{i=1}^{K} \hat{x}_{t,i}, \qquad (4.13)
\end{aligned}$$

since the sets B and C are both subsets of $[K]$ then $\sum_{i \in B} w_{t,i}/W_t \leq 1$, $\sum_{i \in C} w_{t,i}/W_t \leq 1$. Since $p_{t,i} \geq \gamma k/K$ and $x_{t,i} \leq 1$ then $\gamma k \hat{x}_{t,i}/K \leq 1$ and the inequality $\exp(a) \leq 1 + a + (e-2)a^2$ holds for $a \leq 1$ then (4.11) also holds. Since $1 + a \leq \exp(a)$, for $i \in S_t$ we have $w'_{t,i} = \alpha_t \leq w_{t,i}$ and for $i \in [K] \backslash S_t$ we have $w'_{t,i} = w_{t,i}$ then $W'_t/W_t \leq 1$ holds in (4.12). Since $\hat{x}_{t,i} p_{t,i} = x_{t,i} \leq 1$ for $i \in A$ and zero otherwise, and $\sum_{i \in A} \hat{x}_{t,i} \leq \sum_{i \in [K]} \hat{x}_{t,i}$, since $A \subseteq [K]$ in (4.13). Taking logarithms on both sides of (4.13) and summing over $t \in [T]$ gives,

$$\begin{aligned}
\log\left(\frac{W_{T+1}}{W_1}\right) &\leq T \log\left(1 + \beta \exp(\theta)\right) + \frac{\gamma}{K(1-\gamma)} \sum_{t=1}^{T} \sum_{i \in A} x_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)} \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{x}_{t,i} \\
&\leq \frac{\beta T}{1-\theta} \log\left(\frac{K}{k}\right) + \frac{\gamma}{K(1-\gamma)} \sum_{t=1}^{T} \sum_{i \in A} x_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)} \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{x}_{t,i},
\end{aligned} \qquad (4.14)$$

where $\log(a+b) = \log(a) + \log(1 + b/a)$ for $a = 1 + \beta \exp(\theta)$, and b represents the last two terms in (4.13), $1/(1 + \beta \exp(\theta)) \leq 1$, $\exp(\theta) \leq 1/(1-\theta)$ for $\theta < 1$, $\log(a) \leq a$ for $a > 0$ and $\log(K/k) \geq 1$ for $\exp(-1)K \geq k$.

Since $w_{t,i} > 0$ for $i \in [K]$ and $t \in [T]$, we can now give a lower bound on the LHS of (4.14). Consider that the optimal set of arms $I_t^* \subset \mathbb{C}([K], k)$ maximises the received reward $\sum_{i \in I_t^*} \sum_{t \in [T]} x_{t,i}$ such that $\forall i \in I_t^*$, $x_{t,i} > \theta$. Then,

$$\begin{aligned}
\log\left(\frac{W_{T+1}}{W_1}\right) &\geq \log\left(\frac{\sum_{i \in I_t^*} w_{T+1,i}}{W_1}\right) \\
&\geq \frac{1}{k} \sum_{i \in I_t^*} \log(w_{T+1,i}) + \log\left(\frac{k}{K}\right) \qquad (4.15) \\
&= \frac{\gamma}{K} \sum_{t=1}^{T} \sum_{i \in I_t^*} \hat{x}_{t,i} + \log\left(\frac{k}{K}\right), \qquad (4.16)
\end{aligned}$$

where (4.15) uses from [72], the fact that,

$$\sum_{i \in I_t^*} w_{T+1,i} \geq k \left(\prod_{i \in I_t^*} w_{T+1,i} \right)^{\frac{1}{k}},$$

and we recall from the update rule $w_{t+1,i} = w_{t,i} \exp(\gamma k \hat{x}_{t,i}/K)$ for $i \in A$, in (4.16) we have,

$$w_{T+1,i} = \exp\left(\frac{\gamma k}{K} \sum_{t=1}^{T} \sum_{i \in A} \hat{x}_{t,i}\right).$$

Combining (4.14) and (4.16) gives,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i \in I_t^* \setminus S_t} \hat{x}_{t,i} + \frac{K}{\gamma} \log\left(\frac{k}{K}\right) \\
&\leq \frac{\beta K T}{\gamma(1-\theta)} \log\left(\frac{K}{k}\right) + \frac{1}{1-\gamma} \sum_{t=1}^{T} \sum_{i \in A} x_{t,i} + \frac{(e-2)\gamma k}{K(1-\gamma)} \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{x}_{t,i},
\end{aligned}$$

since $|I_t^* \setminus S_t| \leq |I_t^*|$, $\sum_{t=1}^{T} \sum_{i \in A} \hat{x}_{t,i} + \sum_{t=1}^{T} \sum_{i \in B \cup S_t} x_{t,i} \geq \sum_{t=1}^{T} \sum_{i \in I_t} x_{t,i}$ holds trivially, if $A = I_t^* \setminus S_t$ then $B = \emptyset$ and since $\sum_{t=1}^{T} \sum_{i \in I_t^* \cap S_t} x_{t,i} \leq \frac{1}{1-\gamma} \sum_{t=1}^{T} \sum_{i \in I_t^* \cap S_t} x_{t,i}$ holds trivially then,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{i \in I_t^* \setminus S_t} \hat{x}_{t,i} + \sum_{t=1}^{T} \sum_{i \in I_t^* \cap S_t} x_{t,i} + \frac{K}{\gamma} \log\left(\frac{k}{K}\right) \\
&\leq \frac{1}{1-\gamma} G_{T-Exp3.M} + \frac{(e-2)\gamma k}{K(1-\gamma)} \sum_{t=1}^{T} \sum_{i=1}^{K} \hat{x}_{t,i} + \frac{\beta K T}{\gamma(1-\theta)} \log\left(\frac{K}{k}\right). \qquad (4.17)
\end{aligned}$$

Taking expectation on both sides of (4.17) gives,

$$G_{max-k,\theta} \leq \frac{1}{1-\gamma}\mathbb{E}\Big[G_{T-Exp3.M}\Big] + \frac{(e-2)\gamma k}{K(1-\gamma)}\sum_{t=1}^{T}\sum_{i=1}^{K} x_{t,i} + \frac{K(1+\beta T)}{\gamma(1-\theta)}\log\left(\frac{K}{k}\right),$$

where $G_{T-Exp3.M}$ denotes the cumulative gain of the **T-Exp3.M** algorithm, by Definition 3, $G_{max-k,\theta} = \mathbb{E}\Big[\sum_{t\in[T]}\sum_{i\in I_t^*}\hat{x}_{t,i}\Big]$ and the expectation is taken over the randomness in choosing a set of k arms each round. The proof is completed by stating that $\sum_{t\in[T]}\sum_{i\in[K]} x_{t,i} \leq \frac{K}{k}G_{max-k,\theta}$. □

Corollary 3.1. *Set $\beta = 1/T$ and,*

$$\gamma = \min\left\{1, \sqrt{\frac{2K}{(e-1)(1-\theta)kT}\log\left(\frac{K}{k}\right)}\right\},$$

*then an upper bound on the regret of **T-Exp3.M** is,*

$$\overline{\mathcal{R}}_T \leq 2\sqrt{2(e-1)}(k+1)\sqrt{\frac{KT}{(1-\theta)k}\log\left(\frac{K}{k}\right)},$$

and holds for any $T > 0$ and any assignment of rewards.

Corollary 3.1 is achieved by suitably tuning an upper bound for γ, referred to as an *egalitarian* parameter in [10]. The proof of corollary 3.1 has the same structure as that in [10] and [72].

The upper bound of $\mathcal{O}\Big(\sqrt{\frac{kKT}{1-\theta}\log\big(\frac{K}{k}\big)}\Big)$ given in Corollary 3.1 generalises the upper bound $\mathcal{O}(\sqrt{kKT\log(K/k)})$, given for the **Exp3.M** algorithm in [72]. When $\theta = 0$, the thresholding element of ATSBP is discarded and the players' objective reverts to maximising the cumulative feedback from pulling the k best arms each round. When $\theta = 0$ and $k = 1$, Corollary 3.1 resolves to $\mathcal{O}(\sqrt{KT\log(K)})$, which matches the upper bound for **Exp3**. The problem grow asymptotically more difficult as the threshold size increases and becomes degenerate when $\theta = 1$.

However, there are special cases of the adversarial multi-player bandit that still fits the adversarial thresholding semi-bandit model. Considering $\theta = 0$, the only arms not exceeding threshold are those with feedback $y_{t,i} = 0$. Under Assumption 2, there are

exactly k arms with feedback $y_{t,i} > 0$ and $K - k$ arms with $y_{t,i} = 0$ each round. In this scenario and when $k = 1$, the player is tasked with finding the only arm out of K, revealing feedback each round.

The egalitarian parameter γ, dictates how much the player wants to choose selected arms $i \in I_t$, either according to $p_{t,i}$ or uniformly. Playing uniformly gives every arm an opportunity to be observed and can be an advantage if the adversary is playing completely randomly. However, if this is not the case, the player can utilise the history of previous observations for arm i, as a guide for playing it again in future. We consider $0 < \beta < \exp(-\theta)$ (not to be confused with the parameter used in **DepRound**) to be a *penalty* parameter, which costs the player for pulling an arm that is not exceeding threshold.

The time complexity of our **T-Exp3.M** algorithm matches the per-round $\mathcal{O}(K(\log(k)+1))$ of **Exp3.M**, since to obtain the k largest weights each round, a complete binary tree (also known as a *heap*) needs to be constructed in $\mathcal{O}(K)$ time and then extracting the k largest values from the heap in $\mathcal{O}(K \log(k))$ time. The **DepRound** procedure is the bottleneck in our algorithm and will be addressed in Chapter 7. Since the player only receives k rewards each round, the storage requirements of **T-Exp3.M** are $\mathcal{O}(k)$.

4.6 Summary

In this chapter, we introduced the thresholding exponentially weighted exploration and exploitation with multiple plays (**T-Exp3.M**) algorithm for the adversarial thresholding semi-banidt problem, assuming exactly $k \leq \exp(-1)K$ arms exceed a pre-set threshold each round. The algorithm manages a set of weights over all arms, where the weight of an arm observed exceeding threshold increases in relation to the feedback received. Arms observed within threshold are estimated to have zero feedback. This preserves the unbiased nature of the feedback estimator and models the fact that the player learns nothing about another arm that is exceeding threshold from pulling an arm that does not. The weight of an arm observed within threshold is decreased by a fixed amount. Each round k arms are pulled, chosen uniformly according to an exponentially weighted probability distribution over all arms.

We prove an upper bound on the regret for **T-Exp3.M** of order $\mathcal{O}\Big(\sqrt{\frac{kKT}{1-\theta}\log\big(\frac{K}{k}\big)}\Big)$, which generalises that of **Exp3.M** for adversarial bandits with multiple plays and **Exp3** for adversarial bandits. Due to inherent assumptions on adversarial bandits with multiple plays, provable guarantees on performance can only be given for problems where not too many arms exceed threshold each round.

Our approach has computational cost, matching the $\mathcal{O}(K(\log(k)+1))$ and $\mathcal{O}(k)$ of **Exp3.M** in both time complexity and storage, respectively. From this, we demonstrate that **T-Exp3.M** achieves finite regret that converges towards the optimal ATSBP pulling policy over time, for certain sequences of feedback and fulfills Research specification 1. Since the computational requirements of the algorithm are the same as those for **Exp3.M** then specification 3 is partially met, bu may be improved.

We note that the use of **DepRound** to choose k arms according to their associated weighted probabilities each round significantly impacts on the complexity of the algorithm and will be addressed in Chapter 7. However, we also reference the fact that the only restriction on the number of queries a player can make on the environment generating rewards depends on k, the number of arms being selected each round. In Chapter 5, we extend the adversarial thresholding semi-bandit problem to label efficient prediction and consider ATSBP where the player can look at a certain number of rewards throughout the time horizon and must decide whether the query an arm or not.

Chapter 5

Label efficient adversarial thresholding semi-bandits

We now combine thresholding semi-bandits with a variant of prediction with expert advice, developed to provide a theoretical framework for problems relating to repeated games (see [37] and [16] for further details). Label efficient (LE) prediction as proposed by Helmbold and Panizza [39], describes a sequential game where a player predicts the next set of elements in a sequence. The player is not made aware of the true value of the next element set or the accuracy of their prediction unless they specifically ask for it and they are only allowed to ask a certain number of times.

It was shown in [16] and [37] that for a player (referred to as a *forecaster*) to perform optimally, they must utilise randomisation. This is in line with all non-stochastic sequential decision-making problems. In the classical LE prediction problem, a sequence of Bernoulli random variables $B_1, \ldots, B_T$ is used to decide whether to query the feedback from all K arms in some round t, where $\mathbb{P}\big[B_t = 1\big] = \epsilon$ for some $\epsilon \in [0, 1]$. The player makes a prediction by pulling arm I_t in round t, the environment generates feedback x_{t,I_t} and the player accrues some loss l_{t,I_t}, associated with their prediction. However, the player is not made immediately aware of this. If $B_t = 0$, the player receives no information about how well they performed in that round. If $B_t = 1$ then the player is given the loss $l_{t,i}$ for $i \in [K]$.

Blackwell and Hannan discovered that the lower regret (referred to as *excess cumulative loss*) bound for any optimal strategy was $\Omega(\sqrt{T})$. This was refined by Cesa-Bianchi, Freund, Haussler, Schapire and Warmuth [22] to $\Omega(\sqrt{T\log(K)})$.

The *label efficient exponentially weighted average forecaster* proposed in [23] is state-of the-art for LE prediction and achieves a regret bound of $\mathcal{O}\left(T\sqrt{\frac{1}{m}\log\left(\frac{K}{\delta}\right)}\right)$ with probability at least $1-\delta$, where $m \leq T$ is the maximum number of queries the player can make on the environment and the probability of making a query is given by,

$$\epsilon = \max\left\{0, \frac{m - \sqrt{2m\log(4/\delta)}}{T}\right\}.$$

Hence for $\epsilon \approx m/T$, the upper regret bound in [23] becomes,

$$\overline{\mathcal{R}}_T = \mathcal{O}\left(\sqrt{\frac{T}{\epsilon}\log\left(\frac{K}{\delta}\right)}\right). \tag{5.1}$$

In Section 5.1, we extend label efficient prediction to the adversarial thresholding semi-bandit problem. We decompose the act of pulling an arm and observing feedback into an arm-selection process and a selected-arm observation decision. Typically, the observation decision is made by the player in label efficient prediction, where the player is predicting which arms to pull in the next round. However, we also identify applications of label efficient ATSBP can be used to model situations where the feedback collection process is unpredictable. In Section 5.1.1, we present the *label efficient exponentially weighted exploration and exploitation with multiple plays* algorithm **LET-Exp3.M** and we prove an upper bound on its theoretical performance in Section 5.2.

5.1 Label efficient thresholding bandits

The LE prediction problem has been combined with multi-armed bandits in [35], as an application of routing problems in communication networks. An upper bound of,

$$\overline{\mathcal{R}}_T = \mathcal{O}\left(\sqrt{\frac{KT}{\epsilon}\log\left(\frac{K}{\delta}\right)}\right), \tag{5.2}$$

is achieved by György and Ottucsák modifying the LE prediction problem set up in the following way. An arm is chosen each round and the player accrues cumulative regret compared with playing an optimal strategy. If $B_t = 1$ then player is given access to the reward of the arm pulled only (recalling the bandit setting introduced in Chapter 1). The player then updates the associated weight of the pulled arm accordingly. If $B_t = 0$, the player receives no information.

The work presented in this chapter, not only combines thresholding bandits with LE prediction, it generalises the results given in (5.1) and (5.2), to the semi-bandit setting. Furthermore, we propose an alternative view to the player deciding to query feedback or not. Instead of the player making the decision, we consider the model where the adversary (or environment) decides to allow the feedback to be queried or not. The result remains the same (the player either gets to see some feedback or not), but by this subtle change in perspective, we can consider label efficient thresholding semi-bandits as models of ATSBP where the infrastructure used to collect feedback may be unreliable and incomplete. This is a particular challenge faced by any application collecting sequential information.

5.1.1 The LET-Exp3.M algorithm

In this section we generalise the combination of label efficient prediction with multi-armed bandits to the semi-bandit setting. We reference $B_{t,i}$ as an iid Bernoulli random variable such that $\mathbb{P}\big[B_{t,i} = 1\big] = \epsilon_i$, for arm i pulled in round t. After choosing a set of k arms (experts), the player learns of the rewards $x_{t,i}$ received from pulling arms $i \in I_t$ and hence whether they exceeded threshold or not, if and only if she queries (or is allowed access to) the feedback from each arm pulled, determined by $B_{t,i}$. The biased reward estimate becomes the following,

$$x'_{t,i} = \hat{x}_{t,i} + B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon_i}, \tag{5.3}$$

for $i \in [K]$ and where,

$$\hat{x}_{t,i} = \begin{cases} \frac{x_{t,i}}{p_{t,i}\epsilon_i} & \text{if } i \in I_t, B_{t,i} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The **DepRound** algorithm is used to help the player decide which arms to select each round and the set of k Bernoulli random variables are used to decide which of those k arms are queried. The weights of each arm pulled are updated, depending on the size of $x_{t,i}$ and whether $x_{t,i} > \theta$ or not. The pseudo-code for **LET-Exp3.M** is given in Algorithm 10.

A schematic diagram is given in Figure 5.1, illustrating a decomposition of observed arms into an arm-selection process using **DepRound** and a decision to observe selected arms or not. In label efficient multi-armed bandits, the Bernoulli parameters remain fixed and Figure 5.1 depicts the observation decision being made before the environment *generates* feedback for all arms, recalling that an oblivious adversary may have already generated feedback for every arm in every round and simply be accessing feedback for the relevant round. This shows the decision to observe being made by the player. However, if the Bernoulli parameters represent the reliability of the method recording feedback then the decision to allow feedback to be revealed is made by the environment.

In this section, we extend the combination of label efficient prediction with multi-armed bandits to the adversarial threshold semi-bandit setting. We identify that the process of observing feedback from a pulled arm has been decomposed into an arm-selection process and an observation decision for selected arms. To the best of our knowledge, label efficient bandits have only used to model scenarios where the player makes the decision to query an arm or not. However, we believe that this model can also be used in applications where the observation decision is made by the adversary.

5.2 Analysis of LET-Exp3.M

The general approach for analysing our *label efficient thresholding exponentially weighted exploration and exploitation with multiple plays* algorithm follows that presented in Chapter 4. Lemma 4 is known as a Bernstein concentration inequality, given in [14] and [35]. As our reward estimate is biased, we use a Chernoff bounding technique (see [14] for details) in Lemma 5, to bound the deviation between the cumulative sum of biased reward estimates with the cumulative sum of actual rewards over T rounds, in the multi-player bandit setting. Theorem 6 combines the results given in Lemma 4 and Lemma 5, to bound the theoretical performance of the **LET-Exp3.M** algorithm.

Algorithm 10 LET-Exp3.M

Input: Number of arms K, subset size $k \leq \exp(-1)K$, $w_{1,i} = 1$, $\epsilon_i \in [0,1]$ for $i = 1, \ldots, K$, threshold $\theta \in (0,1)$, $\gamma \in (0,1]$, $0 < \beta < \exp(-\theta)$.

for $t = 1, \ldots, T$ **do**
  if $\operatorname{argmax}_{j \in K} w_{t,j} \geq \frac{K - \gamma k}{kK(1-\gamma)} \sum_{i=1}^{K} w_{t,i}$ **then**
    Determine α_t such that,

$$\frac{\alpha_t}{\sum_{w_{t,i} \geq \alpha_t} \alpha_t + \sum_{w_{t,i} < \alpha_t} w_{t,i}} = \frac{K - \gamma k}{kK(1-\gamma)}.$$

    Let $S_t = \{i \in [K] \,|\, w_{t,i} \geq \alpha_t\}$.
    Let $w'_{t,i} = \alpha_t$ for $i \in S_t$.
  else
    Let $S_t = \emptyset$.
  Let $w'_{t,i} = w_{t,i}$ for $i \in [K] \backslash S_t$.
  Let

$$p_{t,i} = k\left((1-\gamma)\frac{w'_{t,i}}{\sum_{j=1}^{K} w'_{t,j}} + \frac{\gamma}{K}\right).$$

  Let $I_t = \textbf{DepRound}(k, (p_{t,1}, \ldots, p_{t,K}))$.
  Adversary generates feedback $x_{t,i}$ for $i \in [K]$.
  for $i = 1, \ldots, K$, **do**
    Draw Bernoulli random variables $B_{t,i}$ for such that,

$$\mathbb{P}\big[B_{t,i} = 1\big] = \epsilon_i,$$

.

    Observe $B_{t,i} x_{t,i}$.
    Let,

$$x'_{t,i} = \begin{cases} \frac{x_{t,i} + \upsilon}{p_{t,i}\epsilon_i} & \text{if } i \in I_t, B_{t,i} = 1 \\ \frac{\upsilon}{p_{t,i}\epsilon_i} & \text{if } i \notin I_t, B_{t,i} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

    Let,

$$w_{t+1,i} = \begin{cases} w_{t,i} \exp\big(\theta \log(\beta)\big) & \text{if } i \in \{i \in [K] \,|\, x_{t,i} \leq \theta\} \cap I_t \backslash S_t, B_{t,i} = 1, \\ w_{t,i} \exp\Big(\frac{\gamma k}{K} x'_{t,i}\Big) & \text{otherwise.} \end{cases}$$

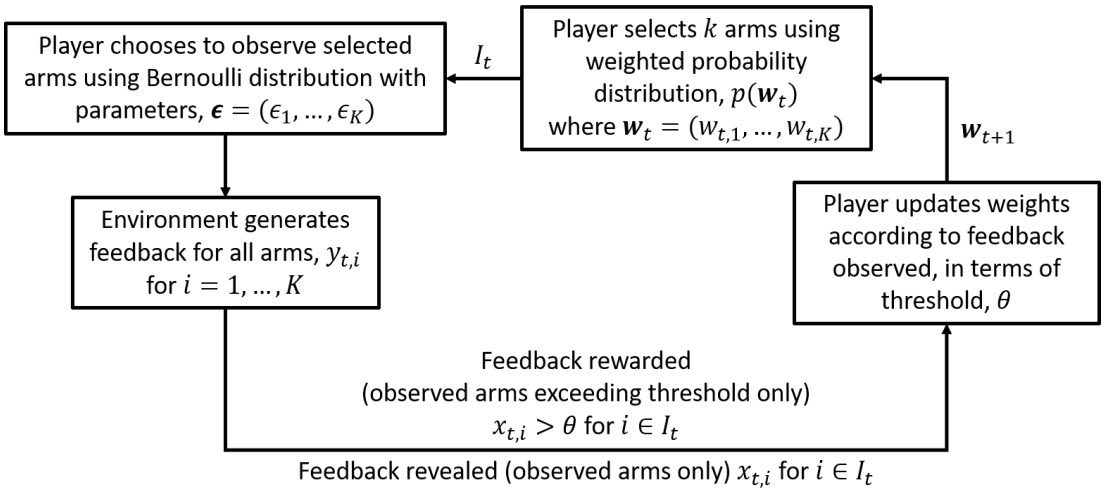


Figure 5.1: A schematic diagram combining label efficient prediction with adversarial thresholding semi-bandits.

Lemma 4 (Bernstein concentration inequality)**.** *Let* $\sum_{i\in I_t} Y_{1,i}, \ldots, \sum_{i\in I_t} Y_{T,i}$ *be a martingale difference sequence where* $\sum_{i\in I_t} Y_{t,i} \in [a, b]$ *with probability one* $\forall t \in [T]$*. Assuming that,*

$$\mathbb{E}\left[\sum_{i\in I_t} Y_{t,i}^2 \,\middle|\, I_1, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\right] \leq \sigma^2,$$

with probability one. Then for any $0 < \epsilon < 1$*,*

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in I_t} Y_{t,i} > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2 T + 2\epsilon(b-a)/3}\right),$$

and,

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in I_t} Y_{t,i} > \sqrt{2\sigma^2 T \log\left(\frac{1}{\delta}\right)} + \frac{2(b-a)}{3}\log\left(\frac{1}{\delta}\right)\right] \leq \delta.$$

The proof of Lemma 4 can be found in [14] and is combined with Lemma 5 to given an upper bound on the regret of **LET-Exp3.M**.

Lemma 5. *For real numbers* $0 < \delta < 1$, $0 < \epsilon_i < 1$, *where* $\epsilon = \min_{i\in[K]} \epsilon_i$ *and* $\sqrt{\frac{1}{\epsilon kKT}\log\left(\frac{2K}{\delta}\right)} \leq \upsilon \leq 1$ *we have,*

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} > \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i} + 4\upsilon kKT\right] \leq \frac{\delta}{2K}.$$

Proof of Lemma 5. From the Chernoff bounding technique [14] we can state for any $a > 0$ and $b > 0$ the following,

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} > \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i} + a\right] \\
&\leq \exp(-ab)\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right]. \qquad (5.4)
\end{aligned}$$

Let $a = 4\upsilon k^2 KT$ and $b = \epsilon\upsilon/4k$ to give,

$$\begin{aligned}
&\exp(-ab)\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right] \\
&= \exp\Big(-\epsilon\upsilon^2 kKT\Big)\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right] \\
&\leq \exp\Big(-\log\Big(\frac{2K}{\delta}\Big)\Big)\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right] \\
&= \frac{\delta}{2K}\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right].
\end{aligned}$$

Hence, (5.4) will be satisfied if we bound,

$$\mathbb{E}\left[\exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right)\right] \leq 1.$$

For any $t \in [T]$ we define the random variable,

$$Z_t = \exp\left(b\left(\sum_{t=1}^{T}\sum_{i\in I_t} x_{t,i} - \sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}\right)\right),$$

which gives,

$$Z_t = Z_{t-1} \exp\left(b\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t} x'_{t,i}\right)\right). \tag{5.5}$$

From (5.3) we have,

$$\begin{aligned} b\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t} x'_{t,i}\right) &\leq b\left(k - \sum_{i\in I_t} \hat{x}_{t,i} + B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\right) \\ &\leq 1 - b\sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon} - b\sum_{i\in I_t} B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon} \\ &\leq 1, \end{aligned}$$

since $b = \epsilon\upsilon/4k < 1$. We can now bound the expectation of the random variable Z_t for all subsequent rounds $t = 2\ldots, T$ such that,

$$\begin{aligned} \mathbb{E}\big[Z_t\big] &= \mathbb{E}\big[Z_t\,|I_{t_1}, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\big] \\ &= Z_{t-1}\mathbb{E}\left[\exp\left(b\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t} x'_{t,i}\right)\right)\right] \\ &\leq Z_{t-1}\mathbb{E}\left[1 + b\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t}\frac{x_{t,i}+B_{t,i}\upsilon}{p_{t,i}\epsilon}\right) + b^2\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t}\frac{x_{t,i}+B_{t,i}\upsilon}{p_{t,i}\epsilon}\right)^2\right], \end{aligned} \tag{5.6}$$

where the expectation is taken over the randomness in choosing arms, I_t denotes the arms $i \in I_t$ in round t and since $\exp(a) \leq 1 + a + a^2$ for $a \leq 1$. Since $p_{t,i} \leq 1$ and $\epsilon \leq 1$ then (5.6) holds. Due to the linearity of expectations we have,

$$\mathbb{E}\left[\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon}\right] = 0,$$

and,

$$\begin{aligned} \mathbb{E}\left[\left(\sum_{i\in I_t} x_{t,i} - \sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon}\right)^2\right] &\leq k\sum_{i\in I_t}\frac{1}{p_{t,i}\epsilon} \\ &\leq \frac{K}{\gamma\epsilon}, \end{aligned}$$

Then from (5.6) we have,

$$\begin{aligned}
\mathbb{E}\big[Z_t\big] &\leq Z_{t-1}\mathbb{E}\left[1 - b\sum_{i\in I_t} B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon} + \frac{b^2K}{\gamma\epsilon}\right] \\
&\quad + Z_{t-1}\mathbb{E}\left[b^2\sum_{i\in I_t} B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\left(2\sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon} - 2\sum_{i\in I_t} x_{t,i} + \sum_{i\in I_t} B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\right)\right] \\
&\leq Z_{t-1}\left[1 - b\sum_{i\in I_t}\frac{\upsilon}{p_{t,i}} + \frac{b^2K}{\gamma\epsilon} + \frac{2b^2\upsilon K}{\gamma} + \frac{b^2\upsilon^2K}{\gamma}\right] \\
&\leq Z_{t-1}\left[1 + \frac{bK}{\gamma}\left(-\upsilon + \frac{b}{\epsilon} + b\upsilon(2+\upsilon)\right)\right], \qquad (5.7)
\end{aligned}$$

where (5.7) holds since,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in I_t} B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\left(\sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon} - \sum_{i\in I_t} x_{t,i}\right)\right] &= \sum_{i\in I_t}\mathbb{E}\left[B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\left(\sum_{i\in I_t}\frac{x_{t,i}}{p_{t,i}\epsilon} - \sum_{i\in I_t} x_{t,i}\right)\right] \qquad (5.8) \\
&\leq \sum_{i\in I_t}\mathbb{E}\left[x_{t,i}B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon}\sum_{i\in I_t}\frac{1}{p_{t,i}\epsilon}\right] \\
&\leq \sum_{i\in I_t}\mathbb{E}\left[B_{t,i}\frac{\upsilon K}{\gamma\epsilon k}\right] \qquad (5.9) \\
&\leq \frac{\upsilon K}{\gamma}
\end{aligned}$$

where (5.8) holds due to the linearity of expectations and (5.9) holds since $x_{t,i} \leq 1$ and $\sum_{\in I_t} 1/p_{t,i} \leq K/\gamma k$. From (5.7) and since $b = \epsilon\upsilon/4k$ we have,

$$\begin{aligned}
-\upsilon + \frac{b}{\epsilon} + b\upsilon(2+\upsilon) &\leq \frac{-3\upsilon + \epsilon\upsilon^2(2+\upsilon)}{4k} \\
&\leq 1,
\end{aligned}$$

since $\upsilon \leq 1$ and $\epsilon \leq 1$. This means that, $\mathbb{E}\big[Z_t\big] \leq Z_{t-1}$. Hence, $\mathbb{E}\big[Z_t\big] \leq \mathbb{E}\big[Z_{t-1}\big]$ and with $\mathbb{E}\big[Z_1\big] \leq 1$, this completes the proof. □

Theorem 6 provides an upper bound on the regret of **LET-Exp3.M** and generalises a combination of the adversarial thresholding semi-bandit and label efficient prediction.

Theorem 6. *For integers $T > 0$ and $K > k \geq 1$ and real numbers $0 < \theta < 1$, $\delta > 0$, $\epsilon > 0$, $0 < \beta < \exp(-\theta)$ and $\upsilon \leq \gamma \leq \frac{1}{2}$ where $\upsilon = \sqrt{\frac{1}{\epsilon k K T} \log\left(\frac{2K}{\delta}\right)}$ then,*

$$\overline{\mathcal{R}}_T \leq \frac{(2e-3)\gamma}{\epsilon} G_{max-k,\theta} + (1+4k)\upsilon k T + \frac{K(1+\beta T)}{\gamma(1-\theta)} \log\left(\frac{K}{k}\right) + \sqrt{\frac{9kT}{2\epsilon} \log\left(\frac{2K}{\delta}\right)} + \frac{k}{\epsilon} \log\left(\frac{2K}{\delta}\right),$$

holds for any assignment of rewards with probability at least $1 - \delta/K$.

The proof of Theorem 6 proceeds by bounding the ratio of the sum of weights between successive rounds, according to the weight update rules for rewards satisfying different conditions.

Proof of Theorem 6. Let $W_t = \sum_{i \in [K]} w_{t,i}$ and $W'_t = \sum_{i \in [K]} w'_{t,i}$ for any round $t \in [T]$ then,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{i \in D} \frac{w_t}{W_t} \exp\left(\theta \log(\beta)\right) + \sum_{i \in [K] \setminus D} \frac{w_t}{W_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) \\
&\leq \beta \exp(\theta) + \sum_{i \in [K] \setminus D} \frac{w_t}{W_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) \\
&\leq \beta \exp(\theta) + \sum_{i \in [K] \setminus D} \frac{w_t}{W_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) && (5.10) \\
&= \beta \exp(\theta) + \frac{W'_t}{W_t} \sum_{i \in [K] \setminus D} \frac{w'_{t,i}}{W'_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) && (5.11) \\
&\leq \beta \exp(\theta) + \sum_{i \in [K] \setminus D} \frac{w'_{t,i}}{W'_t} \left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] && (5.12) \\
&= \beta \exp(\theta) + \sum_{i \in [K] \setminus D} \frac{\frac{p_{t,i}}{k} - \frac{\gamma}{K}}{1-\gamma} \left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \\
&\leq 1 + \beta \exp(\theta) + \frac{\gamma}{K(1-\gamma)} \sum_{i \in [K] \setminus D} x'_{t,i} p_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)} \sum_{i \in [K] \setminus D} x'^2_{t,i} p_{t,i}, && (5.13)
\end{aligned}$$

where $D = \{i \in I_t \setminus S_t \,|\, B_{t,i} = 1, x_{t,i} \leq \theta\}$ and (5.10) holds since $\sum_{i \in D} \frac{w_{t,i}}{W_t} \leq \sum_{i \in [K] \setminus D} \frac{w_t}{W_t} \leq 1$. We have that (5.11) holds from steps 5 and 7 in the algorithm and (5.12) holds since $\exp(a) \leq 1 + a + (e-2)a^2$ for $a \leq 1$ and $W'_t / W_t \leq 1$.

Since $\log(a+b) = \log(a) + \log(1+b/a)$, $1/(1+\beta\exp(\theta)) \leq 1$, $\exp(a) \leq 1/(1-a)$ for $a \leq 1$, $\log(1+a) \leq a$ and $\log(K/k) \geq 1$, for $k \leq \exp(-1)K$. Then,

$$\begin{aligned}\log\left(\frac{W_{t+1}}{W_t}\right) &\leq \frac{\beta}{1-\theta} + \frac{\gamma}{K(1-\gamma)}\sum_{i\in[K]\setminus D} x'_{t,i}p_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)}\sum_{i\in[K]\setminus D} x'^2_{t,i}p_{t,i} \\ &\leq \frac{\beta}{1-\theta}\log\left(\frac{K}{k}\right) + \frac{\gamma}{K(1-\gamma)}\sum_{i\in[K]\setminus D} x'_{t,i}p_{t,i} + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)}\sum_{i\in[K]\setminus D} x'^2_{t,i}p_{t,i}.\end{aligned} \tag{5.14}$$

We now bound the second term in (5.14) to give,

$$\begin{aligned}\sum_{i\in[K]\setminus D} x'_{t,i}p_{t,i} &= \sum_{i\notin I_t} x'_{t,i}p_{t,i} + \sum_{i\in I_t\cap S_t} x'_{t,i}p_{t,i} + \sum_{i\in D_1} x'_{t,i}p_{t,i} + \sum_{i\in D_2} x'_{t,i}p_{t,i} \\ &= \sum_{i\notin I_t} B_{t,i}\frac{\upsilon}{\epsilon} + \sum_{i\in I_t\cap S_t} B_{t,i}\frac{x_{t,i}+\upsilon}{\epsilon} + \sum_{i\in D_1} B_{t,i}\frac{x_{t,i}+\upsilon}{\epsilon} \qquad (5.15) \\ &\leq \sum_{i=1}^{K} B_{t,i}\left(\frac{x_{t,i}\mathbb{I}\{i\in I_t\}+\upsilon}{\epsilon}\right), \qquad (5.16)\end{aligned}$$

where $\mathbb{I}\{a\}$ denotes the indicator function, which equals 1 when condition a is true and 0 when a is false, $D_1 = \{i \in I_t\setminus S_t \,|\, x_{t,i} > \theta\}$ and $D_2 = \{i \in I_t\setminus S_t \,|\, B_{t,i} = 0, x_{t,i} \leq \theta\}$, $I_t\setminus S_t = D \cup D_1 \cup D_2$, (5.15) holds since $x'_{t,i} = 0$ if $B_{t,i} = 0$ and (5.16) holds since and $I_t \cap S_t \cup D_1 \subseteq I_t$.

The last term in (5.14) is bound by,

$$\begin{aligned}\sum_{i\in[K]\setminus D} x'^2_{t,i}p_{t,i} &= \sum_{i=1}^{K} B_{t,i}x'_{t,i}\left(\frac{x_{t,i}\mathbb{I}\{i\in I_t\}+\upsilon}{p_{t,i}\epsilon}\right)p_{t,i} \qquad (5.17) \\ &= \sum_{i=1}^{K} B_{t,i}\frac{x'_{t,i}x_{t,i}\mathbb{I}\{i\in I_t\}}{\epsilon} + \sum_{i=1}^{K} B_{t,i}\frac{\upsilon x'_{t,i}}{\epsilon} \\ &\leq 2\sum_{i=1}^{K} B_{t,i}\frac{x'_{t,i}}{\epsilon}, \qquad (5.18)\end{aligned}$$

where (5.17) holds using the same arguments as (5.16) and (5.18) holds since $x_{t,i} \leq 1$, for $i \in [K]$ and $\upsilon \leq 1$.

Combining (5.14), (5.16) and (5.18) gives,

$$\log\left(\frac{W_{t+1}}{W_t}\right) \leq \frac{\beta}{1-\theta}\log\left(\frac{K}{k}\right) + \frac{\gamma}{K(1-\gamma)}\left(\sum_{i\in I_t} B_{t,i}\frac{x_{t,i}}{\epsilon} + \sum_{i=1}^{K} B_{t,i}\frac{\upsilon}{\epsilon}\right) + \frac{2(e-2)\gamma^2 k}{K^2(1-\gamma)}\sum_{i=1}^{K} B_{t,i}\frac{x'_{t,i}}{\epsilon}. \tag{5.19}$$

Summing (5.19) over $t = 1, \ldots, T$ rounds gives,

$$\log\left(\frac{W_{T+1}}{W_1}\right) \leq \frac{\beta T}{1-\theta}\log\left(\frac{K}{k}\right) + \frac{\gamma}{K(1-\gamma)}\sum_{t=1}^{T}\sum_{i\in I_t} B_{t,i}\frac{x_{t,i}}{\epsilon} + \frac{\gamma}{K(1-\gamma)}\sum_{t=1}^{T}\sum_{i=1}^{K} B_{t,i}\frac{\upsilon}{\epsilon} + \frac{2(e-2)\gamma^2 k}{K^2(1-\gamma)}\sum_{t=1}^{T}\sum_{i=1}^{K} B_{t,i}\frac{x'_{t,i}}{\epsilon}. \tag{5.20}$$

Since $W_{T+1} = \sum_{i\in[K]} w_{T+1,i}$ then we have,

$$\begin{aligned} W_{T+1} &= \sum_{i\in D_T} w_{T,i}\exp(\theta\log(\beta)) + \sum_{i\in[K]\setminus D_T} w_{T,i}\exp\left(\frac{\gamma k}{K}x'_{T,i}\right) \\ &= \sum_{i\in D_T}\exp(\theta\log(\beta))\left(\sum_{i\in D_{T-1}} w_{T-1,i}\exp(\theta\log(\beta)) + \sum_{i\in[K]\setminus D_{T-1}} w_{T-1,i}\exp\left(\frac{\gamma k}{K}x'_{T-1,i}\right)\right) \\ &+ \sum_{i\in[K]\setminus D_T}\exp\left(\frac{\gamma k}{K}x'_{T,i}\right)\left(\sum_{i\in D_{T-1}} w_{T-1,i}\exp(\theta\log(\beta)) + \sum_{i\in[K]\setminus D_{T-1}} w_{T-1,i}\exp\left(\frac{\gamma k}{K}x'_{T-1,i}\right)\right) \\ &= \beta^T\exp(\theta T) + \cdots + \prod_{t=1}^{T}\sum_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right), \end{aligned} \tag{5.21}$$

where $D_t = \{i \in I_t \setminus S_t \,|\, B_{t,i} = 1, x_{t,i} \leq \theta\}$ in round t.

Hence,

$$\begin{aligned}
\log\left(\frac{W_{T+1}}{W_1}\right) &\geq \log\left(\frac{1}{K}\prod_{t=1}^{T}\sum_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right)\right) \\
&= \sum_{t=1}^{T}\log\left(\sum_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right)\right) + \log\left(\frac{1}{K}\right) \\
&\geq \sum_{t=1}^{T}\log\left(k\left(\prod_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right)\right)^{\frac{1}{k}}\right) + \log\left(\frac{1}{K}\right) \qquad (5.22) \\
&\geq \frac{1}{k}\sum_{t=1}^{T}\sum_{i\in[K]\setminus D_t}\frac{\gamma k}{K}x'_{t,i} + \log\left(\frac{k}{K}\right), \qquad (5.23)
\end{aligned}$$

where (5.22) holds using the fact that,

$$\sum_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right) \geq k\left(\prod_{i\in[K]\setminus D_t}\exp\left(\frac{\gamma k}{K}x'_{t,i}\right)\right)^{\frac{1}{k}},$$

and (5.23) holds since $T\log(k) \geq \log(k)$ for $T > 0$. For a player playing an optimal strategy each round such that $i \in I_t^*$ where $I_t^* = \{i \in I_t \,|\, B_{t,i} = 1, x_{t,i} > \theta\}$ and $B_{t,i} = 0$ for $i \notin I_t^*$, then (5.23) gives,

$$\log\left(\frac{W_{T+1}}{W_1}\right) \geq \frac{\gamma}{K}\sum_{t=1}^{T}\sum_{i\in I_t^*}x'_{t,i} + \log\left(\frac{k}{K}\right). \qquad (5.24)$$

Combining (5.24) and (5.20) gives,

$$\begin{aligned}
(1-\gamma)\sum_{t=1}^{T}\sum_{i\in I_t^*}x'_{t,i} + \frac{K(1-\gamma)}{\gamma}\log\left(\frac{k}{K}\right) \leq \frac{\beta KT}{\gamma(1-\theta)}\log\left(\frac{K}{k}\right) + \sum_{t=1}^{T}\sum_{i\in I_t}B_{t,i}\frac{x_{t,i}}{\epsilon} \\
+ \sum_{t=1}^{T}\sum_{i=1}^{K}B_{t,i}\frac{\upsilon}{\epsilon} + \frac{2(e-2)\gamma}{\epsilon}\sum_{t=1}^{T}\sum_{i\in I_t^*}x'_{t,i}. \qquad (5.25)
\end{aligned}$$

Thus,

$$\sum_{t=1}^{T}\sum_{i\in I_t} B_{t,i}\frac{x_{t,i}+\upsilon}{\epsilon} \geq \left(1-\gamma-\frac{2(e-2)\gamma}{\epsilon}\right)\sum_{t=1}^{T}\sum_{i\in I_t^*} x'_{t,i} - \frac{K(1+\beta T)}{\gamma(1-\theta)}\log\left(\frac{K}{k}\right). \tag{5.26}$$

We now let,

$$Y_{t,i} = B_{t,i}\frac{x_{t,i}+\upsilon}{\epsilon} - (x_{t,i}+\upsilon),$$

for $t \in [T]$ and each $i \in I_t$. Since,

$$\mathbb{E}\big[Y_{t,i} \,|\, I_1, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\big] = 0, \tag{5.27}$$

then Y_t is a martingale difference sequence [66] with respect to the sequence of subsets selected by **DepRound** and the realisations of the Bernoulli random variables for each arm in the selection, up to round $t-1$. We also have,

$$\begin{aligned}
\mathbb{E}\big[Y_{t,i}^2 \,|\, I_1, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\big] &= \mathbb{E}\left[\frac{B_{t,i}}{\epsilon^2}(x_{t,i}+\upsilon)^2 + (x_{t,i}+\upsilon)^2 - 2\frac{B_{t,i}}{\epsilon}(x_{t,i}+\upsilon)^2\right] \\
&\leq \mathbb{E}\left[\frac{B_{t,i}}{\epsilon^2}(x_{t,i}+\upsilon)^2 \,\middle|\, I_1, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\right] \\
&\leq \frac{(1+\upsilon)^2}{\epsilon} \\
&\leq \frac{9}{4\epsilon}, \qquad (5.28)
\end{aligned}$$

where (5.28) holds since $\upsilon \leq \gamma \leq 1/2$ for arms $i \in I_t$. We then have,

$$\mathbb{E}\left[\sum_{i\in I_t} Y_{t,i}^2 \,\middle|\, I_1, B_{1,i}, \ldots, I_{t-1}, B_{t-1,i}\right] \leq \frac{9k}{4\epsilon}, \tag{5.29}$$

where $\sum_{i\in I_t} Y_{t,i}$ is also a martingale difference sequence.

Since $0 < \epsilon < 1$ and $k \geq 1$, then we have,

$$\sum_{i \in I_t} Y_{t,i} \in \left[-\frac{3k}{2}, \left(\frac{1}{\epsilon} - 1\right)\frac{3k}{2} \right],$$

for $t = 1, \ldots, T$. From Lemma 4 we have,

$$\mathbb{P}\left[\sum_{t=1}^{T} \sum_{i \in I_t} Y_{t,i} > \sqrt{\frac{9kT}{2\epsilon} \log\left(\frac{2K}{\delta}\right)} + \frac{k}{\epsilon} \log\left(\frac{2K}{\delta}\right) \right] \leq \frac{\delta}{2K},$$

giving,

$$\mathbb{P}\left[\sum_{t=1}^{T} \sum_{i \in I_t} B_{t,i} \frac{x_{t,i} + \upsilon}{\epsilon} \leq \sum_{t=1}^{T} \sum_{i \in I_t} x_{t,i} + \upsilon kT + \sqrt{\frac{9kT}{2\epsilon} \log\left(\frac{2K}{\delta}\right)} + \frac{k}{\epsilon} \log\left(\frac{2K}{\delta}\right) \right] \geq 1 - \frac{\delta}{2K}. \tag{5.30}$$

From Lemma 5 we have,

$$\mathbb{P}\left[\sum_{t=1}^{T} \sum_{i \in I_t^*} x_{t,i} \leq \sum_{t=1}^{T} \sum_{i \in I_t^*} x'_{t,i} + 4\upsilon kKT \right] \geq 1 - \frac{\delta}{2K}. \tag{5.31}$$

By using the fact that for events A and B,

$$\mathbb{P}\big[A \cap B\big] \geq \max\big\{0, \mathbb{P}\big[A\big] + \mathbb{P}\big[B\big] - 1\big\},$$

and combining (5.26) with (5.30) and (5.31) we get,

$$\begin{aligned} \sum_{t=1}^{T} \sum_{i \in I_t} x_{t,i} \geq & \left(1 - \left(\gamma + \frac{2(e-2)\gamma}{\epsilon}\right)\right)\left(\sum_{t=1}^{T} \sum_{i \in I_t^*} x_{t,i} - 4\upsilon kKT\right) \\ & - \frac{K(1+\beta T)}{\gamma(1-\theta)} \log\left(\frac{K}{k}\right) - \upsilon kT - \sqrt{\frac{9kT}{2\epsilon} \log\left(\frac{2K}{\delta}\right)} - \frac{k}{\epsilon} \log\left(\frac{2K}{\delta}\right), \end{aligned}$$

with probability at least $1 - \delta/K$. Taking expectations and recalling that $G_{max-k,\theta} = \mathbb{E}\big[\sum_{t=1}^{T} \sum_{i \in I_t^*} x_{t,i}\big]$ and since $\log(2K/\delta) \geq 1$, this completes the proof. □

Corollary 6.1 follows from Theorem 6, and considering that $G_{max-k,\theta} \leq kT$.

Corollary 6.1. *Set* $\beta = 1/T$,

$$\gamma = \min\left\{1, \sqrt{\frac{2\epsilon K}{(2e-3)(1-\theta)kT}\log\left(\frac{K}{k}\right)}\right\},$$

and,

$$\upsilon = \sqrt{\frac{1}{\epsilon kKT}\log\left(\frac{K}{k}\right)},$$

then an upper bound on the regret of ***LET-Exp3.M*** *is,*

$$\overline{\mathcal{R}}_T \leq 2\sqrt{\frac{2(2e-3)kKT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{2K}{\delta}\right)} + 4(1+k)\sqrt{\frac{kT}{2\epsilon}\log\left(\frac{2K}{\delta}\right)} + \frac{k}{\epsilon}\log\left(\frac{2K}{\delta}\right),$$

and holds for any assignment of rewards and any $T > 0$*, with probability at least* $1 - \delta/K$.

In this section, we analysed the theoretical performance of the **LET-Exp3.M** algorithm. The deviation between cumulative rewards and cumulative biased reward estimates is bound with high probability and used to upper bound the regret of the algorithm.

5.3 Summary

In this chapter, we extended the literature on combining label efficient prediction with multi-armed bandits to the thresholding semi-bandit game against an oblivious adversary. The assumption that exactly $k \leq \exp(-1)K$ arms are known to exceed threshold each round, but not which k, still holds from Chapter 4. The semi-bandit aspect of the problem requires that the player has access to information from no more than k arms each round.

The label efficient feature requires that an average of ϵkT arms are actually observed over a sequence of T rounds and for $0 < \epsilon < 1$. In the event that the player receives feedback from an arm that exceeds a pre-defined threshold, she updates the exponential weighting for selecting that arm again in future. If the player observes feedback from an arm that is within threshold, she penalises the exponential weighting for that arm

as a function of some tunable parameter $0 < \beta < \exp(-\theta)$. The game proceeds over a sequence of T rounds and the objective of the player is to observe as many arms that are exceeding threshold as possible.

We introduced the **LET-Exp3.M** algorithm, which selects subsets of k arms using a distribution over arms and the **DepRound** procedure, as in Chapter 4. A sequence of Bernoulli random variables for each arm determines whether to query said arm (or allow an arm to be queried) and a biased reward estimate is computed depending on which set each arm corresponds to: selected and queried, selected but not queried and not selected. It is noted that the semi-bandit setting remains satisfied since arms selected but not queried are estimated without knowledge of their feedback. An importance weighted distribution is used, where the weights for each arm are exponentially updated. For arms observed with feedback exceeding threshold or not queried at all in round t, weights are exponentially updated depending on their biased reward estimate. Arms selected and queried but observed within threshold have their weight penalised by $0 < \beta < \exp(-\theta)$, which ensures that the arms we do not want to observe are less likely to be selected in future. Arms selected but not queried return a fixed and positive biased reward estimate, increasing their weight when updated. Combined with decreasing weights for undesirable arms, this allows for arms not selected over several round to have more opportunity to be selected.

Using several concentration bounds, we prove an upper bound on the regret of our **LET-Exp3.M** algorithm of $\mathcal{O}\Big(\sqrt{\frac{kKT}{\epsilon(1-\theta)}\log\big(\frac{K}{k}\big)\log\big(\frac{K}{\delta}\big)}\Big)$ with high probability. This further generalises that of **T-Exp3.M**, **Exp3.M** and **Exp3**, up to log factors in K. The computational complexity of **LET-Exp3.M** matches the $\mathcal{O}\Big(KT(\log(k)+1)\Big)$ of **T-Exp3.M** and **Exp3.M** but the space requirements are only $\mathcal{O}(\epsilon kT)$.

As in Chapter 4, we are assuming knowledge of the number of arms exceeding threshold each round, but do not know which of the K arms and set $q_t = k_t = k$. We note it is not always reasonable that Assumption 2 holds both in terms of exactly k arms exceeding threshold each round and the player having knowledge of such information. With static Bernoulli parameters, the player may decide to not query an arm even if the associated importance weighting indicates that the arm is highly likely to exceed threshold. We will turn our attention to relaxing these constraints in Chapter 6.

Chapter 6

Dynamic label efficient adversarial thresholding semi-bandits

Now we are in a position to introduce the problem where the player decides to query an arm based the history of observations for that arm, and may select any number of arms each round. The importance weighted distribution over arms remains the same and a k-armed subset is selected, based on the distribution and using **DepRound**, [33]. However, the player also maintains a distinct set of arms that will be queried with probability one, if some condition is met.

Up to now we have used Assumption 2 where exactly $k \leq \exp(-1)K$ arms exceed threshold each round. This is obviously not sustainable as we may have no idea how many arms will exceed threshold each round. However, without making any assumptions on the structure of feedback from arms, the problem is even more challenging. We can only base any assumption on whether an arm will exceed threshold in the next round in terms of its previous behaviour. As such, we introduce the idea of having two sets of arms to pull. The first subset is the one used previously with k arms and chosen randomly by **DepRound**.

The second subset is dynamically maintained each round and only contains those arms that were observed exceeding threshold in the previous round. Arms in the first subset are queried using their associated Bernoulli variable (as in Chapter 5) and arms in the second subset are observed with certainty.

As briefly mentioned in the previous chapter, when playing the label efficient adversarial thresholding semi-bandit game, a player selecting an optimal arm $i^* \in \{I_t \mid x_{t,i} > \theta\}$ in round t will only get to observe $x_{t,i}$ with probability ϵ_{i^*}. Our aim is to ensure that if the player selects an optimal arm for querying such that $i^* \in I_t$, then the likelihood of querying is high. This corresponds with minimising the difference between the total number of times every arm exceeds threshold over T rounds and the total number of queries the player is allowed to make.

6.1 Dynamic label efficient prediction

In this section, we introduce the reader to the concept of *dynamic label efficient prediction* and discuss the differences and similarities with the setup given in both [39] and [35].

A player of the original label efficient prediction game proceeds each round to choose an expert from a set of experts, where an expert refers to a specific arm. The player receives some loss for the chosen expert and the player decides whether to query all experts or not, with probability ϵ. On average there are $m \approx \epsilon T$ rounds where the player queries every expert. In this setup, ϵ is constant and the decision to query is uniformly distributed over the time horizon. This means that the decision to query or not could actually be made prior to the first round, analogous to an oblivious adversary choosing the sequences of feedback for each arm. This is also the case for when combining label efficient prediction with adversarial thresholding semi-bandits in Chapter 5. However, the premise of online learning games is for a player to use the information gathered as rounds progress to inform their decision-making in future.

In dynamic label efficient prediction, the player still queries labels efficiently, but does so intelligently by utilising the history of observations from each arm. More formally, let K and T be positive integers as before, where $[K] = \{1, \ldots, K\}$.

For a given arm $i \in [K]$, the arm is queried if the realisation of a Bernoulli random variable $B_{t,i} = 1$ and where $\mathbb{P}\big[B_{t,i} = 1\big] = \epsilon_{t,i}$. Since the player has no prior information to base their decisions on in round $t = 1$, we have $\epsilon_{1,i} = 1$ for every arm $i \in [K]$ and,

$$\epsilon_{t,i} = \zeta^{N_{t,i}}, \tag{6.1}$$

where $0 < \zeta < 1$, represents the decay factor for incorrectly pulling an arm when their feedback is within threshold and $N_{t,i} = \sum_{s=2}^{t-1} \mathbb{I}\{i \in I_s \,|\, B_{s,i} = 1, x_{s,i} \leq \theta\}$ denotes the frequency with which arm i is selected, queried and observed within threshold up to round $t-1$. We initially set $N_{1,i} = 0$ for $i = 1, \ldots, K$.

It is noted that the sequence of Bernoulli random variables for each arm is no longer independent and we cannot utilise any theoretical framework that relies on such a fact. We also note that if an arm is seen within threshold on several successive queries, then the probability of querying will become small and may hamper the chances of querying that arm again if it does indeed begin to exceed threshold in the future. Our empirical approach does take this into account.

6.2 Adaptive k and threshold lists

In this section we discuss our approach for relaxing Assumption 2, which assumes that exactly $k < K$ arm will exceed threshold each round but the player does not know which of those k. In order to do this, we define the following,

Assumption 3. *If arm i is observed exceeding threshold in round t, then arm i will definitely be observed in round $t+1$.*

As a result of relaxing Assumption 2, we no longer assume that $q_t = k_t = k$ and have $\Delta_t \geq 1$, recalling that, $\Delta_t = |k_t - q_t| + 1$. Since we have no prior knowledge of $q_1, \ldots, q_T$, we estimate a lower bound for Δ_t as $\hat{\Delta}_t = K + 1$, where $|k_t - q_t| < K$. For arm i to be observed in round t, then we must have $i \in I_t$ and $B_{t,i} = 1$. Those arms $\{i \in I_t \,|\, B_{t,i} = 1, x_{t,i} > \theta\}$ observed exceeding threshold in round t are placed on a *Threshold list*.

Definition 4. *A threshold list is defined as the set of arms observed exceeding threshold in round t,*

$$\mathcal{T}_t = \{i \in I_{t-1} \,|\, B_{t-1,i} = 1, x_{t-1,i} > \theta\} \cup \{i \in \mathcal{T}_{t-1} \,|\, x_{t-1,i} > \theta\},$$

where $\mathcal{T}_1 = \emptyset$.

Any arm on the threshold list observed within threshold is removed from the threshold list in the next round. The subset of k arms are selected using **DepRound** as in Chapters 4 and 5. All arms $i \in \mathcal{T}_t$ are queried with probability one. Those arms $i \in I_t \backslash \mathcal{T}_t$ are observed only if the player decides to query them.

6.3 The dLET-Exp3.M algorithm

The *dynamic label efficient thresholding exponentially weighted exploration and exploitation with multiple plays* algorithm (**dLET-Exp3.M**) proceeds similarly to **LET-Exp3.M**, with the extra storage requirement of maintaining the threshold list (Definition 4). However, the exponentially weighted distribution over arms and the weight updates remain unchanged. The pseudo-code for **dLET-Exp3.M** is given in Algorithm 11 and in Figure 6.1, we sketch the learning cycle of the algorithm. As in Figure 5.1, Figure 6.1 shows the decision to observe feedback being made by the player. The player is also illustrated dynamically updating the Bernoulli parameters, depending on the threshold and feedback observed.

The weight update rules for the **dLET-Exp3.M** algorithm concentrate on four different sets of arms, $\mathcal{A}_t, \mathcal{B}_t, \mathcal{C}_t$ and the remainder $[K] \cap \mathcal{A}_t^c \cap \mathcal{B}_t^c \cap \mathcal{C}_t^c$ where,

$$\mathcal{A}_t = (I_t \backslash \mathcal{S}_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 1\} \cup \mathcal{T}_t) \cap \{i \in [K] \,|\, x_{t,i} > \theta\}, \tag{6.2}$$

which represents arms selected in round t with not too big weights ($w_{t,i} < \alpha_t$, where α_t is determined from (4.4)), queried, not on the threshold list and observed exceeding threshold or those arms on the threshold list in round t and also seen again outside of

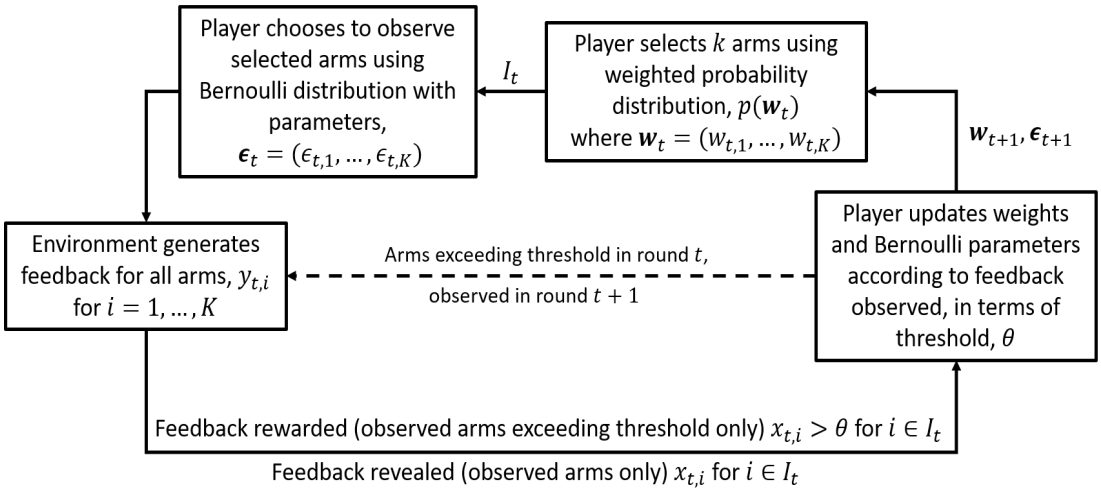


Figure 6.1: A schematic diagram of the dynamic exponentially weighted exploration and exploitation with multiple plays algorithm.

the threshold. We have,

$$\mathcal{B}_t = I_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 0\}, \tag{6.3}$$

which corresponds to all arms selected, not on the threshold list in round t and not queried. Set $\mathcal{C}_t$ is given by,

$$\mathcal{C}_t = (I_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 1\} \cup \mathcal{T}_t) \cap \{i \in [K] \,|\, x_{t,i} \leq \theta\}, \tag{6.4}$$

which is the set of arms selected by **DepRound**, queried, not on the threshold list and observed within threshold or the set of arms on the threshold list and also within threshold. The biased reward estimators for each arm $i \in [K]$ used in **dLET-Exp3.M** are determined, depending on whether they belong to sets $\mathcal{A}_t, \mathcal{B}_t$ or neither.

6.4 Analysis of dLET-Exp3.M

Analysis of the dynamic label efficient thresholding exponentially weighted exploration and exploitation approach concentrates on bounding the rate at which the sum of weights varies. For the thresholding semi-bandit setting, this required taking into account each of the weight update rules. As we move through the label efficient and now the dynamic label efficient version, the analysis needs to pay even more careful attention to the characteristics of each weight update rule.

Algorithm 11 dLET-Exp3.M algorithm

Input: Number of arms K, subset size $k \leq \exp(-1)K$, $w_{1,i} = 1$, $\epsilon_{1,i} = \epsilon_i$ for $i = 1, \ldots, K$, threshold $\theta \in (0,1)$, $\gamma \in (0,1]$, $0 < \beta < \exp(-\theta)$, $\zeta \in (0,1)$, $\mathcal{T}_1 = \emptyset$. Let,

$$\mathcal{A}_t = (I_t \backslash \mathcal{S}_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 1\} \cup \mathcal{T}_t) \cap \{i \in [K] \,|\, x_{t,i} > \theta\}$$
$$\mathcal{B}_t = I_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 0\}$$
$$\mathcal{C}_t = (I_t \cap \mathcal{T}_t^c \cap \{i \in [K] \,|\, B_{t,i} = 1\} \cup \mathcal{T}_t) \cap \{i \in [K] \,|\, x_{t,i} \leq \theta\}$$

for $t = 1, \ldots, T$ **do**
  if $\operatorname{argmax}_{j \in K} w_{t,j} \geq \frac{K - \gamma k}{kK(1-\gamma)} \sum_{i=1}^{K} w_{t,i}$ **then**
    Determine α_t such that,

$$\frac{\alpha_t}{\sum_{w_{t,i} \geq \alpha_t} \alpha_t + \sum_{w_{t,i} < \alpha_t} w_{t,i}} = \frac{K - \gamma k}{kK(1-\gamma)}.$$

    Let $S_t = \{i \in [K] \,|\, w_{t,i} \geq \alpha_t\}$.
    Let $w'_{t,i} = \alpha_t$ for $i \in S_t$.
  else
    Let $S_t = \emptyset$.
  Let $w'_{t,i} = w_{t,i}$ for $i \in [K] \backslash S_t$.
  Let

$$p_{t,i} = k\left((1-\gamma)\frac{w'_{t,i}}{\sum_{j=1}^{K} w'_{t,j}} + \frac{\gamma}{K}\right).$$

  Let $I_t = \textbf{DepRound}(k, (p_{t,1}, \ldots, p_{t,K}))$.
  Adversary generates feedback $x_{t,i}$ for $i \in [K]$.
  Observe $x_{t,i}$ for $i \in \mathcal{T}_t$.
  for $i \in [K] \backslash \mathcal{T}_t$, **do**
    Draw Bernoulli random variables $B_{t,i}$, such that $\mathbb{P}\big[B_{t,i} = 1\big] = \epsilon_{t,i}$.
    Observe $B_{t,i} x_{t,i}$.
  for $i = 1, \ldots, K$ **do**
    Let,

$$x'_{t,i} = \begin{cases} \frac{x_{t,i} + \upsilon}{p_{t,i}\epsilon_i} & \text{if } i \in \mathcal{A}_t \\ \frac{\upsilon}{p_{t,i}\epsilon_i} & \text{if } i \in \mathcal{B}_t \\ 0 & \text{otherwise.} \end{cases}$$

    Let,

$$w_{t+1,i} = \begin{cases} w_{t,i} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) & \text{if } i \in \mathcal{A}_t \cup \mathcal{B}_t \\ w_{t,i} \exp(\theta \log(\beta)) & \text{if } i \in \mathcal{C}_t \\ w_{t,i} & \text{otherwise.} \end{cases}$$

    if $i \in \mathcal{C}_t$ **then**
      $\epsilon_{t+1,i} = \max\{\frac{1}{k}, \zeta \epsilon_{t,i}\}$.
  Let $\mathcal{T}_{t+1} = \{i \in I_t \,|\, B_{t,i} = 1, x_{t,i} > \theta\} \cup \{i \in \mathcal{T}_t \,|\, x_{t,i} > \theta\}$

The set of Bernoulli random variables the player uses to decide whether to query an arm or not remains independent with other arms, but is now conditional on the observed history for each arm. In Lemma 7 we modify Lemma 5, where we take expectations only in terms of the randomness in selecting arms to be observed and consider the expectation for querying arms in the worst-case. Theorem 8 uses Lemma 7 to provide a theoretical guarantee on the performance of **dLET-Exp3.M** and Corollary 8.1 follows by the appropriate tuning of parameters.

Lemma 7. *For real number* $0 < \delta < 1$, *let* $J_t = I_t \cup \mathcal{T}_t$. *Then we have,*

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in J_t} x_{t,i} > \sum_{t=1}^{T}\sum_{i\in J_t} x'_{t,i} + 4\upsilon kKT\right] \leq \frac{\delta}{2K}.$$

The proof of Lemma 7 follows the same Chernoff bounding used on Lemma 5. This allows a bound on the difference between the cumulative feedback from observed arms and cumulative biased feedback estimates to hold with a given probability.

Proof of Lemma 7. We begin by stating that,

$$\upsilon \geq \sqrt{\frac{1}{\epsilon kKT}\log\left(\frac{2K}{\delta}\right)}.$$

By letting $a = 4\upsilon k^2K^2T$, $b = \epsilon\upsilon/4kK$ and considering pulled arms regardless of being queried then from (5.5) we have,

$$Z_t = \exp\left(b\left(\sum_{i\in J_t}(x_{t,i} - x'_{t,i})\right)\right)Z_{t-1}.$$

From Algorithm 11 we have that,

$$\begin{aligned} x'_{t,i} &= \tilde{x}_{t,i} + B_{t,i}\frac{\upsilon}{p_{t,i}\epsilon_{t,i}} \\ &\leq \tilde{x}_{t,i} + \frac{\upsilon k}{p_{t,i}}, \end{aligned}$$

where $B_{t,i} = \{0, 1\}$, $\epsilon_{t,i} \geq 1/k$ and,

$$\tilde{x}_{t,i} = \begin{cases} \frac{x_{t,i}}{p_{t,i}\epsilon_{t,i}} & \text{if } i \in \mathcal{A}_t \\ 0 & \text{otherwise.} \end{cases}$$

This gives,

$$\begin{aligned} b\left(\sum_{i\in J_t}(x_{t,i} - x'_{t,i})\right) &\leq b\left(\sum_{i\in J_t}\left(x_{t,i} - \tilde{x}_{t,i} - \frac{\upsilon k}{p_{t,i}}\right)\right) \\ &\leq \frac{\upsilon}{4k} - \frac{\upsilon}{4kK}\sum_{i\in J_t}\tilde{x}_{t,i} - \frac{\upsilon^2}{4K}\sum_{i\in J_t}\frac{1}{p_{t,i}} \qquad (6.5) \\ &\leq \frac{1}{4k} \qquad (6.6) \\ &< 1, \end{aligned}$$

where (6.5) holds since $\epsilon \leq 1$, $x_{t,i} \leq 1$, $|J_t| \leq K$ and by substituting b and (6.6) holds since $\upsilon \leq 1$ and the last two terms in (6.5) are always positive. From (6.6) we get,

$$\begin{aligned} &\mathbb{E}\big[Z_t \,|\, J_1, \ldots, J_{t-1}\big] = \mathbb{E}\left[\exp\left(b\left(\sum_{i\in J_t}\left(x_{t,i} - \tilde{x}_{t,i} - \frac{\upsilon k}{p_{t,i}}\right)\right)\right) \,\middle|\, J_1, \ldots, J_{t-1}\right] Z_{t-1} \\ &\leq \mathbb{E}\left[1 + b\left(\sum_{i\in J_t}\left(x_{t,i} - \tilde{x}_{t,i} - \frac{\upsilon k}{p_{t,i}}\right) + b^2\left(\sum_{i\in J_t}\left(x_{t,i} - \tilde{x}_{t,i} - \frac{\upsilon k}{p_{t,i}}\right)\right)^2\right) \,\middle|\, J_1, \ldots, J_{t-1}\right] Z_{t-1}, \end{aligned} \qquad (6.7)$$

where (6.7) holds since $\exp(a) \leq 1 + a + a^2$ for $a \leq 1$. Since $\sum_{i\in J_t} x_{t,i} \leq K$, and $\epsilon_{t,i} \geq 1/k$ for $t = 1, \ldots, T$ and $i = 1, \ldots, K$ then,

$$\begin{aligned} \mathbb{E}\left[\sum_{i\in J_t}(x_{t,i} - \tilde{x}_{t,i}) \,\middle|\, J_1, \ldots, J_{t-1}\right] &= \sum_{i\in J_t} x_{t,i}\left(1 - \frac{1}{\epsilon_{t,i}}\right) \\ &\leq (1-k)K, \qquad (6.8) \end{aligned}$$

where (6.8) holds since $\sum_{i\in J_t}\tilde{x}_{t,i} = \sum_{i\in\mathcal{A}_t}\tilde{x}_{t,i}$ because $\tilde{x}_{t,i} = 0$ for $i \notin \mathcal{A}_t$ and $\mathcal{A}_t \subseteq J_t$.

We also have,

$$\mathbb{E}\left[\sum_{i\in J_t}\frac{\upsilon k}{p_{t,i}}\,\middle|\, J_1,\ldots,J_{t-1}\right] \leq \upsilon k K, \tag{6.9}$$

since $\mathbb{E}\Big[\sum_{i\in J_t} 1/p_{t,i}\,\Big|\, J_1,\ldots,J_{t-1}\Big] \leq K$. Hence from (6.7), (6.8) and (6.9) we have,

$$\begin{aligned}\mathbb{E}\left[b\left(\sum_{i\in J_t}\left(x_{t,i}-\tilde{x}_{t,i}-\frac{\upsilon k}{p_{t,i}}\right)\right)\,\middle|\, J_1,\ldots,J_{t-1}\right] &\leq \frac{\upsilon}{4kK}\big[(1-k)K-\upsilon kK\big]\\ &\leq \frac{1-(1+\upsilon)k}{4k},\end{aligned} \tag{6.10}$$

where (6.10) is non-positive since, $(1+\upsilon)k \geq 1$, for $\upsilon \geq 0$ and $k \geq 1$. Then,

$$\mathbb{E}\big[Z_t \,\big|\, J_1,\ldots,J_{t-1}\big] \leq Z_{t-1}, \tag{6.11}$$

Taking expectations on both sides of (6.11) and from the fact that $\mathbb{E}\big[Z_1 \,\big|\, J_1,\ldots,J_{t-1}\big] \leq 1$ gives $\mathbb{E}\big[Z_t \,\big|\, J_1,\ldots,J_{t-1}\big] \leq 1$ for $t = 1,\ldots,T$. This completes the proof. □

Theorem 8 proves a theoretical bound on the performance of the **dLET-Exp3.M** algorithm with high probability, using the result proved in Lemma 7.

Theorem 8. *For integers $T > 0$, $K \geq k \geq 1$ and $t \in [T]$, real numbers $2K\exp(-KT) < \delta < 1$, $\frac{1}{k} \leq \epsilon_i \leq 1$ for $i \in [K]$, $\epsilon = \min_{i\in[K]}\{\epsilon_i\}$, $0 < \beta < \exp(-\theta)$, $0 < \theta < 1$, $0 < \gamma < 1$ and,*

$$\sqrt{\frac{1}{\epsilon k K T}\log\left(\frac{2K}{\delta}\right)} \leq \upsilon \leq 1,$$

then,

$$\begin{aligned}\overline{\mathcal{R}}_T \leq{}& (e-1)(1+\upsilon)\gamma k G_{max-k,\theta} + \frac{K(1+T(\beta+\upsilon(3k+K)))}{\gamma(1-\theta)}\log\left(\frac{K}{k}\right)\\ &+ 4(e-1)(1+\upsilon)\upsilon\gamma k^2 K T,\end{aligned}$$

holds for any assignment of rewards with probability at least $1-\delta/2K$.

By careful consideration of the weight update rules given in Algorithm 11, we are now in a position to prove the theoretical performance guarantees of the **dLET-Exp3.M** algorithm. The proof follows similar lines to the theorems given in Chapters 4 and 5, however, due to the lack of independence between successive queries of the same arm, we consider the worst-case in terms of queries and take expected regret over the randomness in selecting arms each round, whether queried or not.

Proof of Theorem 8. Let $W_t = \sum_{i\in[K]} w_{t,i}$ then we have,

$$W_t = \sum_{i\in\mathcal{A}_t} w_{t,i} + \sum_{i\in\mathcal{B}_t} w_{t,i} + \sum_{i\in\mathcal{C}_t} w_{t,i} + \sum_{i\in\mathcal{D}_t} w_{t,i}, \tag{6.12}$$

where $\mathcal{D}_t = [K] \cap \mathcal{A}_t^c \cap \mathcal{B}_t^c \cap \mathcal{C}_t^c$. The ratio of the sum of weights for successive rounds is given by,

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \frac{1}{W_t}\left[\sum_{i\in\mathcal{A}_t} w_{t+1,i} + \sum_{i\in\mathcal{B}_t} w_{t+1,i} + \sum_{i\in\mathcal{C}_t} w_{t+1,i} + \sum_{i\in\mathcal{D}_t} w_{t+1,i}\right] \\
&\leq 1 + \sum_{i\in\mathcal{A}_t} \frac{w_{t,i}}{W_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) + \sum_{i\in\mathcal{B}_t} \frac{w_{t,i}}{W_t} \exp\left(\frac{\gamma k}{K} x'_{t,i}\right) + \sum_{i\in\mathcal{C}_t} \beta \exp(\theta) \frac{w_{t,i}}{W_t} \qquad (6.13) \\
&\leq 1 + \beta \exp(\theta) + \frac{W'_t}{W_t} \sum_{i\in\mathcal{A}_t} \frac{w'_{t,i}}{W'_t}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \\
&\quad + \frac{W'_t}{W_t} \sum_{i\in\mathcal{B}_t} \frac{w'_{t,i}}{W'_t}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \qquad (6.14) \\
&\leq 1 + \beta \exp(\theta) + \sum_{i\in\mathcal{A}_t} \frac{w'_{t,i}}{W'_t}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \\
&\quad + \sum_{i\in\mathcal{B}_t} \frac{w'_{t,i}}{W'_t}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \qquad (6.15) \\
&= 1 + \beta \exp(\theta) + \sum_{i\in\mathcal{A}_t} \frac{\frac{p_{t,i}}{k} - \frac{\gamma}{K}}{1-\gamma}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \\
&\quad + \sum_{i\in\mathcal{B}_t} \frac{\frac{p_{t,i}}{k} - \frac{\gamma}{K}}{1-\gamma}\left[1 + \frac{\gamma k}{K} x'_{t,i} + (e-2)\left(\frac{\gamma k}{K} x'_{t,i}\right)^2\right] \\
&\leq 1 + \beta \exp(\theta) + \frac{\gamma}{K(1-\gamma)}\left[\sum_{i\in\mathcal{A}_t} x'_{t,i} p_{t,i} + \sum_{i\in\mathcal{B}_t} x'_{t,i} p_{t,i}\right] \\
&\quad + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)}\left[\sum_{i\in\mathcal{A}_t} x'^2_{t,i} p_{t,i} + \sum_{i\in\mathcal{B}_t} x'^2_{t,i} p_{t,i}\right]. \qquad (6.16)
\end{aligned}$$

where (6.13) holds since $\sum_{i\in\mathcal{C}_t} w_{t,i}/W_t \leq 1$ and $\sum_{i\in\mathcal{D}_t} w_{t,i}/W_t \leq 1$, (6.14) holds since $\exp(a) \leq 1 + a + (e-2)a^2$ for $a \leq 1$ and (6.15) holds since $W'_t/W_t \leq 1$. Since $x'_{t,i} = \upsilon/p_{t,i}\epsilon_{t,i}$ for $i \in \mathcal{B}_t$ then from (6.16) we have,

$$\frac{W_{t+1}}{W_t} \leq 1 + \beta\exp(\theta) + \frac{\gamma}{K(1-\gamma)}\left[\sum_{i\in\mathcal{A}_t} x'_{t,i}p_{t,i} + \sum_{i\in\mathcal{B}_t}\frac{\upsilon}{\epsilon_{t,i}}\right] + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)}\left[\sum_{i\in\mathcal{A}_t} x'^2_{t,i}p_{t,i} + \sum_{i\in\mathcal{B}_t}\frac{x'_{t,i}\upsilon}{\epsilon_{t,i}}\right]. \tag{6.17}$$

Taking logarithms on both sides of (6.17) gives and summing over $t = 1, \ldots, T$,

$$\log\left(\frac{W_{T+1}}{W_1}\right) \leq \frac{\beta T}{1-\theta}\log\left(\frac{K}{k}\right) + \frac{\gamma}{K(1-\gamma)}\sum_{t=1}^{T}\left[\sum_{i\in\mathcal{A}_t} x'_{t,i}p_{t,i} + \sum_{i\in\mathcal{B}_t}\frac{\upsilon}{\epsilon_{t,i}}\right] + \frac{(e-2)\gamma^2 k}{K^2(1-\gamma)}\sum_{t=1}^{T}\left[\sum_{i\in\mathcal{A}_t} x'^2_{t,i}p_{t,i} + \sum_{i\in\mathcal{B}_t}\frac{x'_{t,i}\upsilon}{\epsilon_{t,i}}\right], \tag{6.18}$$

where (6.18) holds since $\log(a+b) = \log(a) + \log(1+b/a)$, $1/(1+\beta\exp(\theta)) \leq 1$, $\exp(a) \leq 1/(1-a)$ for $a \leq 1$, $\log(1+a) \leq a$ and $\log(K/k) \geq 1$ for $\exp(-1)K \geq k$. We now consider that,

$$\begin{aligned}\sum_{i\in\mathcal{A}_t} x'_{t,i}p_{t,i} &= \sum_{i\in\mathcal{A}_t}\frac{x_{t,i}+\upsilon}{\epsilon_{t,i}} \\ &\leq \sum_{i\in I_t}\frac{x_{t,i}+\upsilon}{\epsilon_{t,i}} + \sum_{i\in\mathcal{T}_t}\frac{x_{t,i}+\upsilon}{\epsilon_{t,i}},\end{aligned} \tag{6.19}$$

where $\mathcal{A}_t \subseteq I_t \cup \mathcal{T}_t$. By setting $\epsilon_{min} = \min_{t\in[T],i\in[K]}\{\epsilon_{t,i}\}$. Then,

$$\sum_{t=1}^{T}\sum_{i\in\mathcal{A}_t} x'_{t,i}p_{t,i} \leq \frac{1}{\epsilon_{min}}\sum_{t=1}^{T}\left[\sum_{i\in I_t} x_{t,i} + \sum_{i\in\mathcal{T}_t} x_{t,i}\right] + \frac{\upsilon(k+K)T}{\epsilon_{min}}\log\left(\frac{K}{k}\right), \tag{6.20}$$

where $|I_t| = k$ and $|\mathcal{T}_t| \leq K$. We also consider that,

$$\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\frac{\upsilon}{\epsilon_{t,i}} \leq \frac{\upsilon kT}{\epsilon_{min}}\log\left(\frac{K}{k}\right), \tag{6.21}$$

where $|\mathcal{B}_t| \leq k$. By the same principle we have,

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i\in\mathcal{A}_t} x_{t,i}'^2 p_{t,i} &= \sum_{t=1}^{T}\sum_{i\in\mathcal{A}_t} x'_{t,i}\frac{x_{t,i}+\upsilon}{\epsilon_{min}} \\ &= \frac{1}{\epsilon_{min}}\sum_{t=1}^{T}\left[\sum_{i\in I_t} x'_{t,i} + \sum_{i\in\mathcal{T}_t} x'_{t,i}\right] + \frac{\upsilon(k+K)T}{\epsilon_{min}}\log\left(\frac{K}{k}\right), \qquad (6.22)\end{aligned}$$

where (6.22) holds since $x_{t,i} \leq 1$. For the final term in (6.18) we have,

$$\sum_{t=1}^{T}\sum_{i\in\mathcal{B}_t}\frac{x'_{t,i}\upsilon}{\epsilon_{t,i}} \leq \frac{\upsilon}{\epsilon_{min}}\sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i}. \qquad (6.23)$$

Substituting (6.20), (6.21), (6.22) and (6.23) into (6.18) gives,

$$\begin{aligned}\log\left(\frac{W_{T+1}}{W_1}\right) \leq \frac{(\upsilon(3k+K)+\beta)T}{\epsilon_{min}(1-\theta)}\log\left(\frac{K}{k}\right) + \frac{\gamma}{\epsilon_{min}K(1-\gamma)}\sum_{t=1}^{T}\left[\sum_{i\in I_t} x_{t,i} + \sum_{i\in\mathcal{T}_t} x_{t,i}\right] \\ + \frac{(e-2)\gamma^2 k}{\epsilon_{min}K^2(1-\gamma)}\left[(1+\upsilon)\sum_{t=1}^{T}\sum_{i\in I_t} x'_{t,i} + \sum_{t=1}^{T}\sum_{i\in\mathcal{T}_t} x'_{t,i}\right]. \qquad (6.24)\end{aligned}$$

From the weight update rules in Algorithm 11 and recalling (6.12) we have,

$$\begin{aligned}W_{T+1} &= \sum_{i=1}^{K} w_{T+1,i} \\ &= \sum_{i\in\mathcal{A}_{T+1}} w_{T+1,i} + \sum_{i\in\mathcal{B}_{T+1}} w_{T+1,i} + \sum_{i\in\mathcal{C}_{T+1}} w_{T+1,i} + \sum_{i\in\mathcal{D}_{T+1}} w_{T+1,i} \\ &= \sum_{i\in\mathcal{A}_{T+1}} \exp\left(\sum_{t=1}^{T}\frac{\gamma k}{K}x'_{t,i}\right) + \sum_{i\in\mathcal{B}_{T+1}} \exp\left(\sum_{t=1}^{T}\frac{\gamma k}{K}x'_{t,i}\right) \\ &\quad + \sum_{i\in\mathcal{C}_{T+1}} \beta^T\exp(\theta T) + \sum_{i\in\mathcal{D}_{T+1}} w_{T+1,i} \\ &\geq \sum_{i\in I^*_{T+1}\setminus S_{T+1}} \exp\left(\sum_{t=1}^{T}\frac{\gamma k}{K}x'_{t,i}\right) + \sum_{i\in\mathcal{T}^*_{T+1}} \exp\left(\sum_{t=1}^{T}\frac{\gamma k}{K}x'_{t,i}\right) \qquad (6.25) \\ &= \sum_{i\in J^*} \exp\left(\sum_{t=1}^{T}\frac{\gamma k}{K}x'_{t,i}\right), \qquad (6.26)\end{aligned}$$

where (6.25) holds since $\mathcal{A}_t \cup \mathcal{B}_t = I^*_t \setminus S_t \cup \mathcal{T}^*_t$ and $\mathcal{C}_t = \emptyset$ when $I_t = I^*_t$. We denote $\mathcal{T}_t = \mathcal{T}^*_t$ when $I_t = I^*_t$ as the optimal threshold list and optimal queried subset, respectively.

We denote $J^* = I_t^* \backslash S_t \cup \mathcal{T}_t^*$ as the optimal set of arms in (6.26) for $t = 1, \ldots, T$. When the player plays optimally, the threshold list contains every arm that exceeded threshold in round $t-1$ and proceeds to continue exceeding threshold in round t. The remaining arms that are selected and queried account for those that were within threshold in round $t-1$ but exceed it in the following round. Recalling from [72] the fact that,

$$\sum_{i \in J^*} w_{T+1,i} \geq k \left(\prod_{i \in J^*} w_{T+1,i} \right)^{\frac{1}{k}},$$

then we have,

$$\begin{aligned} \log\left(\frac{W_{T+1}}{W_1}\right) &\geq \log\left(\prod_{i \in J^*} \exp\left(\sum_{t=1}^{T} \frac{\gamma k}{K} x'_{t,i} \right) \right) + \log\left(\frac{k}{K}\right) \\ &= \sum_{t=1}^{T} \sum_{i \in J^*} \frac{\gamma k}{K} x'_{t,i} + \log\left(\frac{k}{K}\right). \end{aligned} \tag{6.27}$$

Combining (6.24) and (6.27) gives,

$$\begin{aligned} \sum_{t=1}^{T} \sum_{i \in J^*} \frac{\gamma k}{K} x'_{t,i} + \log\left(\frac{k}{K}\right) &\leq \frac{(\upsilon(3k+K)+\beta)T}{\epsilon_{min}(1-\theta)} \log\left(\frac{K}{k}\right) + \frac{\gamma}{\epsilon_{min} K (1-\gamma)} \sum_{t=1}^{T} \left[\sum_{i \in I_t} x_{t,i} + \sum_{i \in \mathcal{T}_t} x_{t,i} \right] \\ &\quad + \frac{(e-2)\gamma^2 k}{\epsilon_{min} K^2 (1-\gamma)} \left[(1+\upsilon) \sum_{t=1}^{T} \sum_{i \in I_t} x'_{t,i} + \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} x'_{t,i} \right] \\ &\leq \frac{(\upsilon(3k+K)+\beta)T}{\epsilon_{min}(1-\theta)} \log\left(\frac{K}{k}\right) + \frac{\gamma}{\epsilon_{min} K (1-\gamma)} \sum_{t=1}^{T} \sum_{i \in J_t} x_{t,i} \\ &\quad + \frac{(1+\upsilon)(e-2)\gamma^2 k}{\epsilon_{min} K^2 (1-\gamma)} \sum_{t=1}^{T} \sum_{i \in J^*} x'_{t,i}, \end{aligned} \tag{6.28}$$

where $J_t = I_t \cup \mathcal{T}_t$ denotes the set of arms pulled by the **dLET-Exp3.M** player and,

$$(1+\upsilon) \sum_{t=1}^{T} \sum_{i \in I_t} x'_{t,i} + \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} x'_{t,i} \leq (1+\upsilon) \sum_{t=1}^{T} \sum_{i \in J^*} x'_{t,i}.$$

From (6.28) we have,

$$(1-(e-1)(1+\upsilon)\gamma k)\sum_{t=1}^{T}\sum_{i\in J^*}x'_{t,i} \leq G_{dLET-Exp3.M} + \frac{K+(\upsilon(3k+K)+\beta)KT}{\gamma(1-\theta)}\log\left(\frac{K}{k}\right), \quad (6.29)$$

where $\epsilon_{min} \geq 1/k$ and $(e-2)(1+\upsilon)+1 \leq (e-1)(1+\upsilon)$. By considering Lemma 7 only in terms of the randomness in selecting arms $i \in J_t^*$ we have,

$$\mathbb{P}\left[\sum_{t=1}^{T}\sum_{i\in J^*}x'_{t,i} \geq \sum_{t=1}^{T}\sum_{i\in J^*}x_{t,i} - 4\upsilon kKT\right] \geq 1-\frac{\delta}{2K},$$

where $0 < \delta < 1$ and $\sqrt{\frac{1}{\epsilon_{min}kKT}\log\left(\frac{2K}{\delta}\right)} \leq \upsilon \leq 1$. Then by taking expectations with respect to the randomness in selecting subset each round and where $G_{max-k,\theta} = \mathbb{E}\left[\sum_{t\in[T]}\sum_{i\in J^*}x_{t,i}\right]$, this complete the proof. □

By the careful tuning of parameters, Corollary 8.1 follows from Theorem 8.

Corollary 8.1. *Set* $\beta = 1/T$,

$$\upsilon = \sqrt{\frac{1}{\epsilon kKT}\log\left(\frac{2K}{\delta}\right)},$$

and,

$$\gamma = \sqrt{\frac{K+(\upsilon(3k+K)+\beta)KT}{(e-1)(1-\theta)(1+\upsilon)(1+4\upsilon k)kKT}\log\left(\frac{K}{k}\right)},$$

then an upper bound on the regret of ***dLET-Exp3.M*** *is,*

$$\overline{\mathcal{R}}_T \leq 2\Bigg((e-1)\Bigg(\frac{(4k+1)(3k+K)KT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{2K}{\delta}\right) + \frac{(3k+K)KT}{1-\theta}\log\left(\frac{K}{k}\right)\left(\frac{kKT}{\epsilon}\log\left(\frac{2K}{\delta}\right)\right)^{\frac{1}{2}} + \frac{2kK^2T}{1-\theta}\log\left(\frac{K}{k}\right) + \frac{4(3k+K)}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{2K}{\delta}\right)\left(\frac{kKT}{\epsilon}\log\left(\frac{2K}{\delta}\right)\right)^{\frac{1}{2}} + \frac{2K(4k+1)}{1-\theta}\log\left(\frac{K}{k}\right)\left(\frac{kKT}{\epsilon}\log\left(\frac{2K}{\delta}\right)\right)^{\frac{1}{2}} + \frac{8K}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{2K}{\delta}\right)\Bigg)\Bigg)^{\frac{1}{2}},$$

and holds for any assignment of rewards and any $T > 0$*, with probability at least* $1 - \delta/2K$*.*

Theorem 9, gives an upper bound on the expected total number of observations the player will make when following the **dLET-Exp3.M** approach, in terms that quantify the total length of consecutive rounds where an arm's feedback exceeds threshold and the complexity of the adversarial thresholding semi-bandit problem.

Theorem 9. *On average the total number of observations made from playing the* ***dLET-Exp3.M*** *algorithm is given by,*

$$\mathcal{M}_{dLET-Exp3.M} \leq H_\theta + kT + 2\sum_{t=1}^{T-1} Q_t,$$

Proof of Theorem 9. For bounding the expected total number of observations when playing **dLET-Exp3.M**, we consider the number of arms we are expected to observe in round $t = 1$, then define the expected number of observations in round $t = 2, \ldots, T$ in terms of round $t - 1$.

From Algorithm 11 we see that $\epsilon_{1,i} \leq 1$ for arms $i = 1, \ldots, K$, $|I_1| = k$ and $\mathcal{T}_1 = \emptyset$. Hence we have,

$$\mathcal{M}_1 \leq k,$$

where,

$$\mathcal{M}_t = \sum_{i=1}^{K} \mathbb{I}\{i \in \mathcal{T}_{t-1}\} + \sum_{i \in I_t \setminus \mathcal{T}_{t-1}} \epsilon_{t,i},$$

denotes the expected number of observations from playing **dLET-Exp3.M** in round t. Hence,

$$\sum_{t=1}^{T} \mathcal{M}_t \leq \sum_{t=2}^{T} \left(\sum_{i=1}^{K} \mathbb{I}\{i \in \mathcal{T}_{t-1}\} + \sum_{i \in I_t \setminus \mathcal{T}_{t-1}} \epsilon_{t,i} \right) + k,$$

where $\sum_{i \in I_t \setminus \mathcal{T}_{t-1}} \epsilon_{t,i}$ represents the expected number of queries made in rounds $t \in \{2, \ldots, T\}$ that is governed by arms not on the threshold list. Since arms are only added to the threshold list after being observed exceeding threshold then from (3.1),

$$q_t \geq \sum_{i=1}^{K} \mathbb{I}\{i \in \mathcal{T}_{t-1}\},$$

where $q_t = \sum_{i \in [K]} \mathbb{I}\{y_{t,i} > \theta\}$ relates the complexity of adversarial thresholding bandit problems with expected number of observations. Hence,

$$\mathcal{M}_{dLET-Exp3.M} \leq \sum_{t=2}^{T} q_t + \sum_{t=2}^{T} \sum_{i \in I_t \setminus \mathcal{T}_{t-1}} \epsilon_{t,i} + k,$$

where $\mathcal{M}_{dLET-Exp3.M} = \sum_{t \in [T]} \mathcal{M}_t$. Recalling from (3.1) and,

$$\sum_{t=1}^{T-1} q_t + q_{t+1} = \sum_{t=1}^{T-1} q_t + \sum_{t=2}^{T} q_t,$$

then by Theorem 2,

$$(1-\theta)H_\theta \geq \sum_{t=2}^{T} q_t + \sum_{t=1}^{T-1} q_t - 2Q_t.$$

Since we have that $\epsilon_{t,i} \leq 1$ for $t = 1, \ldots, T$ and $i = 1, \ldots, K$, this completes the proof. □

In this section, we prove an upper bound on the regret of the **dLET-Exp3.M** algorithm. We also prove a result, linking the average number of observations, the ATSBP hardness measure from Definition 2, and the total number of rounds with consecutive feedback exceeding threshold for every arm.

6.5 Summary

In this chapter we introduced the dynamic label efficient adversarial thresholding semi-bandit and presented a novel algorithm for selecting and querying arms over a finite time horizon, called **dLET-Exp3.M**. In particular, our approach places no restrictions on the number of arms that may exceed threshold each round. We insist only on the more relaxed Assumption 3, where arms observed exceeding threshold in round t are also observed in round $t+1$. The relaxation is compared with Assumption 2 used in Chapters 4 and 5, where the player knows the exact number of arms that will exceed threshold every round.

To compensate for the increased challenge of Assumption 2 not holding, we dynamically adapt the probability of querying selected arms. Since the sequence of Bernoulli random variables $B_{t,i}$ used to decide whether to query an arm or not is dependent on the history of observations for that arm, our theoretical guarantees do not rely on the expectation of $B_{t,i}$.

Our description of adversarial thresholding semi-bandits involving label efficient prediction has initially centred solely on the player making the decision to query. In Chapter 5, we discussed how label efficient bandits can be used to model applications with an element of unreliable data collection, where the environment controls the decision to allow an arm to be queried. The algorithm developed in this chapter can be used to model situations where a combination of both the environment and player control the

query decision. Under these circumstances, the environment (or adversary) sets the maximum probability of an arm being queried, $\max\{\epsilon_{t,i}\}$. This may depend on the operating conditions experienced or age of the sensor collecting the feedback. The player retains control over querying arms, dependent on previous observations.

Arms are selected for querying according to a weighted probability distribution over arms, based on **DepRound** as in Chapters 4 and 5. The biased reward estimates used to compute such weights are dependent on which partition of $[K]$ each arm belongs to. In each round, every arm belongs to one of four partitions: $\mathcal{A}_t$, $\mathcal{B}_t$, $\mathcal{C}_t$ and whichever arms are not contained in any of the previous three sets.

We prove that the regret of **dLET-Exp3.M** is $\mathcal{O}\Big(\sqrt{\frac{(k+K)kKT}{\epsilon(1-\theta)}\log\big(\frac{K}{k}\big)\log\big(\frac{K}{\delta}\big)}\Big)$ holds for any assignment of rewards and any $T > 0$, with high probability. The cost of not knowing how many arms will exceed threshold each round inhibits the asymptotic performance of **dLET-Exp3.M** only by a constant factor $\sqrt{k+K}$ in comparison to **LET-Exp3.M**. However, for finite time horizons, it is clear from Theorem 8 that theoretical performance guarantees are significantly affected by holding to Assumption 3 and not Assumption 2.

In Theorem 9 we give an upper bound on the expected total number of observations made from playing **dLET-Exp3.M** for T rounds. This result is given in terms of the complexity of the adversarial thresholding semi-bandit problem, which quantifies the cumulative frequency of arms exceeding threshold and a algorithm-specific term that is the cumulative total expected number of observations from querying arms not on the threshold list. It is also governed by the total number of consecutive rounds where feedback from an arm exceeds threshold, as a result of the threshold list.

We note from (6.26) that the value of W_{T+1} when the game has been played optimally, can only be achieved by playing every round $t = 1, \ldots, T$ optimally. This corresponds to the single policy of pulling only those arms exceeding threshold in every round from the space of possible policies where any number of arms from zero to K could be pulled over the sequence of T rounds.

One issue we can foresee is that the player will find difficulty in playing optimally when there are more than k arms exceeding threshold in round t but none of them were on the threshold list in the previous round.

We assume it is likely that arms exceeding threshold in the previous round probably have large weight and so it also highly likely that **DepRound** will repeatedly keep choosing those arms. This is satisfactory in terms of arms behaving above threshold having larger weights and wanting to definitely observe those, but can cause the algorithm (**DepRound** in particular) to get stuck in an apparent stationary point where the weights for some arms are so large relative to the remainder that the others never get a chance to be observed. This can be exploited by an intelligent adversary and leads nicely on to the next chapter.

Chapter 7

AliceBandit

The suite of algorithms proposed in Chapters 4, 5 and 6 have theoretically provable bounds on their performance that are sub-linear in the time horizon, where the player's decisions converge towards optimal decisions as T grows large. However, they all possess the same computational bottleneck due to their reliance on **DepRound** [33], to select the subset of arms I_t, using an importance-weighted probability distribution. We now pay particular attention to developing an algorithm that improves computational efficiency, does not need to maintain a distribution over all arms and achieves sub-linear regret when the hardness of the environment does not scale with the time horizon.

In Section 7.1, we introduce the *randomised round robin* selection policy that the player uses to decide which subset of arms to select each round. We also introduce a novel approach to the semi-bandit setting by randomising the subset size and describe frameworks for the finite time horizon and *anytime* setting, where an anytime algorithm can be run for even an unknown number of rounds. In Section 7.2, we give details of the **AliceBandit** algorithm, for which we present a theoretical analysis in Section 7.3.

7.1 Randomised round robin selection

In Chapters 4, 5 and 6, algorithms **T-Exp3.M**, **LET-Exp3.M** and **dLET-Exp3.M** used an exponentially-weighted distribution over all arms, $p_{t,1}, \ldots, p_{t,K}$ each round, for some integer $K > 1$ denoting the number of arms, pulling probability $p_{t,i}$, for arm $i \in \{1, \ldots, K\}$, in round $t \in \{1, \ldots, T\}$. The **DepRound** procedure is implemented

to select a subset of $k < K$ arms from the set $[K] = \{1, \ldots, K\}$. The computational complexity of these approaches is $\mathcal{O}(K(\log(k) + 1))$, per round (See Section 4.5).

It is well known that optimal performance can only be achieved by randomisation, when playing against an adversary (see [8] for details). Without any additional source of information, the seemingly appropriate course of action is to randomly select arms uniformly over the sequence of T rounds. A *randomised round robin* procedure selecting randomly-sizes subsets of arms each round, ensures that each arm is selected for observation randomly, at lower computational cost and decentralises the requirement to maintain a distribution over all arms. The observation decision is then determined by the history of feedback for each arm.

In this section, we pay particular attention to reducing the computational cost of selecting subsets from a centralised, weighted probability distribution over all arms. In particular, section 7.1.1 introduces the concept of *round robin sub-sequences* and we develop computationally efficient algorithms to ensure that every arm has an opportunity to be selected for observation.

7.1.1 Round robin sub-sequences

An adversary can inflict significant regret on to the player in the event that $q_t > k_t$, when Assumption 2 (see Section 4.1) holds and $q_t > k_t + |\mathcal{T}_{t-1}|$, when Assumption 3 (Section 4.1) holds, recalling that $q_t = \sum_{i \in [K]} \mathbb{I}\{x_{t,i} > \theta\}$ and until Chapter 6, the number of arms pulled has been fixed, $k_t = k$, for $t = 1, \ldots, T$. An extreme example of this is the situation where the environment sets the feedback to $y_{t,i} = 1$, for arms $i = 1, \ldots, K$, yet the player decides not to query any of the selected arms and the player has no arms on the threshold list from the previous round. To avoid this scenario, we allow the player to randomly choose the size of the subset of arms selected for observation, over a given sub-sequence of rounds.

We note that if the player chooses to select all arms every round and rely only on the observation decision, then $k_t = K$ for $t = 1, \ldots, T$. In the event that,

$$(|k_t - q_t| + 1) \sum_{i \in I_t^*} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\} \leq (K - q_t + 1) \sum_{i \in I_t} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\}, \tag{7.1}$$

where I_t^* denotes the optimal subset of arms in round t, then selecting every arm for observation will never be more advantageous. Observing optimal feedback is as a result of the player observing those arms with feedback exceeding threshold. Consider the case where (7.1) holds. Since,

$$\sum_{i \in I_t^*} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\} = \sum_{i=1}^{K} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\},$$

then,

$$\frac{1}{K - q_t + 1} \sum_{i=1}^{K} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\} \leq \frac{1}{|k_t - q_t| + 1} \sum_{i \in I_t} x_{t,i} \mathbb{I}\{x_{t,i} > \theta\}.$$

Hence, the cumulative feedback rewarded to the player from selecting every arm is never bettered when (7.1) holds and in particular, when q_t is small. The randomised round robin selection enables any number of K arms to be potentially observed and the use of a threshold list (introduced in Chapter 6) enables arms exceeding threshold over a sequence of consecutive rounds to be observed. This mitigates the problem of more arms exceeding threshold than the fixed size of k arms pulled (used in Chapters 4 and 5).

The **S3**$_m$ algorithm (see Algorithm 12), randomly generates the m^{th} set $\{\mathcal{N}_m\}_{\tau(t)=1}^{|\mathcal{N}_m|} = \{\mathcal{N}_{m,1}, \ldots, \mathcal{N}_{m,|\mathcal{N}_m|}\}$ of integers, where $T \leq m \leq KT$, counts the number of calls for Algorithm 12, up to round t. The number of rounds in the m^{th} sub-sequence is denoted by $|\mathcal{N}_m|$ and,

$$\tau(t+1) = \begin{cases} 1 & \text{if } \sum_{s=1}^{t} k_s \mod K = 0 \\ \tau(t) + 1 & \text{otherwise,} \end{cases}$$

where $k_t = \mathcal{N}_{m,\tau(t)}$, connecting the notation for the number of arms selected in round t, from Chapter 6, with sub-sequences. The position in the m^{th} sub-sequence is denoted by $\tau(t)$, where $\tau(1) = 1$. If $\sum_{s=1}^{t} k_t \mod K = 0$, then Algorithm 12 is called. From this, we can now define I_t, the subset of selected arms in round t, in terms of the $\tau(t)^{th}$ position, in the m^{th} randomised round robin sub-sequence.

Algorithm 12 $S3_m$, m^{th} Sub-sequence sizes algorithm

Input: Number of arms $K \in \mathbb{N}$, sub-sequence number, $m \in \mathbb{N}$, $\sigma = 0$, $n = 1$, $\mathcal{N}_m = \emptyset$.

while $\sigma < K$ **do**
  Draw $\mathcal{N}_{m,n} \sim \mathcal{U}([K - \sigma])$
  Set $\sigma = \sigma + \mathcal{N}_{m,n}$
  if $\sigma < K - 1$ **then**
    $\mathcal{N}_m = \mathcal{N}_m \cup \{\mathcal{N}_{m,n}\}$
  else
    $\mathcal{N}_m = \mathcal{N}_m \cup \{\mathcal{N}_{m,n} + 1\}$
  Set $n = n + 1$

Output: $\mathcal{N}_m = \{\mathcal{N}_{m,1}, \ldots, \mathcal{N}_{m,|\mathcal{N}_m|}\}$

Definition 5. *The subset of selected arms $i \in I_t$ is defined as the randomised round robin subset of arms in position $\tau(t)$, from the m^{th} sub-sequence.*

$$\mathcal{I}_{m,\tau(t)} = \left\{ i \sim \mathcal{U}([K]) \;\middle|\; \bigcap_{j=1}^{\tau(t)} \mathcal{I}_{m,j} = \emptyset, |\mathcal{I}_{m,\tau(t)}| = \mathcal{N}_{m,\tau(t)} \right\}.$$

The anytime version of **AliceBandit** (**AliceBandit**$(\mathcal{I}_{m,\tau(t)})$) is run each round on the subset $\mathcal{I}_{m,\tau(t)}$, depending on the m^{th} sub-sequence and position $\tau(t)$, in that sub-sequence. When $\sum_{s=1}^{t} k_t \mod K = 0$, $\mathbf{S3}_m$ is called and $I_t = \mathcal{I}_{m,1}$. For the subsequent $|\mathcal{N}_m| - 1$ rounds where $\sum_{s=1}^{t} k_t \mod K > 0$, **AliceBandit** is run for rounds $\tau(t) = 2, \ldots, |\mathcal{N}_m|$ on the remainder of the same sub-sequence.

Since algorithm $\mathbf{S3}_m$ randomly generates a set of integers, the number of times $\mathbf{S3}_m$ is called throughout the time horizon m, is also a random number, even if T is known. The anytime selection framework for **AliceBandit** is run without knowledge of the time horizon T (see Algorithm 13).

The finite time horizon version for selecting randomly-sized subsets of arms is called *finite-time arm selection* (**FTAS**), repeatedly calling $\mathbf{S3}_m$ until T subsets have been generated (see Algorithm 14). Since the randomised round robin procedure is independent of the decision to query arms, **FTAS** can be run in advance. This is analogous to the adversarial thresholding semi-bandit game being played against an oblivious adversary. From Algorithm 14, we see that since $1 \leq |\mathcal{N}_m| \leq K$, **FTAS** can generate up to $T + K - 1$ subsets, but only outputs the first T.

Algorithm 13 AliceBandit($\mathcal{I}_{m,\tau(t)}$), Anytime AliceBandit framework

Input: Number of arms K, $\tau(1) = 1$, $m = 1$, $\mathbf{S3}_1 = \{\mathcal{N}_{1,1}, \ldots, \mathcal{N}_{1,|\mathcal{N}_1|}\}$

for $t = 1, \ldots$ **do**
  Let $k_t = \mathcal{N}_{m,\tau(t)}$
  AliceBandit$(\mathcal{I}_{m,\tau(t)})$
  if $\sum_{s=1}^{t} k_t \mod K = 0$ **then**
    Set $m = m + 1$
    $\{\mathcal{N}_{m,1}, \ldots, \mathcal{N}_{m,|\mathcal{N}_m|}\} = \mathbf{S3}_m$
    Set $\tau(t+1) = 1$
    for $\tau(t) = 1, \ldots, |\mathcal{N}_m|$ **do**
      $\mathcal{I}_{m,\tau(t)} = \emptyset$
      for $j = 1, \ldots, \mathcal{N}_{m,\tau(t)}$ **do**
        $\chi_j \sim \mathcal{U}\big([K] \backslash \bigcup_{r=1}^{\tau(t)} \mathcal{I}_{m,r}\big)$
        $\mathcal{I}_{m,\tau(t)} = \mathcal{I}_{m,\tau(t)} \cup \{\chi_j\}$
  else
    $\tau(t+1) = \tau(t) + 1$

Algorithm 14 FTAS, Finite time arm selection algorithm

Input: Number of arms $K \in \mathbb{N}$, number of rounds $T \in \mathbb{N}$, $t = 1$, $m = 1$.

while $t < T$ **do**
  $\{\mathcal{N}_{m,1}, \ldots, \mathcal{N}_{m,|\mathcal{N}|}\} = \mathbf{S3}_m$
  for $\tau(t) = 1, \ldots, |\mathcal{N}_m|$ **do**
    Set $I_{m,\tau(t)} = \emptyset$
    for $j = 1, \ldots, \mathcal{N}_{m,\tau(t)}$ **do**
      $\chi_j \sim \mathcal{U}\big([K] \backslash \bigcup_{r=1}^{\tau(t)} \mathcal{I}_{m,r}\big)$
      $\mathcal{I}_{m,\tau(t)} = \mathcal{I}_{m,\tau(t)} \cup \{\chi_j\}$
    $I_t = \mathcal{I}_{m,\tau(t)}$
    Set $t = t + 1$
  Set $m = m + 1$

Output: $I = \{I_1, \ldots, I_T\}$

In terms of computational complexity, since we have $\sum_{\tau(t)=1}^{|\mathcal{N}_m|} \mathcal{N}_{m,\tau(t)} = K$, the computational cost of running **S3**$_m$ (Algorithm 12) is $\mathcal{O}(K)$ and **FTAS** (Algorithm 14) is $\mathcal{O}(KT)$. This is due to the randomised round robin, in which every arm is selected exactly once in each sub-sequence. Hence the computational cost of running **AliceBandit**$(\mathcal{I}_{m,\tau(t)})$, in an anytime setting (Algorithm 13) is also linear in the number of arms, if the complexity of **AliceBandit** is no greater than linear in computational complexity itself (see Section 7.2 for further details).

In this section, we have developed the procedures for selecting subsets of arms, using randomised round robin sub-sequences. Several algorithms have been introduced for selecting subsets of arms, including a framework for running an adversarial thresholding semi-bandit without knowledge of the time horizon, T. The computational cost of each algorithm is also considered.

7.2 The AliceBandit algorithm

We recall from Section 2.4, the classical label efficient prediction player chooses an arm in round t, incurs loss (or reward depending on the player's point of view) associated with pulling that arm (but not necessarily revealed by the adversary) and then decides whether to query all arms or not. The decision to query is based on a sequence of Bernoulli random variables $B_1, \ldots, B_T$, such that $\mathbb{P}\big[B_t = 1\big] = \epsilon$, for some $\epsilon \in [0, 1]$. By querying the entire feedback (or *label*) for all arms, the player can determine exactly how much reward is available for every arm once the player has made their decision. To compensate for the player only expecting to see ϵKT arms over the entire time horizon, she uses an unbiased estimate of the rewards, to guide her choice of arm in the next round.

In Chapter 6, we introduced the concept of dynamically adapting the Bernoulli parameter $\epsilon_{t,i}$, based on the history of previous observations for arm $i \in \{1, \ldots, K\}$, prior to round $t \in \{1, \ldots, T\}$, where $\mathbb{P}\big[B_{t,i} = 1\big] = \epsilon_{t,i}$. The algorithms proposed in Chapters 5 and 6 both relied on **DepRound** [33], and an exponentially-weighted distribution over arms for selecting a subset each round. We now introduce an algorithm called **AliceBandit**, which incurs a lower computational cost, compared with previous finite time horizon algorithms in this thesis. As with **dLET-Exp3.M**, **AliceBandit** uses a threshold list $\mathcal{T}_t$, that relies on Assumption 3 holding. However, **AliceBandit** uses

the output from **FTAS** for the subset I_t, of arms selected and populates a set of arms pulled, based on the realisation of a Bernoulli random variable $B_{t,i}$, denoted by $\mathcal{P}_t$.

In each round, **AliceBandit** is comprised of three modules, dealing with (i) arms belonging to the threshold list from the previous round, $\mathcal{T}_{t-1}$, (ii) arms selected using the randomised round robin and (iii) arms that have not been observed for at least a given number of consecutive rounds:

- **Threshold list arms:** The first module pulls every arm on the threshold list from the previous round. The arms $i \in \mathcal{T}_{t-1}$ observed exceeding threshold in round t are stored on the threshold list $\mathcal{T}_t$, the Bernoulli parameter $p_{t,i}$ is set equal to $p_{max,i}$, where $\mathbb{P}\big[B_{t,i} = 1\big] = p_{t,i}$ and $p_{max,i}$ denotes the maximum probability of the feedback from arm i being revealed to the player ($p_{max,i}$ can be pre-set either by the player or adversary). The player is rewarded with the feedback for the arms observed in I_t. For arms $i \in \mathcal{T}_{t-1}$ within threshold in round t, the player receives no information regarding feedback from the adversary or reward but can count the number of times she makes such a mistake, denoted by $N_{t,i}$. The Bernoulli parameter of arm i decays exponentially as a function of $N_{t,i}$.

- **Selected arms:** The second module takes the t^{th} subset I_t from **FTAS** and decides whether the query each arm $i \in I_t$, using the Bernoulli random variable $B_{t,i}$. Arms queried and exceeding threshold are added to the current threshold list and the probability of accessing feedback is reset to the maximum for arm i, $p_{t,i} = p_{max,i}$. Otherwise, $N_{t,i}$ is incremented and $p_{t,i}$ is decayed.

- **Forgotten arms:** The final module of **AliceBandit** pulls any arm that has not been pulled for at least t_{max} rounds and follows the same procedure depending on which arms exceed threshold or not. This ensures that even if the likelihood of querying an arm becomes very small, it is not forgotten about completely. The size of t_{max} governs the sensitivity of the algorithm. Larger values of t_{max} mean that **AliceBandit** pulls fewer arms overall increasing efficiency but increase the probability of missing arms exceeding threshold. Smaller values of t_{max} lead potentially to more unnecessary arm pulls.

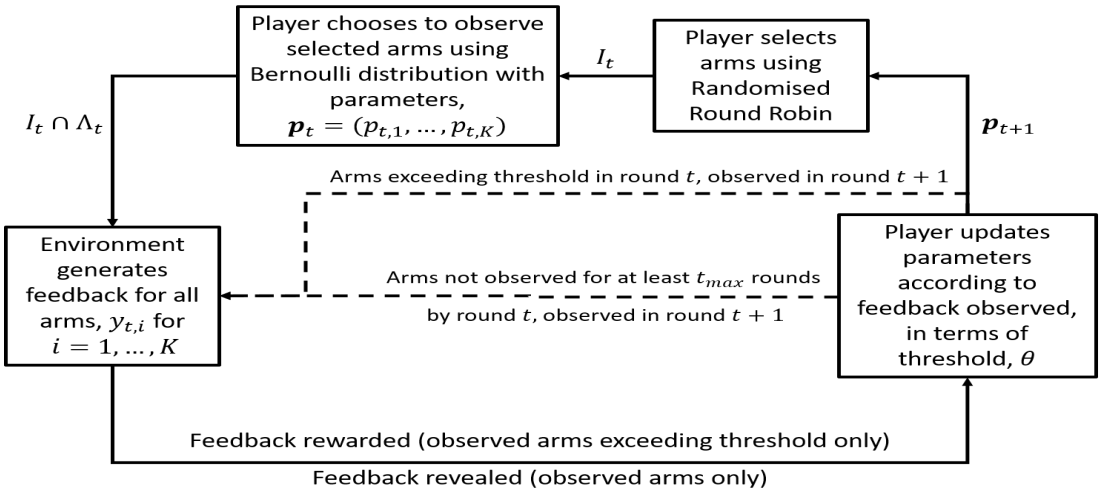


Figure 7.1: A schematic diagram of the AliceBandit algorithm.

In the penultimate paragraph of Section 7.1.1, we determined that the computational complexity of algorithms $\mathbf{S3}_m$ and **FTAS**, were $\mathcal{O}(K)$ and $\mathcal{O}(KT)$, respectively. Since each of the three modules introduced within **AliceBandit** are at most a linear function of K steps each, then the computational cost of the **AliceBandit** algorithm is $\mathcal{O}(K)$, in every round and note that **FTAS** can be run prior to running **AliceBandit**.

The pseudo-code for **AliceBandit** is given in Algorithm 15. A schematic diagram of the learning cycle for the algorithm is also given (see Figure 7.1). The arm-selection process now involves a randomised round robin and the observation-decision is governed by a dynamically updated Bernoulli distribution.

7.3 Analysis of AliceBandit

We now analyse the performance of the **AliceBandit** algorithm. Since the player makes decisions about observing arms, depending on the number of mistakes made by the player since their last correct pull, the analysis takes on a different flavour compared with the analysis given in Chapters 4, 5 and 6. The following result provides a theoretical bound for the performance of **AliceBandit**, in terms of the loss incurred by the player from observing arms with feedback within threshold.

Algorithm 15 AliceBandit

Input: Number of arms K, number of rounds T, I_t for $t = 1, \dots, T$, $\mathcal{T}_0 = \emptyset$, $p_{0,i} = p_{max,i}$, $N_{0,i} = 1$, $n_i = 0$ for $i = 1, \dots, K$, maximum number of rounds between pulls, t_{max}, $\eta \in (0, 1]$, $\theta \in (0, 1)$

for $t = 1, \dots, T$ **do**
  Set $\mathcal{P}_t = \emptyset$
  Set $\mathcal{T}_t = \emptyset$
  for $i \in \mathcal{T}_{t-1}$ **do**
    Adversary returns feedback $x_{t,i}$
    Set $n_i = t$
    if $x_{t,i} > \theta$ **then**
      Let $\mathcal{T}_t = \mathcal{T}_t \cup \{i\}$
      Set $p_{t,i} = p_{max,i}$
    else
      Set $N_{t,i} = N_{t,i} + 1$
      Set $p_{t,i} = p_{t-1,i} \exp\left(-\frac{\eta N_{t,i}}{K}\right)$
  for $i \in I_t$ **do**
    Draw Bernoulli random variables $B_{t,i}$ such that $\mathbb{P}\big[B_{t,i} = 1\big] = p_{t,i}$
    if $B_{t,i} = 1$ **then**
      Adversary returns feedback $x_{t,i}$
      if $i \in I_t \backslash \mathcal{T}_{t-1}$ **then**
        Set $n_i = t$
      Set $\mathcal{P}_t = \mathcal{P}_t \cup \{i\}$
      if $x_{t,i} > \theta$ **then**
        Let $\mathcal{T}_t = \mathcal{T}_t \cup \{i\}$
        Set $p_{t,i} = p_{max,i}$
      else
        Set $N_{t,i} = N_{t,i} + 1$
        Set $p_{t,i} = p_{t-1,i} \exp\left(-\frac{\eta N_{t,i}}{K}\right)$
  for $i \in [K]$ **do**
    if $i \notin \mathcal{P}_t \cup \mathcal{T}_{t-1} \wedge t - n_i > t_{max}$ **then**
      Adversary returns feedback $x_{t,i}$
      Set $n_i = t$
      if $\theta < x_{t,i}$ **then**
        Let $\mathcal{T}_t = \mathcal{T}_t \cup \{i\}$
        Set $p_{t,i} = p_{max,i}$
      else
        Set $N_{t,i} = N_{t,i} + 1$
        Set $p_{t,i} = p_{t-1,i} \exp\left(-\frac{\eta N_{t,i}}{K}\right)$

Theorem 10. *Let $K, T > 0$ be integers, $0 < \eta \leq 1$ and $0 < \theta < 1$ be a real numbers and H_θ be the ATSBP hardness from Definition 2, then the ATSBP regret of Algorithm 15 is given by*

$$\overline{\mathcal{R}}_T \leq \frac{2KT}{\eta} + \frac{3}{2}\eta(1-\theta)H_\theta,$$

and holds for any assignment of feedback.

Notation introduced in Section 7.1.1, is used in the proof of Theorem 10, identifying the subset of arms with observed feedback, in terms of position $\tau(t)$ of the m^{th} subsequence, up to round t. The ATSBP regret defined in Chapter 4 (Definition 1), is transformed to measure the cumulative loss of the player observing feedback within threshold and a bound is then placed on the loss incurred by the player, for pulling an arm when the feedback is within threshold. The upper bound given in Theorem 10, is also a function of the problem complexity, H_θ.

Proof of Theorem 10. We begin by defining the loss incurred by the adversarial thresholding semi-bandit player and then restate Definition 1, given in Chapter 4. A bound is then placed on the expected number of observations revealed to the player each round.

Let the feedback observed by the player in a given round, be the feedback from an arm that has been selected by the randomised round robin procedure and observed according to realisation of an associated Bernoulli random variable, or the feedback from an arm that was assigned to the threshold list in the previous round,

$$x_{t,i} = \begin{cases} y_{t,i} & \text{if } y_{t,i} > \theta, i \in \Lambda_t \cup \mathcal{T}_{t-1} \\ 0 & \text{otherwise,} \end{cases} \tag{7.2}$$

where $x_{t,i} \in [0,1]$. Let the ATSBP loss be denoted $\ell_{t,i} = 1 - x_{t,i}$, then from (7.2) we have,

$$\ell_{t,i} = \begin{cases} 1 - y_{t,i} & \text{if } y_{t,i} > \theta, i \in I_t \\ 1 & \text{otherwise,} \end{cases} \tag{7.3}$$

where $I_t = \Lambda_t \cup \mathcal{T}_t$, denotes the subset of arms that the player receives feedback for, even if the feedback is zero (including no feedback). In bandit terminology, I_t denotes

the set of arms pulled in round t. From (7.3), we see that maximum loss is assigned to arms pulled and returning feedback within threshold. Maximum loss is also given to all arms not observed, encouraging the player to only observe arms with feedback exceeding threshold.

From Definition 1, the adversarial thresholding semi-bandit regret is given by,

$$\overline{\mathcal{R}}_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in I_t^*} x_{t,i} - \sum_{t-1}^{T}\frac{1}{\Delta_t}\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\}\right],$$

where $\Delta_t = |\kappa_t - q_t| + 1$. The number of arms pulled is denoted by $\kappa_t = \sum_{i\in[K]}\mathbb{I}\{i \in I_t\}$ and $q_t = \sum_{i\in[K]}\mathbb{I}\{y_{t,i} > \theta\}$ denotes the number of arms with feedback exceeding threshold in round t. This gives,

$$q_t = \sum_{i=1}^{K}\mathbb{I}\{y_{t,i} > \theta\} = \sum_{i\in I_t^*}\mathbb{I}\{x_{t,i} > \theta\},$$

where every arm $i \in I_t^*$, is observed and exceeding threshold, $x_{t,i} > \theta$. By (7.3), the cumulative feedback from optimally pulling only those arms exceeding threshold gives,

$$\begin{aligned}\sum_{i\in I_t^*} x_{t,i} &= \sum_{i\in I_t^*}(1-\ell_{t,i}) \\ &= q_t - \sum_{i\in I_t^*}\ell_{t,i}. \end{aligned} \tag{7.4}$$

We can also state from (7.3), the cumulative feedback received by the player from pulled arms in terms of the loss incurred by the player and the number of arms pulled,

$$\begin{aligned}\sum_{i\in I_t} x_{t,i}\mathbb{I}\{x_{t,i} > \theta\} &= \sum_{i\in I_t}(1-\ell_{t,i})\mathbb{I}\{x_{t,i} > \theta\} + \mathbb{I}\{x_{t,i} \le \theta\} \\ &= \kappa_t - \sum_{i\in I_t}\ell_{t,i}\mathbb{I}\{x_{t,i} > \theta\}. \end{aligned} \tag{7.5}$$

Substituting (7.4) and (7.5) into Definition 1 gives,

$$\begin{aligned}\overline{\mathcal{R}}_T &= \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\Delta_t}\sum_{i\in I_t}\ell_{t,i}\mathbb{I}\{x_{t,i} > \theta\} - \sum_{t=1}^{T}\left(\frac{1}{\Delta_t}\kappa_t - q_t\right) - \sum_{t=1}^{T}\sum_{i\in I_t^*}\ell_{t,i}\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{\Delta_t}\sum_{i\in I_t}\ell_{t,i}\mathbb{I}\{x_{t,i} > \theta\} + q_t\right]. \end{aligned} \tag{7.6}$$

Since $\Delta_t \leq 1$, we have,

$$\mathbb{E}\left[\frac{1}{\Delta_t}\sum_{i\in I_t}\ell_{t,i}\mathbb{I}\{x_{t,i}>\theta\}\right] \leq \mathbb{E}\left[\frac{1}{\Delta_t}\kappa_t\right] \leq \mathbb{E}[\kappa_t]. \tag{7.7}$$

From (7.7), the expected number of arms observed in a given round, independently depends on the probability of being selected, the probability of being observed and vicariously from being on the threshold list in the previous round,

$$\begin{aligned}
\mathbb{E}[\kappa_t] &= \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]p_{t,i} \\
&= \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]p_{max,i}\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}>\theta\} \\
&\quad + \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]p_{t-1,i}\exp\left(-\frac{\eta}{K}N_{t-1,i}\right)\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}\leq\theta\} \\
&\quad + \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]p_{t-1,i}\mathbb{I}\{i\notin\mathcal{I}_{m,\tau(t)}\} \qquad (7.8) \\
&\leq \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}>\theta\} \\
&\quad + \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]\exp\left(-\frac{\eta}{K}N_{t-1,i}\right)\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}\leq\theta\} \\
&\quad + \sum_{i=1}^{K}\mathbb{P}\big[i\in\mathcal{I}_{m,\tau(t)}\big]\mathbb{I}\{i\notin\mathcal{I}_{m,\tau(t)}\} \qquad (7.9) \\
&\leq \sum_{i=1}^{K}\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}>\theta\} \\
&\quad + \frac{K}{\eta+K}\sum_{i=1}^{K}\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i}\leq\theta\} \\
&\quad + \sum_{i=1}^{K}\mathbb{I}\{i\notin\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{y_{t,i}>\theta\} + \mathbb{I}\{i\notin\mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{y_{t,i}\leq\theta\} \qquad (7.10) \\
&\leq q_t + \frac{K}{\eta+K}\sum_{i=1}^{K}\mathbb{I}\{i\in\mathcal{I}_{m,\tau(t)}\} + \sum_{i=1}^{K}\mathbb{I}\{i\notin\mathcal{I}_{m,\tau(t)}\} \qquad (7.11)
\end{aligned}$$

where (7.8) holds by the update rules given in Algorithm 15, (7.9) holds since $p_{t-1,i} \leq p_{max,i} \leq 1$ for $i = 1,\ldots,K$ and $t = 2,\ldots,T$.

We have that (7.10) holds since $\exp(-a) < 1/(1+a)$ for $a > -1$ and $\mathbb{P}\big[i \in \mathcal{I}_{m,\tau(t)}\big] \leq 1$ and (7.11) holds since $\sum_{i\in[K]} \mathbb{I}\{i \in \mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{x_{t,i} > \theta\} + \mathbb{I}\{i \notin \mathcal{I}_{m,\tau(t)}\}\mathbb{I}\{y_{t,i} > \theta\} = q_t$. Summing (7.11) over $t = 1, \ldots, T$ gives,

$$\sum_{t=1}^{T} \mathbb{E}[\kappa_t] \leq \frac{KT}{\eta} + \sum_{t=1}^{T} q_t, \tag{7.12}$$

where $K \leq K/\eta$, for $0 < \eta \leq 1$. From Definition 2, where $d_{t,i} = |y_{t,i} - y_{t+1,i}| + 1 \geq 1$, we have,

$$H_\theta \leq \frac{2}{\eta(1-\theta)} \sum_{t=1}^{T-1} q_t + q_{t+1} - 2Q_t,$$

for $0 < \eta \leq 1$. This gives,

$$\begin{aligned} \frac{\eta(1-\theta)H_\theta}{2} &\leq \sum_{t=1}^{T-1} q_t + q_{t+1} - 2Q_t \\ &= -q_1 - q_T + 2\sum_{t=1}^{T} q_t - 2\sum_{t=1}^{T-1} Q_t. \end{aligned} \tag{7.13}$$

Also from Definition 2, we have $d_{t,i} \leq 2$. This gives,

$$\eta(1-\theta)H_\theta \geq \sum_{t=1}^{T-1} q_t + q_{t+1} - 2Q_t. \tag{7.14}$$

Combining (7.13) and (7.14) gives,

$$\frac{\eta(1-\theta)H_\theta}{2} \leq q_1 + q_T + 2\eta(1-\theta)H_\theta - 2\sum_{t=1}^{T} q_t.$$

Hence,

$$\sum_{t=1}^{T} q_t \leq K + \frac{3}{4}\eta(1-\theta)H_\theta, \tag{7.15}$$

where $(q_1 + q_T)/2 \leq K$. The proof is completed by combining (7.6), (7.12) and (7.15), and since $K \leq \frac{KT}{\eta}$ for $T > 0$ and $0 < \eta \leq 1$. □

Careful tuning of the parameter η, enables the ATSBP regret of **AliceBandit** to be comparable to the algorithms developed in previous chapters.

Corollary 10.1. *For threshold $\frac{1}{4} < \theta < 1$ and ATSBP hardness H_θ, set,*

$$\eta = \sqrt{\frac{8KT}{3(1-\theta)H_\theta}}.$$

*Then the adversarial thresholding semi-bandit regret of **AliceBandit**,*

$$\overline{\mathcal{R}}_T \leq \sqrt{6KH_\theta T},$$

holds for any assignment of feedback and $T \geq \frac{3}{4\theta-1}$.

An analysis of the theoretical performance of the algorithm developed in the previous section has been carried out. Adversarial thresholding semi-bandit regret has been defined from considering the loss incurred by a player after selecting and observing feedback from a subset of arms. A theoretical bound on the performance of **AliceBandit** has been proved in terms of the time horizon T, the total number of arms K and the hardness of ATSBP, H_θ, measuring the frequency with which feedback deviates about some threshold and the size of such deviations, for every arm in every round.

7.4 Summary

In this section, we introduced the concept of independently selecting subsets of arms each round, using a randomised round robin procedure (see Section 7.1) and allowing the observation of selected arms to be decided by the realisation of a Bernoulli random variable, with the parameter as a function of the history of observations from selected arms.

In Section 7.2, the mathematical framework for the running of the algorithms developed in this chapter, is presented with and without knowledge of the time horizon, T. The computational complexity of our framework has been discussed in detail and demonstrates a reduction in computational cost when compared with the algorithms developed in previous chapters. The **AliceBandit** adversarial thresholding semi-bandit algorithm is shown to have a computational cost of $\mathcal{O}(K)$.

In Section 7.3, a theoretical bound on the performance of **AliceBandit** is proved in terms of the time horizon, threshold, number of arms and ATSBP hardness. In Theorem 10, we show that the regret of **AliceBandit** is $\mathcal{O}\big(\sqrt{KH_\theta T}\big)$, where H_θ is a measure of ATSBP hardness. The regret of **AliceBandit** is dependent on the adversarial nature of the environment. If H_θ scales with the time horizon T, then the ATSBP regret of **AliceBandit** is linear in T. Environments where H_θ is independent of T, achieve sub-linear regret. Having comparable theoretical performance and significantly lower computation cost, **AliceBandit** satisfies Research specifications 1 and 3.

Chapter 8

Online monitoring of ALICE detector controls at the LHC

In Chapters 4, 5, 6 and 7, we developed algorithms for the adversarial thresholding semi-bandit problem, introduced in Chapter 3. The algorithms have been analysed in terms of their theoretical performance and computational complexity, with respect to Research specifications 1 and 3, respectively (See Section 1.3). We now draw attention to the empirical performance of each algorithm (Research specification 2) and their ability to solve a specific application problem (Research specification 4).

In this chapter, we focus on the problem of monitoring electronic device power channels for a particular detector on ALICE, the heavy-ion experiment on the LHC particle accelerator at CERN, the European Organisation for Nuclear Research. A particular operational challenge in the control of high-energy physics experiments is the management of device power channels (see Section 1.2 for further details). Excessive current flow in a single electronic component on the detector is potentially damaging to many others. The detector automatically *trips* (switches off) a device channel when current flow through a device (measured in µA) is detected exceeding given limits, to prevent damage. Unfortunately, this has a negative impact on the experiments' capacity to collect data. Adversarial thresholding semi-bandits are applied to the problem of efficiently monitoring current flow from multiple electronic device channels on a high-energy physics experiment, with the aim of only observing channels when their associated current exceeds (or deviates by more than) a given threshold.

In Section 8.1, we present the particular detector on ALICE used in our application and give a brief overview of the detector's characteristics and purpose. In Section 8.2, we discuss how the experiment is controlled for consistently safe operation and preparations for the next extended period of data-taking are reviewed in Section 8.3. Section 8.4 describes the experimental environment used for carrying out simulations and generating the results used in subsequent analyses, giving the reader an opportunity to perform experiments themselves. In Section 8.4.1, we describe the salient characteristics of the datasets kindly provided by the ALICE DCS team, including any necessary pre-processing and several channels are identified as potentially being of interest to a DCS operator in Section 8.4.2. In Section 8.5, we discuss the metrics used to determine how Research specifications 2 and 4 are satisfied. We present the results from our empirical investigation in Section 8.6 and consider how effectively each algorithm achieves specifications 2 and 4. Finally, we summarise our findings in Section 8.7.

8.1 Transition Radiation Detector

Many of the decaying particles that originate from heavy-ion collisions such as Pb-Pb, involve electrons. Hence, electron identification is an essential requirement, particularly when taking into account the large number of fragments that result from colliding large nuclei (approximately 3000). The ALICE Transition Radiation Detector (TRD) is designed to carry out this task. The TRD has a modular construction with each module consisting of a radiator, drift region and a gaseous xenon chamber for detecting Transition Radiation (TR) (see Figure 8.1). When charged particles cross the boundary between layers of material with different dielectric fields, TR may be produced. The probability of TR occurring increases with the ratio between the particle's energy and it's rest energy. The ultra-relativistic energies experienced in ALICE have a ratio exceeding 1000. This enables the detection of electrons with high transverse momentum and increases the discrimination sensitivity between electrons and pions (see [4] for further details). The beam pipe is covered by five separate stacks of chambers in the longitudinal beam direction and six chambers are layered together to form an azimuth section. Each section is called a super-module and eighteen super-modules surround the beam pipe (labelled $0 - 17$ in Figure 8.2).

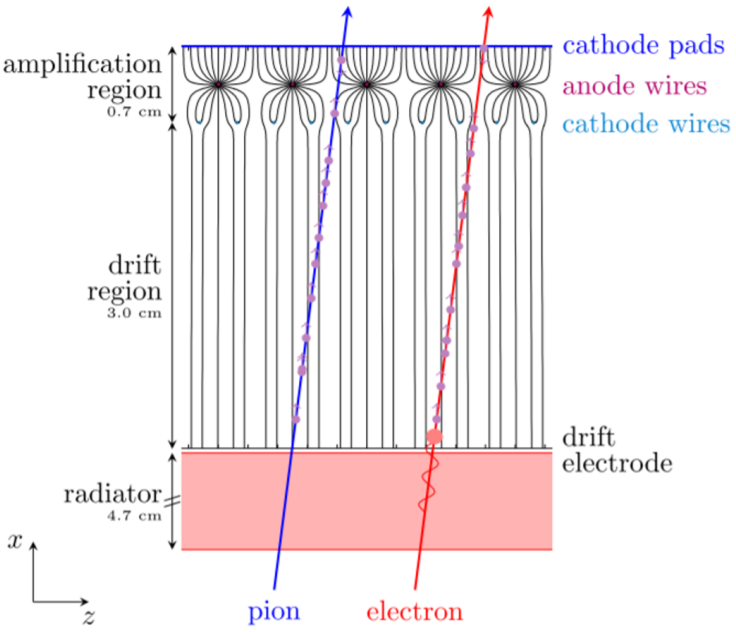


Figure 8.1: Schematic cross-section of a TRD chamber, illustrating a pion and electron local track segment. Large energy deposition due to TR photon absorption (indicated by the large red circle in the drift region) is used to distinguish electron tracks. The solid drift lines are calculated algorithmically and relate to the drift voltages set for nominal chamber operation, [4].

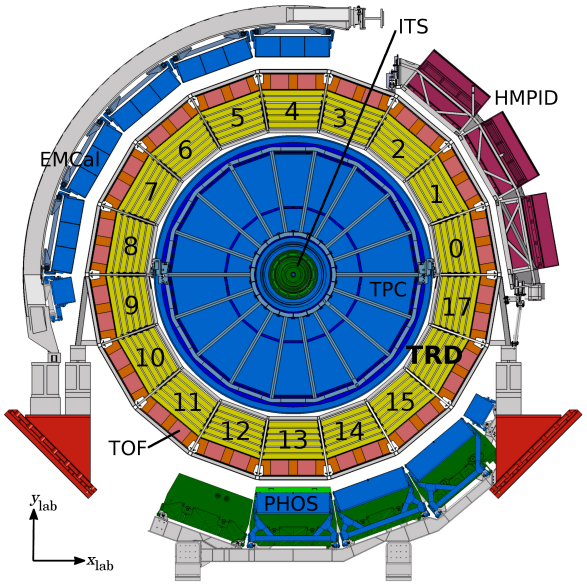


Figure 8.2: A cross-section schematic of ALICE, perpendicular to the LHC beam. The central barrel detectors are located within a solenoid magnet and provide a magnetic flux density of $B = 0.5\,T$, parallel to the beam direction, [4].

For optimised performance of each chamber, a variable negative (necessary to create a drift field) drift high voltage (HV) of order $-2.5\,\mathrm{kV}$ and a variable Anode HV $\sim 1.9\,\mathrm{kV}$ must be applied to each detector module. Variation of both voltages ensures the proposed concentration of gas mixtures in the drift chamber. A master/slave HV distribution system (HVDS) [57], supplies the necessary Anode and drift voltages to every chamber. Each supply channel can be operated independently by switching on/off, regulating voltage, setting maximum current thresholds, ramping up/down procedures and general trend monitoring of both voltage and current.

Much of the TRD infrastructure requires monitoring and service for their operation, wherever accessible. Prior to Long Shutdown 2 (LS2), which began in December 2018 and is scheduled until 2021, ALICE was collecting $1\,\mathrm{nb}^{-1}$ (per nano barn) of Pb-Pb collisions, corresponding to 1 event per $10^{-37}\,\mathrm{m}$ (a $barn = 10^{-28}\,\mathrm{m}$, denotes a measure of area commonly used in particle physics) and a peak luminosity of $L = 10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, corresponding to a collision rate of approximately 8 kHz, [63]. Prior to the previous Long Shutdown, ALICE was limited to reading 500 Hz of Pb-Pb collisions. Post-LS2 (RUN3), the experiment is expected to collect more than $10\,\mathrm{nb}^{-1}$ and with luminosities up to six times larger than before, corresponding to interaction rates of up to 50 kHz. Planning is also in place for collecting $6\,\mathrm{pb}^{-1}$ of proton-ion collisions, with rates up to 200 kHz.

Due to the much higher interaction rates expected after LS2, many of the triggering systems are inefficient for studying physics characterised by high signal-to-background ratios. This has imposed a paradigm shift towards continuous read-out of many detectors on ALICE. Hence, the emergence of a need for online calibration and continuous monitoring of the Detector Control Systems (DCS).

8.2 TRD Detector Controls

The ALICE Detector Control System (DCS) is designed to ensure the safe and correct operation of the experiment. It is the only system to service the experiment throughout the past decade. During this time, the ALICE DCS has guaranteed stable experimental conditions, essential for validating analysis of the physics data. The DCS conditions data is archived in the ALICE central database periodically at the end of each data-taking *run* (defined specifically as the period of time collecting data).

The DCS provides a user interface between LHC operators and each experimental device as part of the Experimental Control System (ECS) and can control the entire detector remotely from a single work station. The DCS logic architecture comprises three levels. The *field layer* collects information directly from each device and is responsible for providing services to each device. The *control layer* is responsible for broadcasting control commands. The machines used to execute DCS commands on the control layer are referred to as Worker Nodes (WN). However, WNs require instruction from Operator Nodes (ONs) on the top level *supervision layer*. The shift towards continuously-triggered operation during the upgrade phase between RUN2 and RUN3 has required the DCS to undergo a complete data-flow re-design.

An objective of the ALICE TRD DCS is to manage the operation of the HVDS. In this setting, the TRD DCS field layer corresponds to the HV channels. A shifter can set a range of hardware variables, such as voltage range, maximum current, ramp procedures and even archive data for future offline analysis, using the TRD HV control panel. The TRD DCS finite state machine (FSM) is used to perform intervening actions, in the event of unusual activity such as overcurrent and channel trips. However, such intervention is only triggered after the current maximum threshold has been breached, leading to a channel trip. Following software upgrades during LS1, the FSM during RUN2 included the option of automatically switching off a channel if an overcurrent exceeds the maximum threshold and the potential for automatic recovery under certain circumstances. Alerts for each FSM action are automatically generated, but it is the task of the shifter to provide overall management of the system. This is already challenging and following the upgrades for RUN3, will be even more so.

8.3 DCS preparations for RUN3

Following upgrades to the detector, ALICE will be challenged to read-out and inspect collisions at a rate of 50 kHz for Pb-Pb and 200 kHz for p-p and p-Pb. The ALICE O^2 project [21], combines the *on-detector* and *off-detector* (referred to as online and offline, respectively, in the high-energy physics community) computing systems. It is estimated that the new O^2 facility will be required to deal with data rates of 1 TB per second and store approximately 60 PB per year. In preparedness for the third extended period of data-taking (RUN3), the ALICE DCS needs to provide O^2 with a continuous

stream of up to 100,000 condition parameters to enable data reconstruction. In this section, we briefly describe the mechanism proposed by the DCS team [48], allowing for detector control in parallel with physics data acquisition.

Throughout data-taking in RUN1 and RUN2, ALICE DCS conditions data has been retrieved at the end of each run (typically corresponding to a period of several hours, depending on technical issues). Detector devices (channels) are read at a frequency of approximately 1 Hz and are triggered by a *change* in value (determined by the DCS operator). As such, conditions data arrival rates are not regular. At the end of a run, the entire block of conditions data is archived and then combined with the corresponding physics data, for event reconstruction. Due to the new requirement of a continuous stream of conditions parameters, DCS data-publishing rates will need to increase by a factor of 5000.

The approach proposed in [48], utilises a process image stored in memory that is sent to the O^2 computing facility for analysis. This block of conditions data is populated with the latest conditions of the experiment and updated by the DCS only when new data arrives. The Alice DAta-POint Service (ADAPOS), manages the process of collecting DCS conditions, updating the process image and publishing the image block to the O^2 computing farm. The data managed by ADAPOS provides an opportunity to apply adversarial thresholding semi-bandit algorithms that search for channels with conditions data that exceeds a threshold interval around a previous value, satisfying the requirement in specification 4.

8.4 Experimental environment

In this section, we describe the environment with which all of our experiments have been conducted and the datasets used in each experiment. The ALICE Transition Radiation Detector conditions data and the RUN3 simulated data (see Section 8.4.1 for details), are kindly provided by the ALICE Detector Control team. In Section 8.4.2, we highlight the characteristics of which channels are of interest to an adversarial thresholding semi-bandit player.

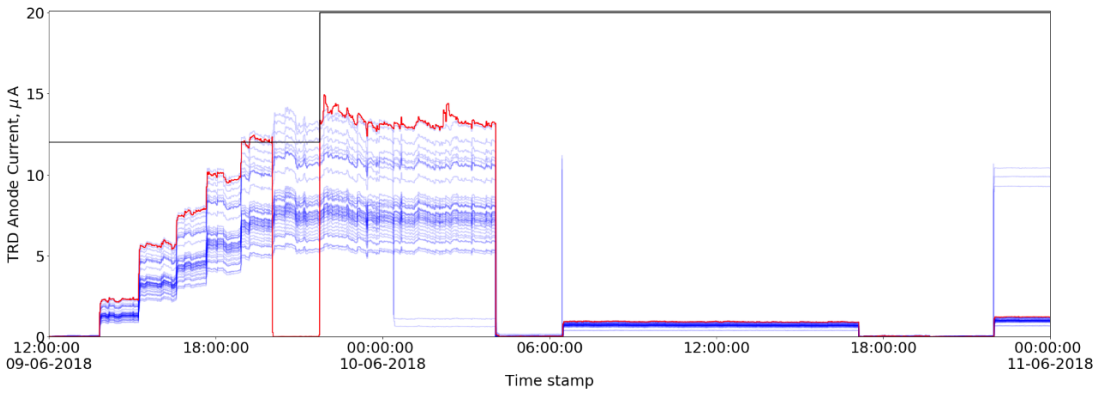


Figure 8.3: Dataset of multiple ($K = 44$) TRD Anode Current channels between 12:00:00, 9^{th} June 2018 and 00:00:00, 11^{th} June 2018. Data is binned into two second intervals, resulting in $T = 64800$ rounds. Arm 3 (red line) indicates the tripped channel, with ITRIPLimit (black line).

The interested reader is directed towards the authors' Github repository[1], which contains all the necessary code to run the experiments. All simulations were performed using a Jupyter notebook [47], version 6.0.3. All algorithms were written in Python 3.6.9, using the Numpy 1.18.1, Pandas 1.0.0 and Matplotlib 3.1.3 libraries.

8.4.1 TRD conditions and RUN3 simulation data

We now give an overview of the TRD high voltage Anode current dataset and a RUN3 simulation dataset, provided by the ALICE DCS team. We discuss the salient properties of both datasets, describing any pre-processing required to perform our analysis.

The time period under investigation for the TRD conditions dataset is between midday on 9^{th} June 2018 and mid-night on 11^{th} June 2018. A total of 44 Anode current channels were provided and the original datasets were time-stamped at approximate intervals of two seconds in length. We chose to bin the data for each channel into two second intervals, selecting the maximum value in each bin (in case of multiple values), to retain as much of the original characteristics as possible, whilst insisting on regularity within the data. This resulted in our data consisting of $K = 44$ columns of feedback with a total of $T = 64800$ rounds (see Figure 8.3). Each channel was labelled as *arm* $0 - 43$.

Overall, the current for each individual channel begins at its minimum base value ($\sim 0\,\mu A$), depending on the amount of noise associated with each channel. Over the course

[1] http://www.github.com/Craig5005/ALICE_deep_learning

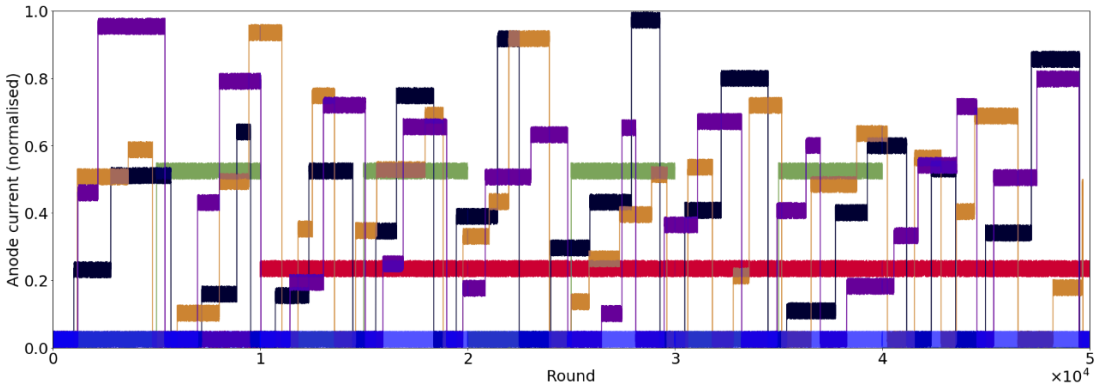


Figure 8.4: Simulated dataset of ($K = 10$) Anode Current channels (blue, purple, orange, green, red and black lines) from the ALICE Cosmic Ray detector, for $T = 50000$ rounds.

of approximately the next 10 hours, the Anode current for most channels increases in relatively regular steps. For the subsequent six hours, current for most channels remains relatively stable. However, at precisely 04:04:14 on 10^{th} June 2018, almost every channel was switched off. Approximately two hours later, all channels were switched back on and nominal current flow was observed for almost eleven hours, with the exception of arms 10, 20 and 34, presenting a short-lived spike (2-4 seconds) in current of 11.2 µA, 10.3 µA and 11.0 µA, respectively before resuming nominal behaviour. All channels were then powered off at 17:07:26 on 10^{th} June 2018 (corresponding with round $t = 52423$). All channels are then switched on again, resuming nominal current flow at 21:57:58 on 10^{th} June 2018, with arms 22 and 25 spiking to 10.7 µA and 9.1 µA respectively for two seconds before returning to nominal. However, simultaneously, arms 4, 5 and 34 present much higher but stable current flow, at 9.9 µA, 8.9 µA and 9.4 µA, respectively. A pre-defined threshold, referred to as the ITRIPLimit is given, which can be manually set by a detector operator.

Anode current is simulated under RUN3 operating conditions (see Figure 8.4), for a period of $T = 50000$ rounds (27 hours, 46 minutes and 40 seconds). A sample of eight channels was generated by the ALICE DCS team, with three channels exhibiting unusual deviations in current, followed by short periods of bound stability (current remains within a relatively small interval throughout that period). The remaining five channels present bound stability for the entire time horizon of $T = 50000$ rounds. A further two channels of data are simulated, one demonstrating periodic deviations and another with a single deviation.

8.4.2 Channels of interest

In this section, we highlight several channels of Anode current data that are of interest to the adversarial thresholding semi-bandit player. For the TRD dataset:

- Channels (arms) 0, 3, 4, 27 and 28 are the only arms that exceed the initial ITRIPLimit of 12 µA at least once between 12:00:00, 9^{th} June 2018 and 00:00:00, 11^{th} June 2018.
- Channel 3 is the only channel to trip, between rounds $t = 14465$ and $t = 17533$.
- Channel 3 exceeds ITRIPLimit on 1471 rounds (49 m 2 s), out the 1989 rounds (1 hr 6 m 18 s) between exceeding ITRIPLimit and tripping.
- Arms 0 and 28 crosses the ITRIPLimit boundary on 14 and 48 occasions, respectively, but all after arm 3 is tripped in round $t = 14465$.
- Arms 3, 4 and 27 cross the 12 µA mark, for 10, 3 and 5 rounds respectively, before arm 3 trips.
- Arms 3, 4 and 27 deviate in current by at least 3 µA on 8, 7 and 6 occasions, compared with a median of 4 deviations and all arms deviate at least twice.

As previously discussed and illustrated in Figure 8.3, the stream of Anode current for all $K = 44$ channels presents a similar sequence of distinct steps, until all channels are switched off in round $t = 28928$. A sequence of four *jumps* can be seen, at approximately rounds (t = 3270-3290), (t = 5810-5858) (actually consisting of two jumps between t = 5810-5814 and t = 5857-5858), (t = 8260-8275) and (t = 10193-10245), before channel 3 trips.

The RUN3 simulated dataset consists of $K = 10$ arbitrary channels. Figure 8.4 illustrates channel 0 (green line) with relatively stable Anode current that does not deviate by more than 3 µA, interspersed by periodic jumps every 5000 rounds, of more than 3 µA. Channel 1 (red line) shows the same relatively stable current behaviour, but with an isolated jump of more than 3 µA, in round $t = 10000$. Channels 2, 3 and 4 (purple, orange and black lines, respectively) show irregular jumps in current and the remaining channels 5-9 (blue lines) all show relatively stable current, not deviating by more than 3 µA.

8.5 ATSBP for ALICE

The two approaches used to quantify application of the adversarial thresholding semi-bandit problem, to monitoring detector conditions on heavy-ion physics experiments are discussed in terms of the ATSBP objective. Recall that the goal of playing the adversarial thresholding semi-bandit game is to pull only those arms satisfying some threshold criteria each round.

Fixed threshold approach: The current approach for the ALICE DCS is with *overcurrent detection*. This relates to the monitoring of channels when their feedback exceeds a fixed threshold. The theoretical analysis performed in previous chapters assumes a fixed threshold approach.

The limitation of a fixed threshold policy is that the behaviour from historic feedback within threshold is not recognised prior to a trip event, since an ATSBP player is learning to disregard channels returning feedback within threshold in preference for other channels, regardless of the instability in current from that channel.

Threshold interval approach: The ALICE DCS will need to be concerned with *current deviation detection* upon commencement of RUN3 (see Section 8.3 for details). This approach defines a fixed interval, centred on the most recent observation from each channel. If a new observation returns feedback outside of this interval then the interval becomes centre about the new observation. The goal of the adversarial thresholding semi-bandit player here is to pull only those arms with feedback deviating by more than the deviation threshold, from their more recent observation. The goal of the ALICE DCS operator is to know about a deviation in current exceeding a deviation threshold as soon as and as efficiently as possible.

In Chapters 4, 5, 6 and 7, upper bounds on theoretical performance were proved in terms of feedback $x_{t,i} \in [0,1]$, from arms $i = 1,\ldots,K$ exceeding a fixed threshold, $0 < \theta < 1$, in round $t = 1,\ldots,T$. A threshold interval is defined as,

$$\left[\max\left\{0, x_{t',i} - \frac{\theta}{2}\right\}, \min\left\{x_{t',i} + \frac{\theta}{2}, 1\right\}\right], \tag{8.1}$$

where t' denotes the most recent round in which arm i was observed, such that $1 \leq t' < t$, $i \in \Lambda_{t'} \cup \mathcal{T}_{t'-1}$ and $i \notin \Lambda_{t''} \cup \mathcal{T}_{t''-1}$ $\forall\, t'' : t' < t'' < t$ and $t = 2,\ldots,T$. The threshold

interval can be manually set and is $[0, \theta/2]$ for $t' = 1$. The adversarial thresholding semi-bandit player wants to observe arms where $x_{t,i}$ does not belong to the interval given in (8.1). This is equivalent to, $|x_{t,i} - x_{t',i}| > \theta$, upon which t' becomes t. Since $|x_{t,i} - x_{t',i}| \in [0, 1]$ and,

$$1 - \theta \leq \left| \left[0, \max \left\{ 0, x_{t',i} - \frac{\theta}{2} \right\} \right) \cup \left(\min \left\{ x_{t',i} + \frac{\theta}{2}, 1 \right\}, 1 \right] \right| \leq 1 - \frac{\theta}{2},$$

then the theoretical upper bounds proved for the fixed threshold regime also hold for the threshold interval regime.

8.6 Empirical analysis

We now present the findings from an empirical analysis on each of the algorithms developed in Chapters 4, 5, 6 and 7, using the datasets described in Section 8.4.1. To measure the empirical performance of adversarial thresholding semi-bandit algorithms and satisfy Research specification 2, our analysis consists of three areas, where all results are averaged over ten repeated experiments to ensure validity.

ATSBP performance: Regret is measured as the cumulative difference between the feedback from pulling arms according to the decisions made by an adversarial thresholding semi-bandit algorithm and the cumulative feedback from pulling optimally. For the TRD conditions dataset and under a fixed threshold regime, the set of optimal arms is,

$$I_t^* \subset \{0, 3, 4, 27, 28\} \cup \emptyset,$$

depending on which elements of I_t^* have feedback exceeding threshold in round t. The empirical performance of adversarial thresholding semi-bandit algorithm for fixed thresholds is compared by computing the cumulative gain, from Definition 3 (see Section 4.1).

The optimality of an arm in a given round is dependent on the historic feedback from that arm, in terms of a threshold interval approach. Hence, every arm is potentially optimal in any given round. By Definition 1, the ATSBP player is rewarded with feedback, exclusively in the round where the deviation from the most recent observation exceeds

threshold. Optimally, this would coincide with the deviation between successive rounds exceeding threshold. However, detecting the deviation in the following round is almost as valuable and the player should be rewarded accordingly. The player is rewarded optimally for pulling an arm when a deviation takes place and the longer it takes for the player to detect the deviation, the smaller the reward received (until the next deviation). For the threshold interval approach, ATSBP performance is compared in terms of cumulative reward, since reward maximisation is equivalent to regret minimisation. The reward received by the player is a function of the most recently observed feedback, $x_{t',i}$ and defined by,

$$r_{t,i} = \begin{cases} x_{t',i} \exp\left(-\frac{\eta}{K}(t-t') \right) & \text{if } i \in \Lambda_t \cup \mathcal{T}_{t-1} \\ 0 & \text{otherwise,} \end{cases} \tag{8.2}$$

where $x_{t',i}$ is set to $x_{t,i}$ and t' is set to t, if $|x_{t,i} - x_{t',i}| > \theta$ (see Section 8.5). Initially, $t' = 0$ and $x_{0,i} = 0$ for $i = 1, \ldots, K$ and $t = 1, \ldots, T$. The cumulative reward is computed by summing for every round, $r_{t,i}$ summed over all arms $i \in \Lambda_t \cup \mathcal{T}_{t-1}$ and scaled by $1/\Delta_t$ each round, where q_t is set to the number of arms deviating most often ($|\{3, 4, 27\}| = 3$ for the TRD dataset and $|\{0, 1, 2, 3, 4\}| = 5$ for the simulated dataset). Hence, the ATSBP performance under threshold interval is measured by $\sum_{t\in[T]} \Delta_t^{-1} \sum_{i\in I_t} r_{t,i}$, where $I_t = \Lambda_t \cup \mathcal{T}_{t-1}$ and $\mathcal{T}_0 = \emptyset$.

Cumulative pull frequency: Relative changes in the frequency with which an algorithm decides to pull an arm gives an indication that the algorithm believes the feedback from that arm exceeds threshold. This information is critical to an ALICE DCS operator and can be used to instigate preventative measures (manually or automatically).

Efficiency ratios: The efficiency of an adversarial thresholding semi-bandit algorithm is measured in terms of false positive rate α (observing feedback within threshold, out of all feedback within threshold), false negative rate β (not observing feedback exceeding threshold, out of all feedback exceeding threshold) and false discovery rate FDR (observing feedback within threshold, out of all observations). Our analysis includes α and β, however, an ALICE DCS operator would only have sufficient information to monitor FDR.

8.6.1 Fixed threshold results

In this section, we present the results from running adversarial thresholding semi-bandit algorithms on the ALICE Transition Radiation Detector conditions dataset (see Section 8.4.1), with a fixed threshold set to the ITRIPLimit (12 µA). Since an ITRIPLimit was not provided with the RUN3 simulated dataset, a fixed threshold analysis has been omitted.

Label efficient adversarial thresholding semi-bandits: An analysis of label efficient ATSBP is performed by comparing **T-Exp3.M** and **LET-Exp3.M** with observation probability $\epsilon \in \{0.9, 0.8, 0.7\}$, where **T-Exp3.M** is **LET-Exp3.M** with $\epsilon = 1$. Label efficiency can be viewed in terms of the adversarial thresholding semi-bandit player deciding not to pull certain arms and still achieve relatively well or the environment not revealing feedback occasionally and the player still achieving relative performance.

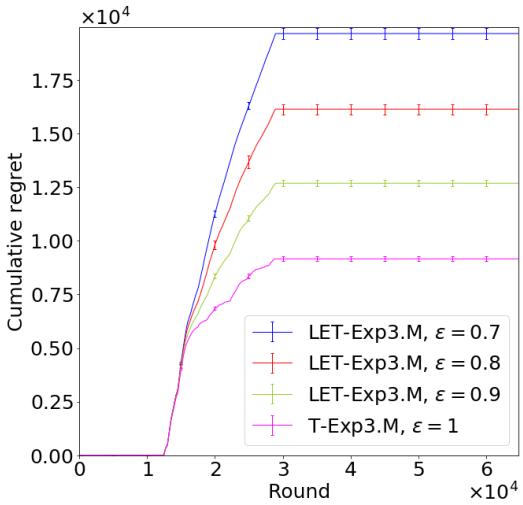


Figure 8.5: Average cumulative regret of label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, in the fixed threshold setting, on the ALICE TRD conditions dataset.

Comparison of average cumulative regret for label efficient adversarial thresholding semi-bandit algorithms indicates that as the player has access to less feedback, regret increases (see Figure 8.5). Periods of constant regret coincide with rounds where no arm exceeds threshold (recall *free-play* from Section 3.1.2). The small variation in performance for each setting of the maximum observation probability suggests the robustness of **LET-Exp3.M** and **T-Exp3.M**.

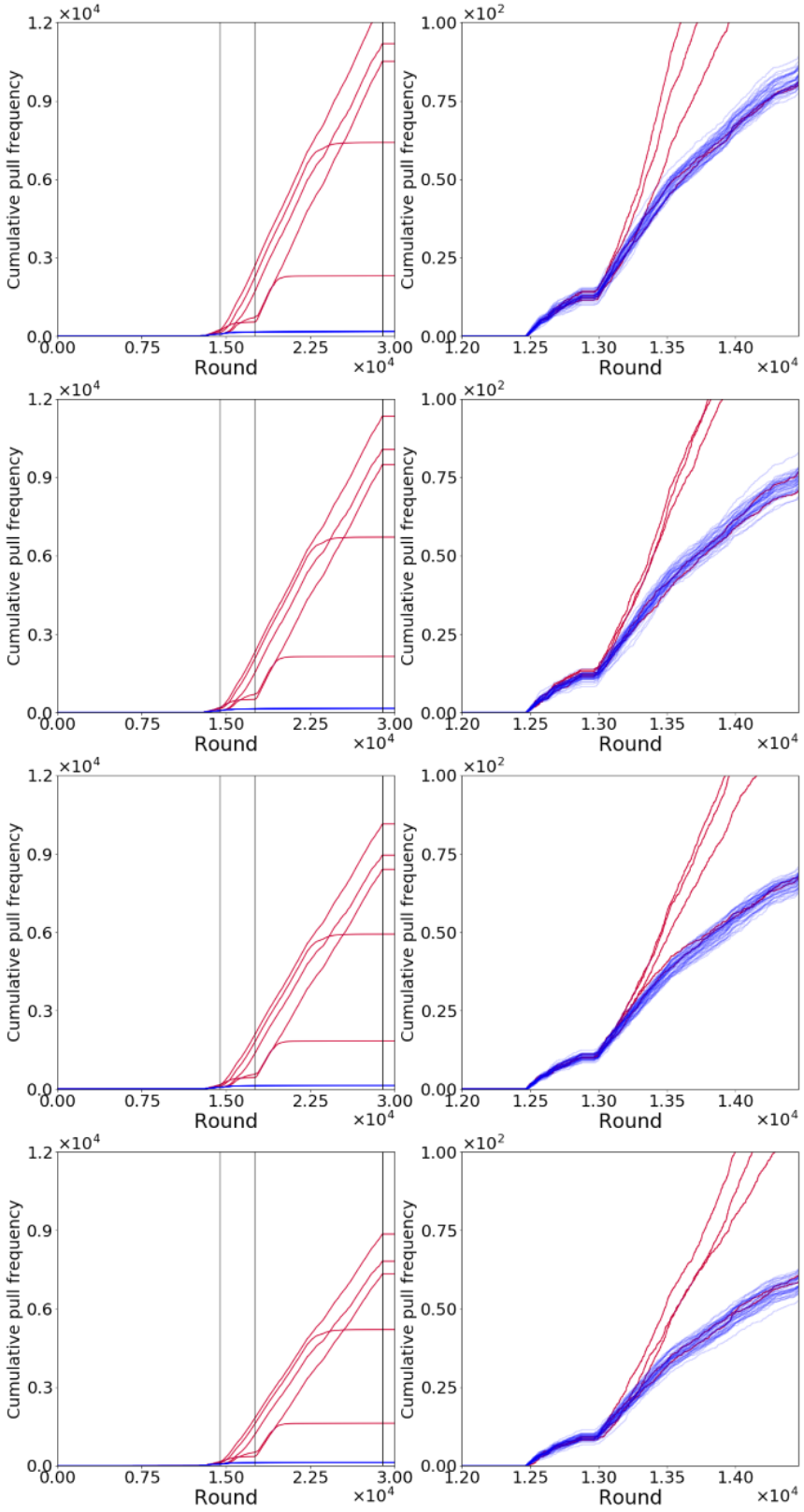


Figure 8.6: Average cumulative pull frequency for **LET-Exp3.M** algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top to bottom). (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 12000, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines.

The relative difference in regret of **LET-Exp3.M**, for $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, shows that the performance of label efficient adversarial thresholding semi-bandit algorithms is dependent on the amount of information received. For monitoring Anode current channels on ALICE, robustness, reliability and empirical performance are required for Research specification 2.

The average cumulative pull frequency of label efficient adversarial thresholding semi-bandit algorithms (given in Figure 8.6), indicates a significant difference between the number of pulls of optimal arms and sub-optimal arms, over the time horizon. The arms exceeding threshold prior to arm 3 tripping (arms 3, 4 and 27) can easily be identified as behaving differently to other arms, almost an hour (49 minutes) before arm 3 trips (round $t = 14465$). The cumulative pull frequencies of the optimal arms present a constant positive gradient, appearing between round $t = 14465$ (trip) and round $t = 28928$ (all arms switched off). This indicates that **LET-Exp3.M** recognises an extended period of time where arms 0, 3, 4, 27 and 28 have TRD Anode current exceeding the ITRIPLimit of $\theta = 12\,\mu\text{A}$. The minimal difference in cumulative pull frequencies for various values of the observation probability ϵ, is evidence that label efficient adversarial thresholding semi-bandit algorithms are robust to randomly distributed drop-out of feedback.

Since label efficient adversarial thresholding semi-bandit algorithms rely on Assumption 2, the player has knowledge of how many arms to pull each round, but not which arms. This explains why the cumulative pull frequency of every arm increases as soon as the first arm to exceed threshold does so (arm 4, in round $t = 12476$). It takes the label efficient player relatively the same amount of time to be able to recognise arms 3, 4 and 27 need observing more often (approximately 1000 rounds), regardless of the proportion of feedback they do not get to observe, since the proportion is uniformly distributed over the time horizon. Arms 0 and 28 are yet to exceed threshold, prior to round $t = 14465$ and remain indistinguishable from sub-optimal arms.

The efficiency of **LET-Exp3.M** is determined by the number of arms pulled, how many arms return feedback within threshold and how many have feedback exceeding threshold over the course of T rounds (see Table 8.1).

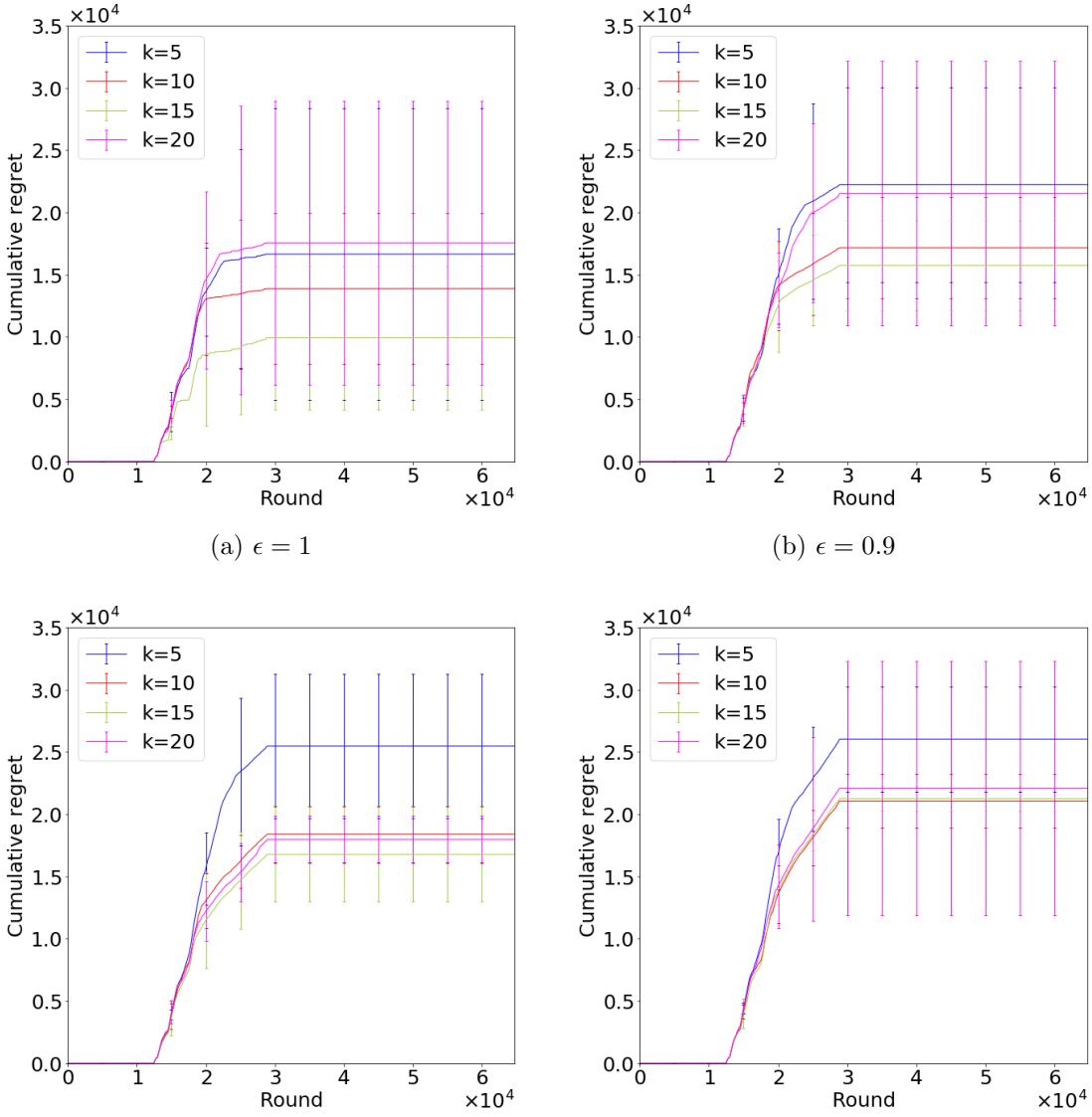
Table 8.1: Comparison of average false positive (α), negative (β) and discovery (FDR) rates of label efficient adversarial thresholding semi-bandit algorithms, in the fixed threshold setting, on the ALICE TRD conditions dataset.

Algorithm		α	β	FDR
T-Exp3.M	(LET-Exp3.M, $\epsilon = 1$)	0.003	0.211	0.211
LET-Exp3.M	$\epsilon = 0.9$	0.003	0.294	0.215
	$\epsilon = 0.8$	0.003	0.312	0.215
	$\epsilon = 0.7$	0.003	0.451	0.215

As ϵ decreases, the player uses (or has access to) a smaller proportion of the feedback they wish to observe, hence, fewer arms are pulled. For the false positive rate α, to remain constant (to three decimal places) throughout $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, the ratio of arms pulled with feedback within threshold to all arms with feedback within threshold remains constant. Since the total number of rounds each arm has feedback within threshold does not change, then those rounds where feedback is not returned from certain arms, mostly consist of arms with feedback exceeding threshold. This indicates that the majority of arms pulled by label efficient thresholding semi-bandit algorithms, have feedback exceeding threshold. The same reasoning explains the increase in false negative rate β, as ϵ decreases. Pulling fewer arms implies more arms are not pulled with feedback exceeding threshold, particularly when most of the arms pulled return feedback exceeding threshold. Furthermore, pulling fewer arms should cause the false discovery rate (FDR) to increase. The fact that FDR only increases slightly, indicates that since fewer arms are pulled in total, the proportion of arms pulled with feedback within threshold must increase. Hence, those arms that would have been pulled, would have returned feedback exceeding threshold.

Dynamic label efficient adversarial thresholding semi-bandits: Since it is unrealistic in any practical application to rely on Assumption 2, the dynamic label efficient adversarial thresholding semi-bandit (introduced in Chapter 6) makes the assumption that arms observed with feedback exceeding threshold in the current round are likely to do so in the next round (see Definition 4 in Section 6.2). The policy for selecting subsets of arms each round remains the same importance-weighted distribution, as for label efficient adversarial thresholding semi-bandits. However, the probability of observing the feedback from selected arms is dynamically updated, depending on the history of observed feedback for each arm. A range of values for the maximum obser-

vation probability are analysed ($\epsilon \in \{1, 0.9, 0.8, 0.7\}$), which it resets to every time the arm is detected exceeding threshold. Since the optimal number of arms to pull each round is unknown, a range of subset sizes are also investigated, where $k \in \{5, 10, 15, 20\}$.



(a) $\epsilon = 1$

(b) $\epsilon = 0.9$

(c) $\epsilon = 0.8$

(d) $\epsilon = 0.7$

Figure 8.7: Average cumulative regret of dynamic label efficient adversarial thresholding semi-bandit algorithms, with subsets of size $k \in \{5, 10, 15, 20\}$, selected each round and maximum observation probability, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right). Experiments on the ALICE TRD conditions dataset, in the fixed threshold setting, were repeated ten times.

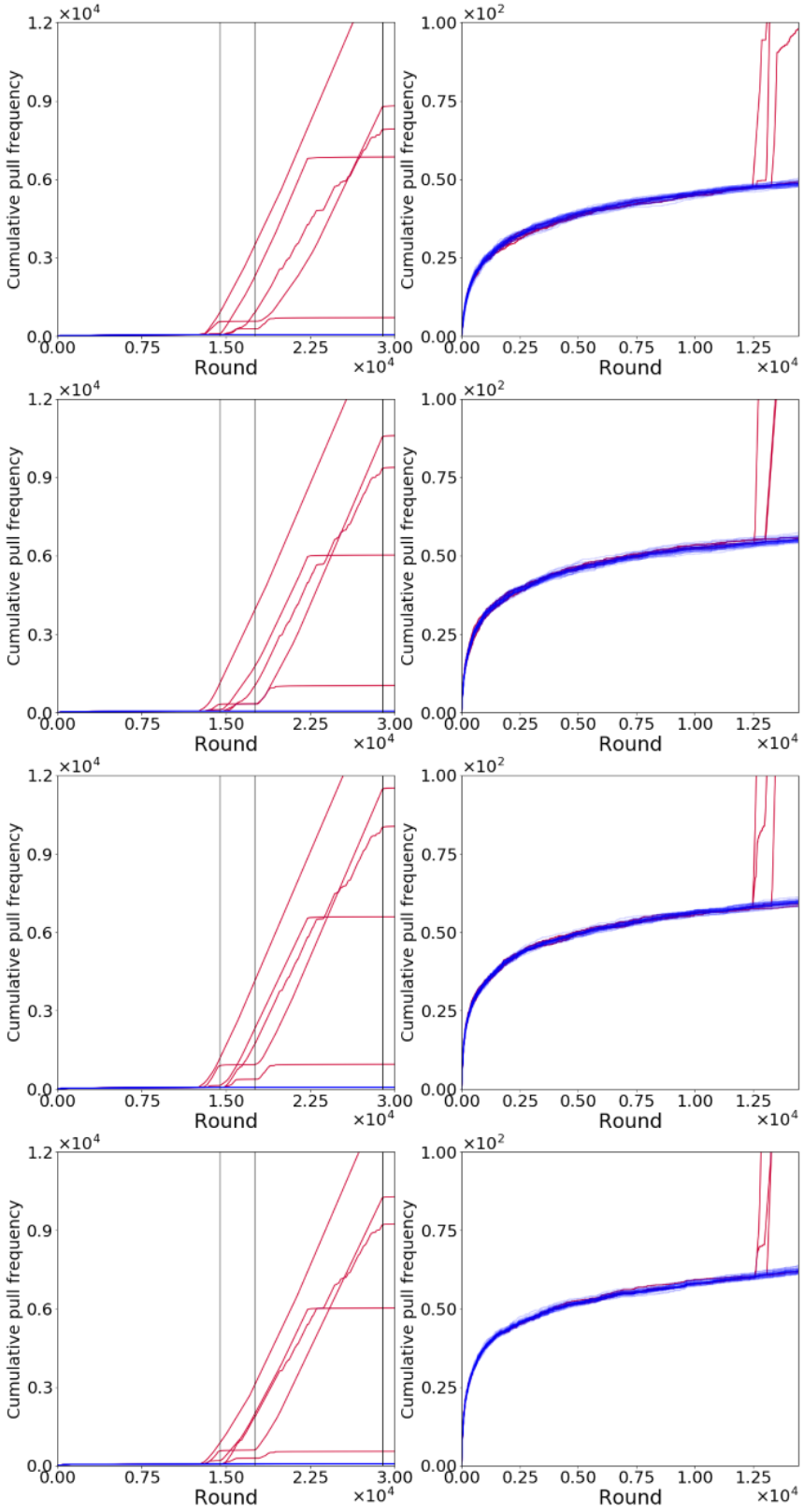


Figure 8.8: Average cumulative pull frequency for **dLET-Exp3.M** algorithms with subsets of $k \in \{5, 10, 15, 20\}$ arms selected each round (top to bottom) and $\epsilon = 1$. (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 1, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines.

Since the optimal number of arms to pull each round is no more than $q_t \leq 5$ (where q_t denotes the actual number of arms exceeding threshold in round t), Figure 8.7 shows that to optimise the performance of **dLET-Exp3.M**, sufficient numbers of arms need to be selected each round, but not excessively so, before the decision to observe is taken.

The average cumulative regret of dynamic label efficient adversarial thresholding semi-bandit algorithms increases as the maximum observation probability ϵ, decreases. This is expected behaviour, since the dynamic label efficient adversarial thresholding semi-bandit player will regret more of their decisions if they have less information to learn from. However, the relative similarities in the evolution of cumulative regret across the range of ϵ values and across the range of subset sizes k, indicates that **dLET-Exp3.M** is robust to feedback loss. Our analysis suggests that there is an opportunity to optimise the subset size, k. In Figure 8.7 (b), (c) and (d), we see that $k = 5$ incurs the most regret. For maximum observations probabilities of $\epsilon \geq 0.9$, $k \in \{5, 20\}$ incur the most regret. For $\epsilon = 1, 0.9$ and 0.7, $k = 15$ incurs the least regret.

Since the dynamic label efficient adversarial thresholding semi-bandit player has no knowledge of the optimal number of arms to pull each round, the average cumulative pull frequencies of all arms increase from the beginning of the game (see Figure 8.8). However, even without this knowledge, the total number of pulls of sub-optimal arms (blue) by the time of the trip event (round $t = 14465$), is comparable with the cumulative pull frequency of sub-optimal arms for **LET-Exp3.M**. All optimal arms (0, 3, 4, 27 and 28) are clearly distinguishable from sub-optimal arms by round $t = 28928$, by the number of times they have been observed, regardless of the proportion of feedback from pulled arms they receive and regardless of the number of arms selected for observation each round. Setting the maximum observation probability ($\epsilon = 1$) and the number of selected arms each round ($k = 15$), **dLET-Exp3.M** can identify all optimal arms by observing them more than sub-optimal arms, more frequently and no earlier than round $t = 14752$ (9 minutes, 34 seconds after the trip event). By round $t = 14465$, **dLET-Exp3.M** ($\epsilon = 1$, $k = 15$) has cumulatively pulled arm 27 (the least pulled optimal arm at that time) 2.3 times more often than arm 20 (the most pulled sub-optimal arm).

The average false positive rate (α), false negative rate (β) and false discovery rate (FDR), for **dLET-Exp3.M** with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, are averaged after repeated

Table 8.2: Comparison of average false positive (α), negative (β) and discovery (FDR) rates of **dLET-Exp3.M** with maximum observation probabilities $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ and subsets selection sizes $k \in \{5, 10, 15, 20\}$, in the fixed threshold setting, on the ALICE TRD conditions dataset.

ϵ	k	α	β	FDR
1	5	0.001	0.248	0.074
	10	0.001	0.191	0.082
	15	0.001	0.190	0.084
	20	0.001	0.231	0.092
0.9	5	0.001	0.320	0.079
	10	0.001	0.273	0.090
	15	0.001	0.274	0.094
	20	0.001	0.340	0.104
0.8	5	0.001	0.390	0.087
	10	0.001	0.320	0.095
	15	0.001	0.337	0.099
	20	0.001	0.340	0.103
0.7	5	0.001	0.467	0.099
	10	0.001	0.438	0.111
	15	0.001	0.414	0.111
	20	0.001	0.487	0.132

experimentation on the ALICE TRD conditions dataset (see Table 8.2). The average false positive rate remains unchanged for all values of (ϵ, k). The metric α, denotes the average ratio of pulled arms with feedback within threshold to arms with feedback within threshold throughout the time horizon of $T = 64800$ rounds. For α to remain constant as the maximum observation probability decreases, the majority of pulled arms must return feedback exceeding threshold. The metric β, denotes the average ratio of arms not pulled with feedback exceeding threshold to arms with feedback exceeding threshold over T rounds. For each value of ϵ, an optimal subset size is suggested, minimises β, as the subset size increases from $k = 5$ to $k = 20$. Larger values of β imply that the average number of non-arm pulls with feedback exceeding threshold dominates over the average number of arm pulls with feedback exceeding threshold. Selecting just enough arms to observe those exceeding threshold relies on correctly selected all optimal arms and correctly deciding to observe all selected. If too many ($K \geq k_t > q_t$) arms are selected, then the player is certain to pull arms with feedback within threshold. The average number of non-arm-pulls with feedback exceeding threshold decreases because more arms are selected and pulled. In general, as fewer arms are pulled (ϵ decreases), the average ratio of non-arm-pulls with feedback

exceeding threshold to arms with feedback exceeding threshold, increases, since more arms are not pulled. In the context of adversarial thresholding semi-bandits, FDR represents the average ratio of arms pulled with feedback within threshold to total arms pulled. FDR increasing as the maximum observation probability decreases is implied by fewer arms revealing their feedback. FDR increasing as the number of arms selected for observation increases implies that the number of arms pulled with feedback within threshold also increases. While pulling few arms may increase the average rate of false negatives, small FDR implies that of the arms pulled, most do exceed threshold. Table 8.2 suggests that pulling too many arms leads to pulling more arms with feedback within threshold and not pulling arms with feedback exceeding threshold than pulling arms with feedback exceeding threshold.

AliceBandit: A motivation for developing the **AliceBandit** algorithm (see Chapter 7) was the adversarial thresholding semi-bandit player not knowing how many arms to pull each round and being unable to assume that this number remains fixed. The player selects subsets of arms by a randomised round robin procedure and then decides whether to observe selected arms. An observation probability $p_{t,i}$, is updated depending on the feedback received. The player utilises a combination of a threshold list (Definition 4, Section 6.2), the randomised round robin subset selection with observation dependent on $p_{t,i}$ and a parameter t_{max}, constraining the maximum number of rounds between successive observations for any arm, regardless of previous selections.

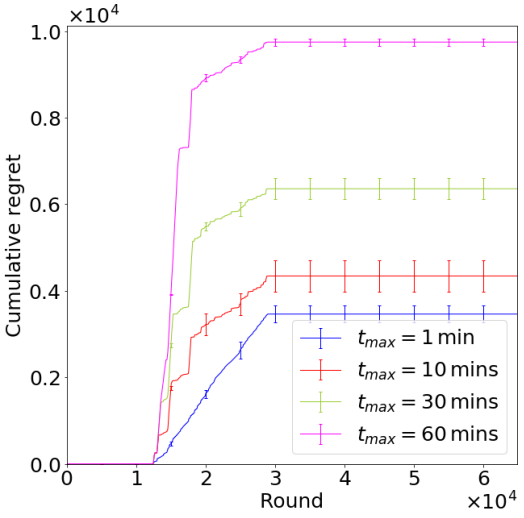


Figure 8.9: Average cumulative regret of **AliceBandit**, with $t_{max} = \{1, 10, 30, 60\}$ minutes. Experiments on the ALICE TRD conditions dataset, in the fixed threshold setting were repeated ten times.

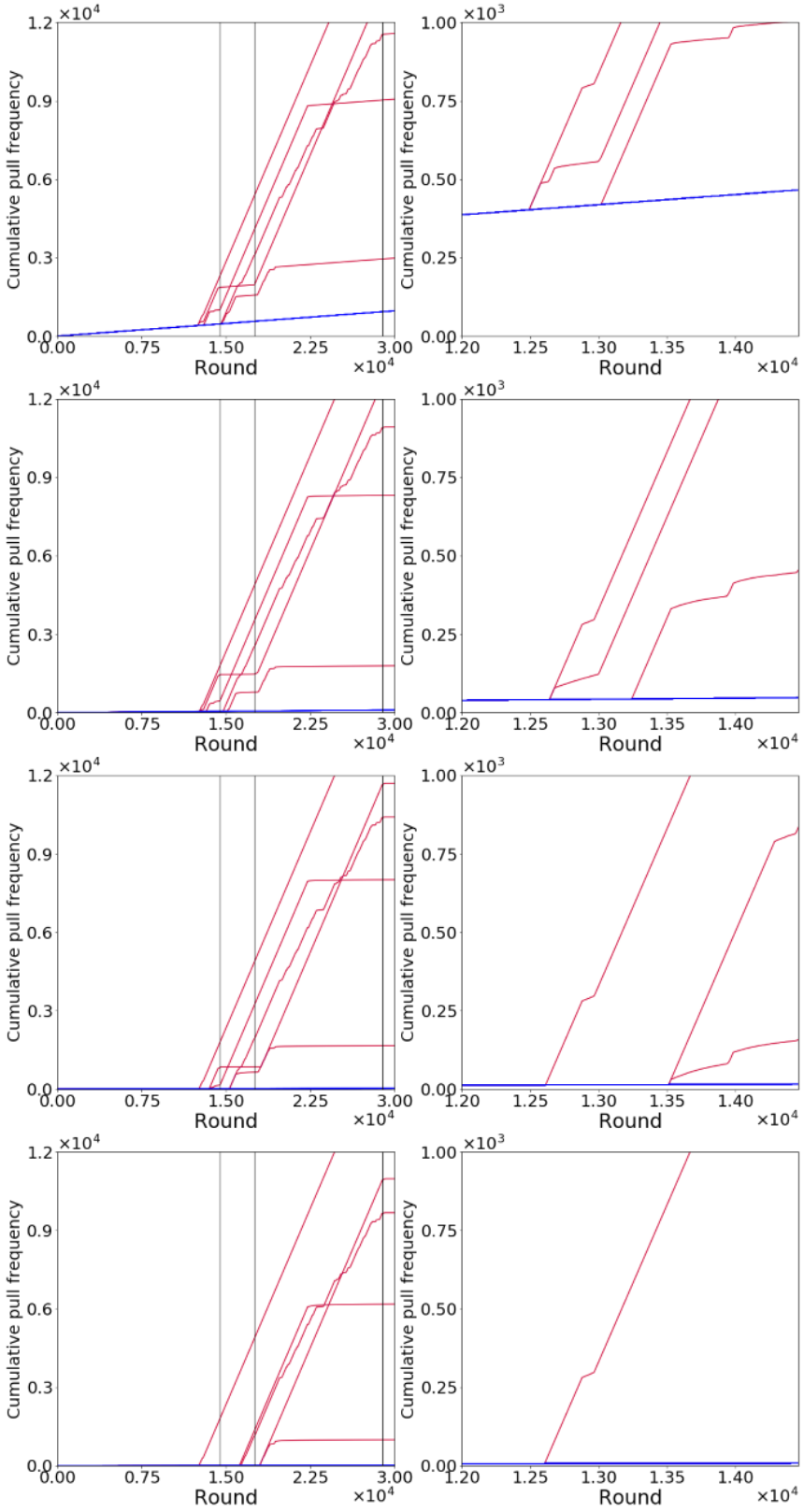


Figure 8.10: Average cumulative pull frequency for **AliceBandit** with $t_{max} = \{1, 10, 30, 60\}$ minutes (top to bottom). (left) ALICE TRD conditions dataset ($t = 1, \ldots, 30000$) with trip, recovery and switch off indicated (black lines). (right) Pulling behaviour prior to trip event ($t = 12000, \ldots, 14465$). Arms $\{0, 3, 4, 27, 28\}$ shown by red lines, remaining arms given by blue lines.

The average cumulative regret of **AliceBandit** for a range of values of the maximum number of rounds between consecutive observations are analysed, where $t_{max} = \{1, 10, 30, 60\}$ minutes (corresponding with $t_{max} = \{30, 300, 900, 1800\}$ rounds). Since larger values of t_{max} allow for arms to be forgotten for longer, the player incurs more regret. Smaller value of t_{max} ensure that arms with feedback exceeding threshold have a better chance of being observed (see Figure 8.9). The player does not want to observe arms that have previously shown little sign of exceeding threshold, too often. However, the player does not want to forget about such arms altogether, because the environment is adversarial. Due to the free-play concept (Section 3.1.2), **AliceBandit** only incurs regret during the period where arms have feedback exceeding threshold. For $t_{max} = 1$ minute, the cumulative regret for rounds $t = 12469, \ldots, 28928$ appears relatively linear. This is due to the regular intervals between pulling arms with feedback consistently within threshold and only incurring regret from pulling these arms. As we increase the size of t_{max}, this linear contribution diminishes and the regret incurred from selecting and deciding to observe arms with feedback within threshold dominates. The rate of increase in average cumulative regret is proportional to the rate of increase in the maximum number of rounds between consecutive observations. **AliceBandit** takes longer to learn which arms have feedback exceeding threshold. Arms pulled and observed within threshold may not be pulled again for another t_{max} rounds, giving the player many opportunities for regret as t_{max} increases.

The average cumulative pull frequencies for **AliceBandit** are illustrated for $t_{max} = \{1, 10, 30, 60\}$ minutes (see Figure 8.10). Whilst the constant rate of increase in cumulative pull frequencies of sub-optimal arms (blue lines) is governed by the size of t_{max}, the periods of constant increase in cumulative pulls of optimal arms is due to those arms having extended periods with feedback exceeding threshold and Assumption 3, where an arm is pulled with certainty in round $t + 1$, if it is observed with feedback exceeding threshold in round t, by the threshold list (see Section 6.2). By round $t = 30000$, **AliceBandit** had identified all optimal arms, for every value of t_{max}. By the time of the trip event ($t = 14465$), **AliceBandit** had identified all arms that had exceeded threshold by then (arms 3, 4 and 27), for $t_{max} = \{1, 10, 30\}$ minutes. By round $t = 14465$, the least pulled of the arms exceeding threshold (arm 27) had been pulled 2.2, 9.5 and 30.2 times that of the most pulled sub-optimal arm (arm 43), for $t_{max} = \{1, 10, 30\}$ minutes, respectively. By the time of the trip event, **AliceBandit**

has only pulled one arm (arm 4) significantly more than the remaining arms, although that arm was pulled 224.3 times more than the next most pulled arm (arm 43). The latest all three arms exceeding threshold before the trip event can be identified as being observed by **AliceBandit**, more than the baseline (once every t_{max} rounds) is $t = \{13021, 13243, 13514, 16208\}$, for $t_{max} = \{1, 10, 30, 60\}$ minutes, respectively.

Table 8.3: Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 10, 30, 60\}$ minutes, in the fixed threshold setting, on the ALICE TRD conditions dataset.

t_{max}	α	β	FDR
1 min	0.032	0.011	0.642
10 mins	0.004	0.039	0.169
30 mins	0.001	0.074	0.074
60 mins	0.001	0.148	0.048

The average efficiency ratios for **AliceBandit** are presented, over the range of maximum rounds between successive pulls (see Table 8.3). The average false positive rate decreases as t_{max} increases. This implies that as fewer arms are pulled, more arms are not pulled when their feedback would have been within threshold. This also implies that those arms that would have been pulled due to a smaller t_{max}, would have yielded feedback within threshold. The fact that the average false negative rate increases with t_{max}, suggests that the number of arms not pulled with feedback exceeding threshold, dominates the number of arms pulled with feedback exceeding threshold, as t_{max} increases. The average false discovery rate decreases as t_{max} increases, suggesting that the number of arms pulled with feedback exceeding threshold, dominates those with feedback within threshold, as the maximum number of rounds between consecutive pulls increases. The relationship between α and t_{max}, and β and t_{max}, coincides with larger t_{max} implying that fewer arms are pulled in general. However, the relationship between FDR and t_{max}, infers that the number of arms pulled with feedback within threshold must be decreasing more than the number of arms pulled with feedback exceeding threshold. Unfortunately, increasing t_{max} too much results in too many arms not pulled with feedback exceeding threshold, which is contrary to the objective of ATSBP.

8.6.2 Fixed threshold summary

In this section, we summarise the empirical performance of the algorithms developed in Chapters 4, 5, 6 and 7, for the problem of detecting overcurrent in power channels from the Transition Radiation Detector, on ALICE.

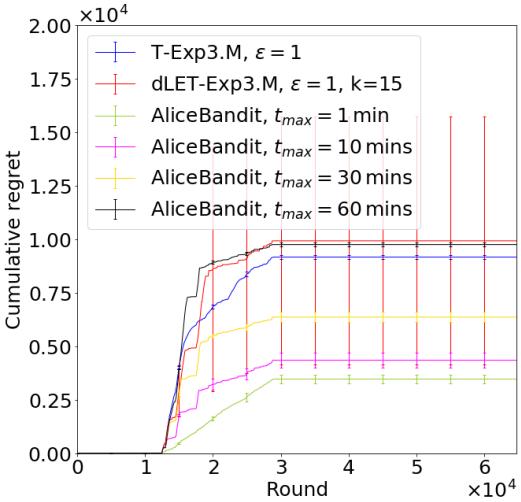


Figure 8.11: Average cumulative regret of selected adversarial thresholding semi-bandit algorithms, on the ALICE TRD conditions dataset, in the fixed threshold setting, where experiments were repeated ten times.

The average cumulative regret of adversarial thresholding semi-bandit algorithms, with specific configuration parameters yielding better performance are compared (see Figure 8.11). The **AliceBandit** algorithm with $t_{max} = 1$ minute, achieves approximately four times better performance than exponentially-weighted adversarial thresholding semi-bandit algorithms, even if the number of arms with feedback exceeding threshold each round, is known. The maximum number of consecutive rounds between pulls can be extended to at least $t_{max} = 60$ minutes, before the performance of **AliceBandit** is comparable with the best performing configuration of dynamic label efficient adversarial thresholding semi-bandit.

The number of sub-optimal arm pulls prior to the trip event for **LET-Exp3.M** with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (including **T-Exp3.M**), increases from approximately 50, to approximately 80, as ϵ decreases. While **dLET-Exp3.M** increases from approximately 50, to approximately 60, for the same decrease in maximum observation probability. However, the number of sub-optimal arm pulls prior to trip, for **AliceBandit** is 466, 48, 16 and 8, with $t_{max} = \{1, 10, 30, 60\}$ minutes, respectively. Hence, **AliceBandit**

with $t_{max} = 10$ minutes, achieves performance approximately 60% better than the other best performing adversarial thresholding semi-bandit and with a comparable number of sub-optimal arm pulls prior to trip.

The empirical analysis of **AliceBandit** has been demonstrated to achieve better performance than other adversarial thresholding semi-bandit algorithms, in the fixed threshold setting, with the ALICE TRD conditions dataset. Thus **AliceBandit** satisfies Research specification 2, in terms of empirical performance. Whilst exponentially-weighted adversarial thresholding semi-bandits partially satisfy Research specification 2, when also factoring efficiency ratios. However, achievement of specification 4 has yet to be demonstrated.

8.6.3 Threshold interval results

In this section, we present the results from running adversarial thresholding semi-bandit algorithms on the ALICE Transition Radiation Detector conditions dataset and RUN3 simulated dataset (see Section 8.4.1), with a threshold interval set to deviations of at least 3 µA. Again, all results are the mean from ten repeated experiments.

Since every arm deviates by more than 3 µA, over the course of $T = 64800$ rounds, in the ALICE TRD conditions dataset, all K arms could in this context be considered optimal. Hence, in the threshold interval setting, we utilise the equivalence between regret minimisation and reward maximisation. Adversarial thresholding semi-bandits are analysed empirically, in terms of average cumulative reward, G (see Definition 3 in Section 4.1), average cumulative pull frequencies and average efficiency ratios. The cumulative pull frequencies of arms 3, 4 and 27 are highlighted due to their greater frequency of deviation, particularly prior to the trip event. The reward received by the player is a function of the deviation between the current and most recent observation (recall (8.2)).

Label efficient adversarial thresolding semi-bandits: Label efficiency in the threshold interval setting has been analysed over maximum observation probabilities, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, as in the fixed threshold setting.

The average cumulative reward of label efficient adversarial thresholding semi-bandit algorithms increases as the maximum observation probability decreases (see Figure

8.12). This implies that observing fewer arms improves performance and that the adversarial thresholding semi-bandit player is struggling to detect deviations in Anode current, larger than 3 µA.

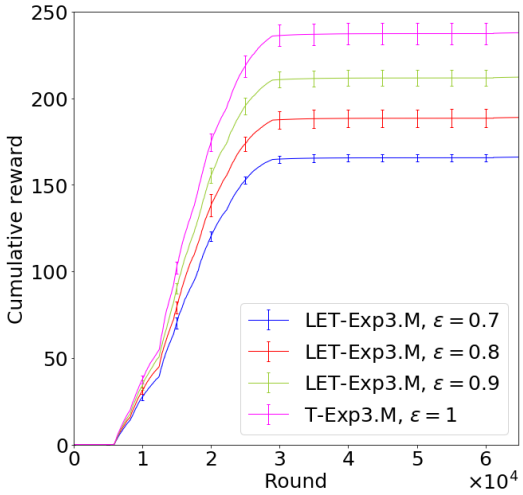


Figure 8.12: Average cumulative reward of label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, in the threshold interval setting, on the ALICE TRD conditions dataset.

Due to rewards decaying the later an observation is made after the deviation occurs and due to the importance-weighted probabilities of such algorithms not being able to react quickly, label efficient adversarial thresholding semi-bandits find the threshold interval setting a challenge.

The average cumulative pull frequencies of **T-Exp3.M** and **LET-Exp3.M** with $\epsilon \in \{0.9, 0.8, 0.7\}$ confirm a distinct lack of learning any arms deviating from previous observations by more than 3 µA (see Figure 8.13). As the maximum observation probability decreases, the total number of arm pulls decreases. The linear behaviour is due to label efficient adversarial thresholding semi-bandit algorithms selecting a fixed number of arms for observation each round.

The efficiency ratios for label efficient adversarial thresholding semi-bandits were averaged after running experiments on the ALICE TRD conditions dataset, ten times (see Table 8.4), in the threshold interval setting. As ϵ decreases, fewer arms are pulled and the ratio of arms pulled within threshold to total arms within threshold also decreases, but not significantly. The ratio of arms not pulled exceeding threshold to total arms exceeding threshold, rises slightly with the decrease in ϵ.

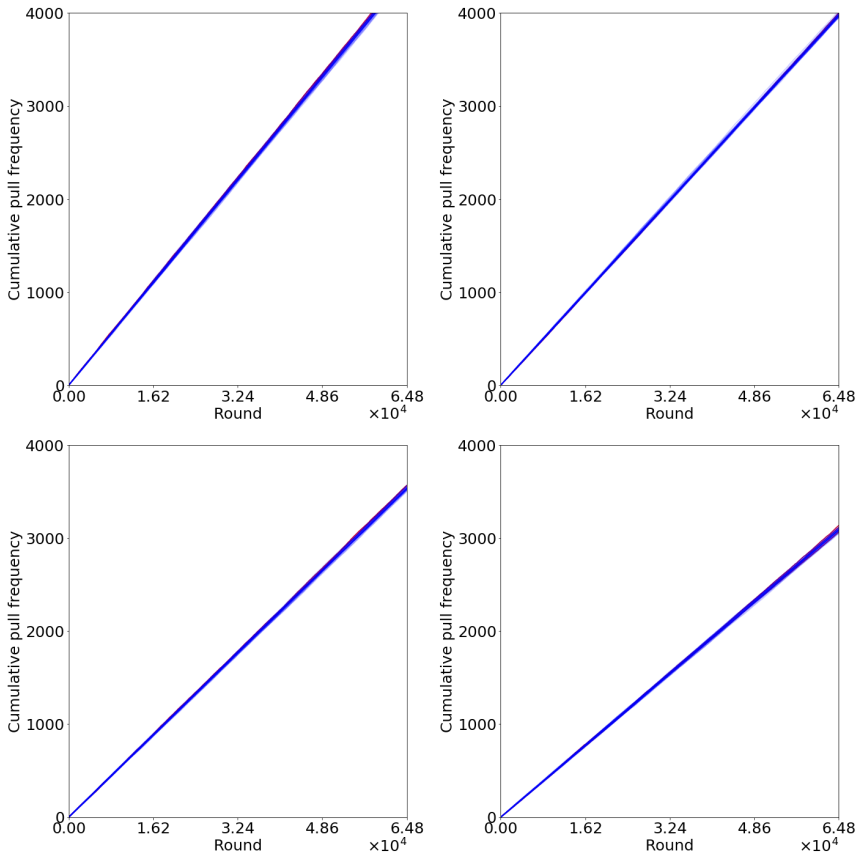


Figure 8.13: Average cumulative pull frequency for label efficient adversarial thresholding semi-bandit algorithms with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right), in the threshold interval setting, on the ALICE TRD conditions dataset. Arms $\{3, 4, 27\}$ are highlight by red lines, with remaining arms given by blue lines.

Table 8.4: Comparison of average false positive (α), negative (β) and discovery (FDR) rates of label efficient adversarial thresholding semi-bandit algorithms, in the threshold interval setting, on the ALICE TRD conditions dataset.

Algorithm		α	β	FDR
T-Exp3.M	(LET-Exp3.M, $\epsilon = 1$)	0.068	0.931	0.931
LET-Exp3.M	$\epsilon = 0.9$	0.061	0.939	0.932
	$\epsilon = 0.8$	0.055	0.945	0.932
	$\epsilon = 0.7$	0.048	0.952	0.932

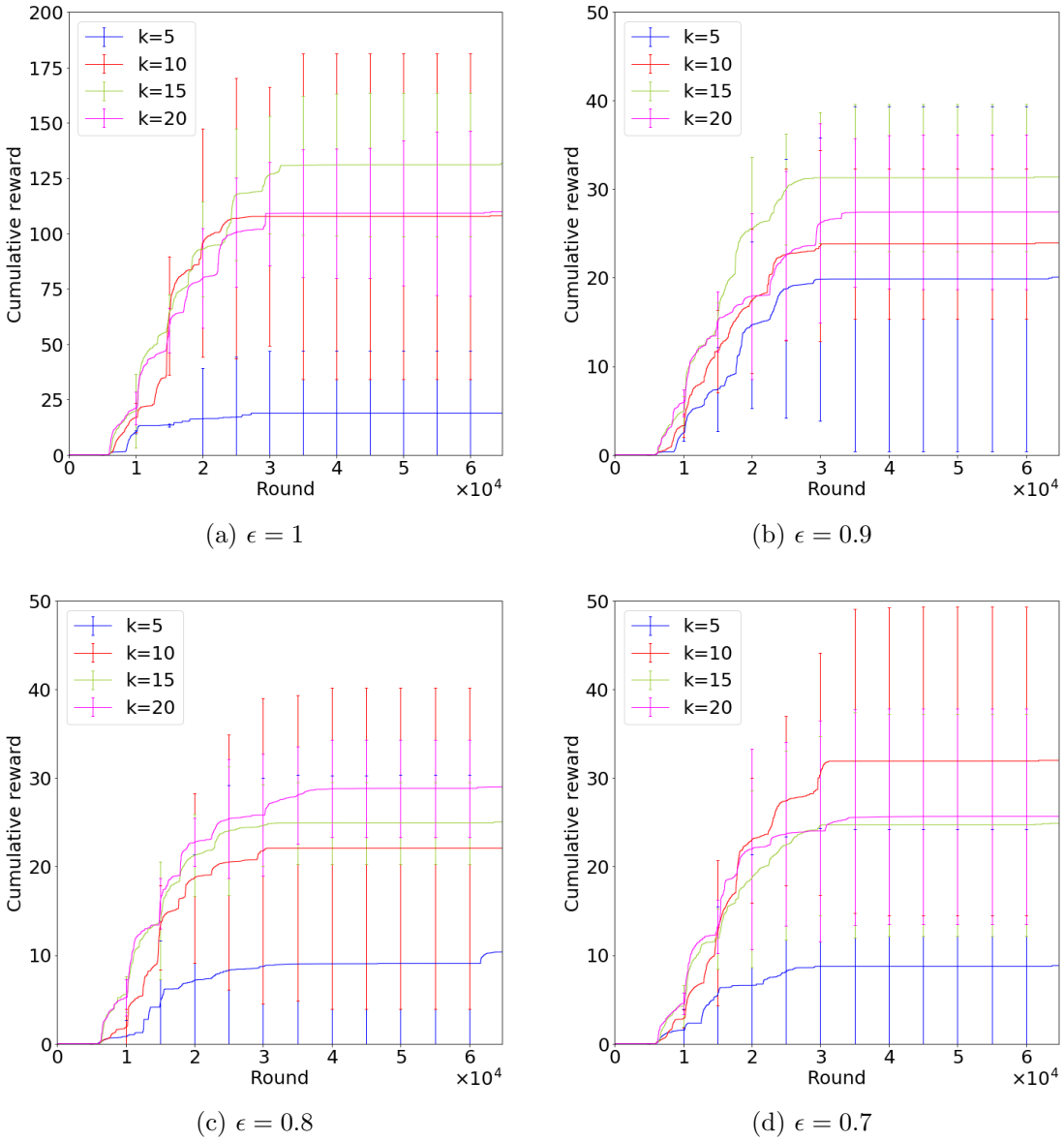
The average false positive rate α, remaining low implies that few arms are pulled within threshold, compared with those not pulled and within threshold. The average false negative rate β, remaining relatively high suggests that many arms are not pulled whilst exceeding threshold, compared with those pulled exceeding threshold. The fact that the average false discovery rate FDR remains high infers that many arms are pulled within threshold, compared with those pulled exceeding threshold. The empirical

results given in Figures 8.12, 8.13 and Table 8.4, give clear evidence that the label efficient adversarial thresholding semi-bandit algorithms introduced in Chapters 4 and 5, do not satisfy Research specification 4.

Dynamic label efficient adversarial thresholding semi-bandits: In the threshold interval setting, deviations are determined by comparison of the current and the previously most recent observations, from pulling the same arm. Dynamic label efficient adversarial thresholding semi-bandits select a fixed number of arms k, for observing each round and use the Bernoulli observation probability $\epsilon_{t,i}$, associated with arm $i \in [K]$, in round $t \in [T]$. Arms where consecutive pulls (not necessarily in consecutive rounds) are observed with feedback deviating by more than the deviation threshold are assigned to the threshold list (see Definition 4 in Section 6.2), the observation probability is reset to the maximum observation probability, $\epsilon_{t+1,i} = \epsilon$. Since the player's objective is to pull an arm precisely when feedback deviates by more than the threshold, the player receives more reward for pulling that arm closer in time (but not before) to the actual deviation.

The average cumulative reward of **dLET-Exp3.M** with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ decreases for all subsets selection sizes ($k \in \{5, 10, 15, 20\}$) evaluated, with the effect becoming less significant as ϵ decreases (see Figures 8.14 (a), (b), (c) and (d)). The repeated increase, followed by decay behaviour reflects the stepped nature of the ALICE TRD conditions dataset, where a deviation in Anode current is detected but no immediate deviations occur. As fewer arms are likely to be pulled, our analysis suggests that it becomes more difficult to identify which subset size optimises average cumulative reward.

Figure 8.15 gives the average cumulative pull frequencies for **dLET-Exp3.M**, with $k \in \{5, 10, 15, 20\}$ arms selected each round (top left to bottom right) and maximum observation probability, $\epsilon = 1$. For each subset size, arms deviating by more than the threshold most often (six, seven or the maximum of eight times) are illustrated (red) and compared with arms deviating no more than five times (light blue). The relative behaviour across all subset sizes is due to almost all arms presenting periods of relatively constant current, interspersed with relatively short periods of change in value.



(a) $\epsilon = 1$ (b) $\epsilon = 0.9$

(c) $\epsilon = 0.8$ (d) $\epsilon = 0.7$

Figure 8.14: Average cumulative reward of **dLET-Exp3.M**, with subsets of size $k \in \{5, 10, 15, 20\}$, selected each round and maximum observation probability, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ (top left to bottom right). Experiments on the ALICE TRD conditions dataset, in the threshold interval setting, were repeated ten times.

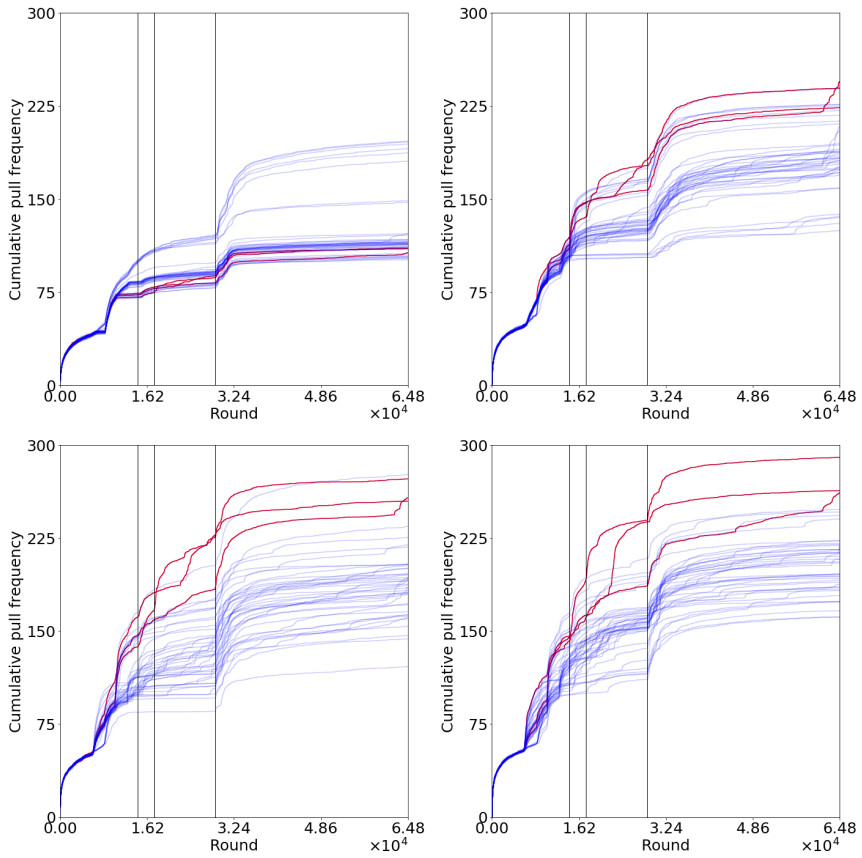


Figure 8.15: Average cumulative pull frequency for **dLET-Exp3.M** with $k \in \{5, 10, 15, 20\}$ (top left to bottom right) and $\epsilon = 1$, in the threshold interval setting, on the ALICE TRD conditions dataset. Arms $\{3, 4, 27\}$ are highlight by red lines, with remaining arms given by blue lines. Black lines indicate trip event, recovery and switch off times.

As the subset size increases, the adversarial thresholding semi-bandit player pulls all arms more often, almost as soon as a significant deviation in current occurs and then learns to not pull arms after no further deviations are observed. Also as k increases, the player becomes more effective in identifying those arms with more deviations, by observing them more often. Evidence suggests that selecting fewer arms for potential observation, gives the player fewer opportunities to make the correct observation decision. Critically, as k increases, the player takes less time to identify deviations. For $k = 10, 15, 20$, **dLET-Exp3.M** with $\epsilon = 1$, captures the switch-on of arm 4, at round $t = 61142$. Even after an extended period of stable feedback, the dynamic label efficient adversarial thresholding semi-bandit player can identify the change within 12 minutes, for $k = 20$ and within 40 minutes for $k = 15$. However, it is only possible to distinguish arms 3, 4 and 27 from sub-optimal arms, prior to the trip event (round $t = 14465$), for $k = 15$.

The false positive, false negative and false discovery rates (α, β and FDR, respectively), for **dLET-Exp3.M** with $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, are averaged after repeated experimentation on the ALICE TRD conditions dataset (see Table 8.5). The ratio of arms pulled within threshold to all arms within threshold and ratio of arms not pulled but exceeding threshold to all arms exceeding threshold, remain almost constant throughout all values of ϵ and k. The small false positive rates indicate that the number of arms not pulled with feedback deviating within threshold dominate the number of arms pulled with feedback deviating by more than the deviation threshold.

The large false negative rates show that the number of arms not pulled with feedback deviating by more than the threshold dominate the number of arms pulled with feedback deviating by more than the threshold. Both characteristics indicate that few arms have been pulled. The false discovery rate is affected by the number of arms selected for pulling each round, but remains large for all $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ and $k \in \{5, 10, 15, 20\}$. This suggests that the number of arms pulled with feedback deviating within the deviation threshold, dominates the number of arms pulled with feedback deviating by more than the deviation threshold.

Table 8.5: Comparison of average false positive (α), negative (β) and discovery (FDR) rates of **dLET-Exp3.M** with maximum observation probabilities $\epsilon \in \{1, 0.9, 0.8, 0.7\}$ and subsets selection sizes $k \in \{5, 10, 15, 20\}$, in the threshold interval setting, on the ALICE TRD conditions dataset.

ϵ	k	α	β	FDR
1	5	0.002	0.998	0.937
	10	0.003	0.996	0.914
	15	0.003	0.996	0.914
	20	0.003	0.996	0.911
0.9	5	0.002	0.997	0.921
	10	0.003	0.997	0.917
	15	0.003	0.996	0.913
	20	0.003	0.996	0.921
0.8	5	0.002	0.998	0.927
	10	0.003	0.997	0.913
	15	0.003	0.996	0.916
	20	0.003	0.996	0.914
0.7	5	0.002	0.998	0.929
	10	0.002	0.997	0.906
	15	0.003	0.996	0.912
	20	0.003	0.996	0.920

AliceBandit: Empirical analysis of **AliceBandit** with maximum number of rounds between consecutive pulls, $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes, in the threshold interval setting, has been performed using the ALICE TRD conditions dataset for comparison with other adversarial thresholding semi-bandit algorithms and using the RUN3 simulation dataset, to further demonstrate the potential for supporting sustained operation of the upgraded ALICE experiment, at the LHC.

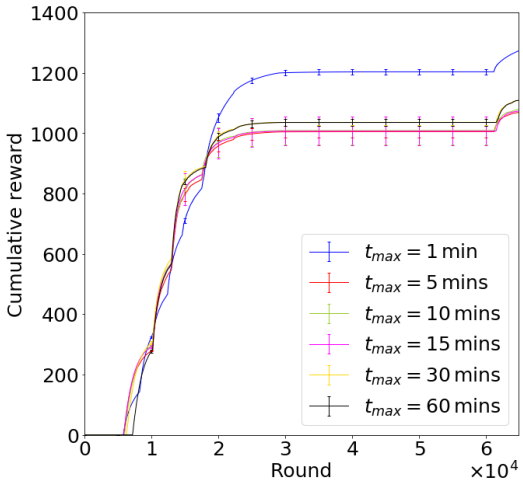


Figure 8.16: Average cumulative reward of **AliceBandit**, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes. Experiments on the ALICE TRD conditions dataset, in the threshold interval setting were repeated ten times.

The average cumulative reward of **AliceBandit** is compared over a range values for the maximum number of rounds between consecutive pulls, where $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (see Figure 8.16). The player achieves maximum cumulative reward when setting $t_{max} = 1$ minute. Minimising the maximum number of rounds between consecutive pulls increases the chances of the player identifying deviations, even after a period of time where feedback from an arm gives no indication of deviating. However, relatively high reward can be accrued, even for larger values of t_{max}. For $t_{max} = 5, 10, 15, 30, 60$ minutes, the evolution of cumulative reward appears almost identical, with corresponding *jumps* in reward at regular intervals.

The average cumulative pull frequencies for **AliceBandit** are compared with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes, on the ALICE TRD conditions dataset (see Figure 8.17). The frequency with which an **AliceBandit** player appears to progress with relative stability, implying that the algorithm is robust to all but the most significant of deviations.

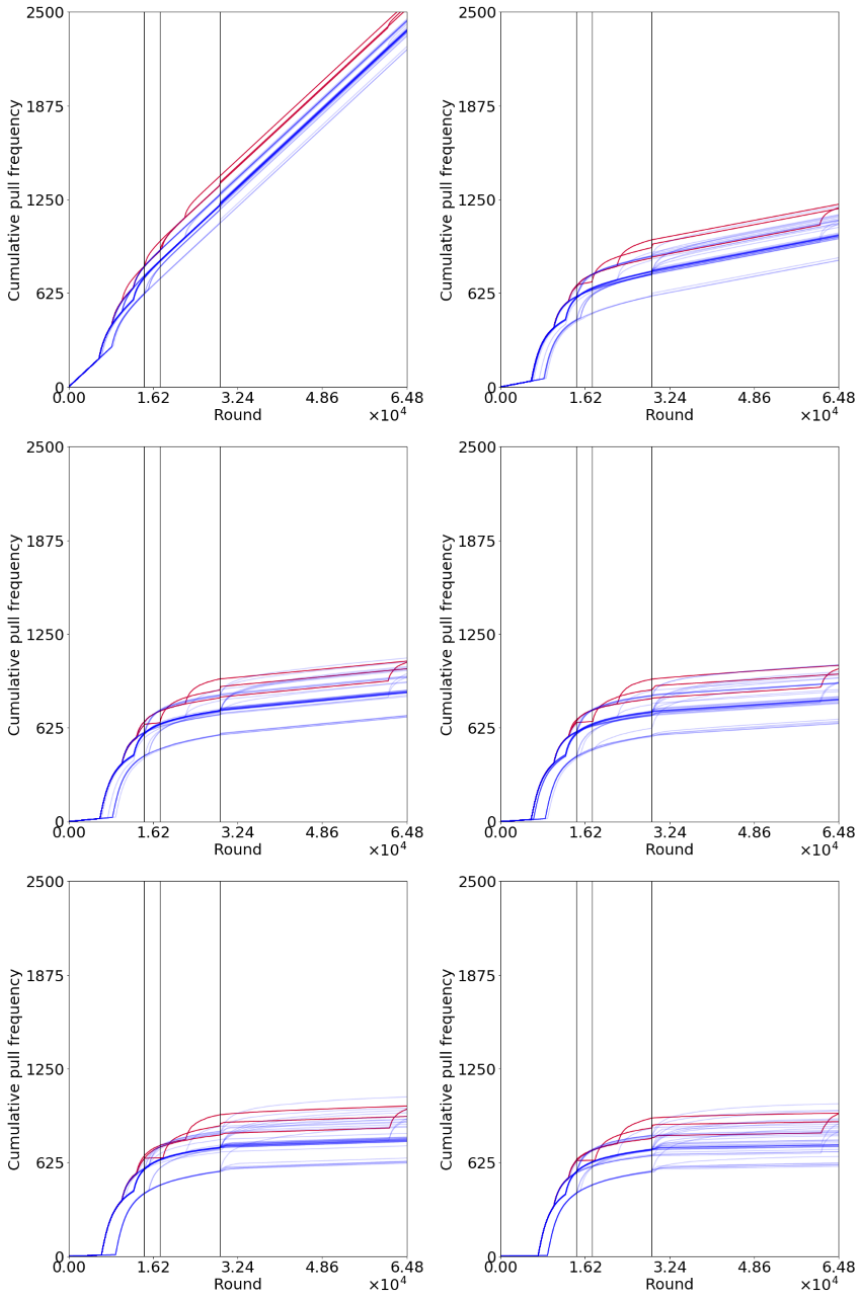


Figure 8.17: Average cumulative pull frequency for **AliceBandit** on the ALICE TRD conditions dataset, with $t_{max} = \{1, 5, 10, 15, 30, 60\}$ minutes (top to bottom) with trip, recovery and switch off indicated (black lines). Arms $\{0, 1, 2, 3, 4\}$ shown by red lines, remaining arms given by blue lines.

The fewer rounds allowed between consecutive pulls of an arm, the more effective **AliceBandit** is in identifying arms 3, 4 and 27. However, even for $t_{max} = 60$ minutes, these arms are in the eight most pulled arms over the time horizon. The linear behaviour of pull frequencies is a direct result of constraining the number of rounds between consecutive pulls, with larger t_{max} giving smaller rates of pull. Arms following a linear trend throughout the time horizon indicate no deviation being detected. When the pull frequency of an arm departs from linearity, a deviation exceeding threshold has been detected. Initial detection of deviation from linearity, recedes in time as t_{max} increases. Deviations are detected as early as round $t = 5825$ for $t_{max} = 1$ minute (4 hours 48 minutes before trip) and no later than round $t = 7205$ for $t_{max} = 60$ minutes (4 hours 2 minutes before trip).

Table 8.6: Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 10, 30, 60\}$ minutes, in the threshold interval setting, on the ALICE TRD conditions dataset.

t_{max}	α	β	FDR
1 min	0.037	0.962	0.981
5 mins	0.016	0.978	0.976
10 mins	0.014	0.980	0.974
15 mins	0.013	0.981	0.974
30 mins	0.013	0.981	0.972
60 mins	0.012	0.980	0.970

The average efficiency ratios from running **AliceBandit** in the threshold interval setting, on the ALICE TRD conditions dataset are compared for $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (see Figure 8.6). As t_{max} increases, the false positive rate decreases. Although the decrease from $t_{max} = 5$ minutes to $t_{max} = 60$ minutes is relatively small compared with the difference between $t_{max} = 1$ minute and $t_{max} = 5$ minutes. Decreasing α suggests that either the number of arms pulled with feedback deviating by less than the deviation threshold is decreasing, the number of arms pulled with feedback deviating by more than the deviation threshold is increasing, or both. The false negative rate increases slightly as t_{max} increases. This implies that either the number of arms not pulled with feedback deviating by more than the deviation threshold is increasing but only slightly, the number of arms pulled with feedback deviating by more than the deviation threshold is decreasing but only slightly, or both but even more slightly. As t_{max} increases, the false discovery rate decreases slightly, indicating that either the

number of arms pulled with feedback deviating by less than the deviation threshold is decreasing slightly, the number of arms pulled with feedback deviating by more than the deviation threshold is increasing slightly, or both but even more slightly. The fact that α remains relatively small for all values of t_{max} suggests that the number of arms not pulled with feedback deviating by less than the deviation threshold dominates the number of arms pulled with feedback deviating by more than the deviation threshold. The number of arms not pulled with feedback deviating by more than the deviation threshold appears to dominate the number of arms pulled with feedback deviating by more than the deviation threshold, since β remains relatively high. The fact that the false discovery rate remains also high, indicates that the number of arms pulled with feedback deviating by less than the deviation threshold dominates the number of arms pulled with feedback deviating by more than the deviation threshold.

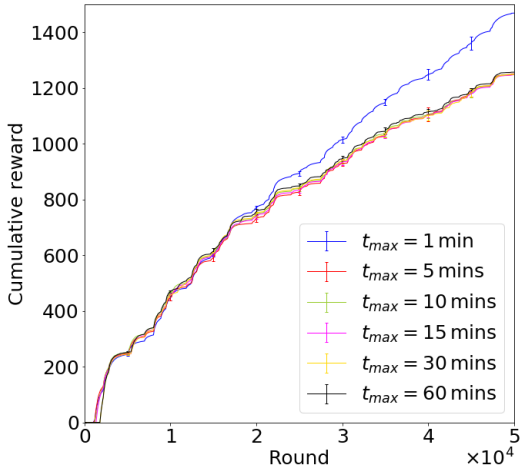


Figure 8.18: Average cumulative reward of **AliceBandit**, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes. Experiments on the RUN3 simulated dataset, in the threshold interval setting were repeated ten times.

The average cumulative reward from running **AliceBandit** on RUN3 simulated data, is compared for $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (see Figure 8.18). The cumulative reward for all values of t_{max} continues to increase over the entire time horizon, due to several arms frequently deviating by more than the deviation threshold. Whilst there appears to be no performance benefit between running **AliceBandit** with $t_{max} = 5$ minutes or $t_{max} = 60$ minutes, there is a clear advantage for $t_{max} = 1$ minute. This minimises the maximum time taken to detect deviations, but increases the inefficiency of pulling arms unnecessarily.

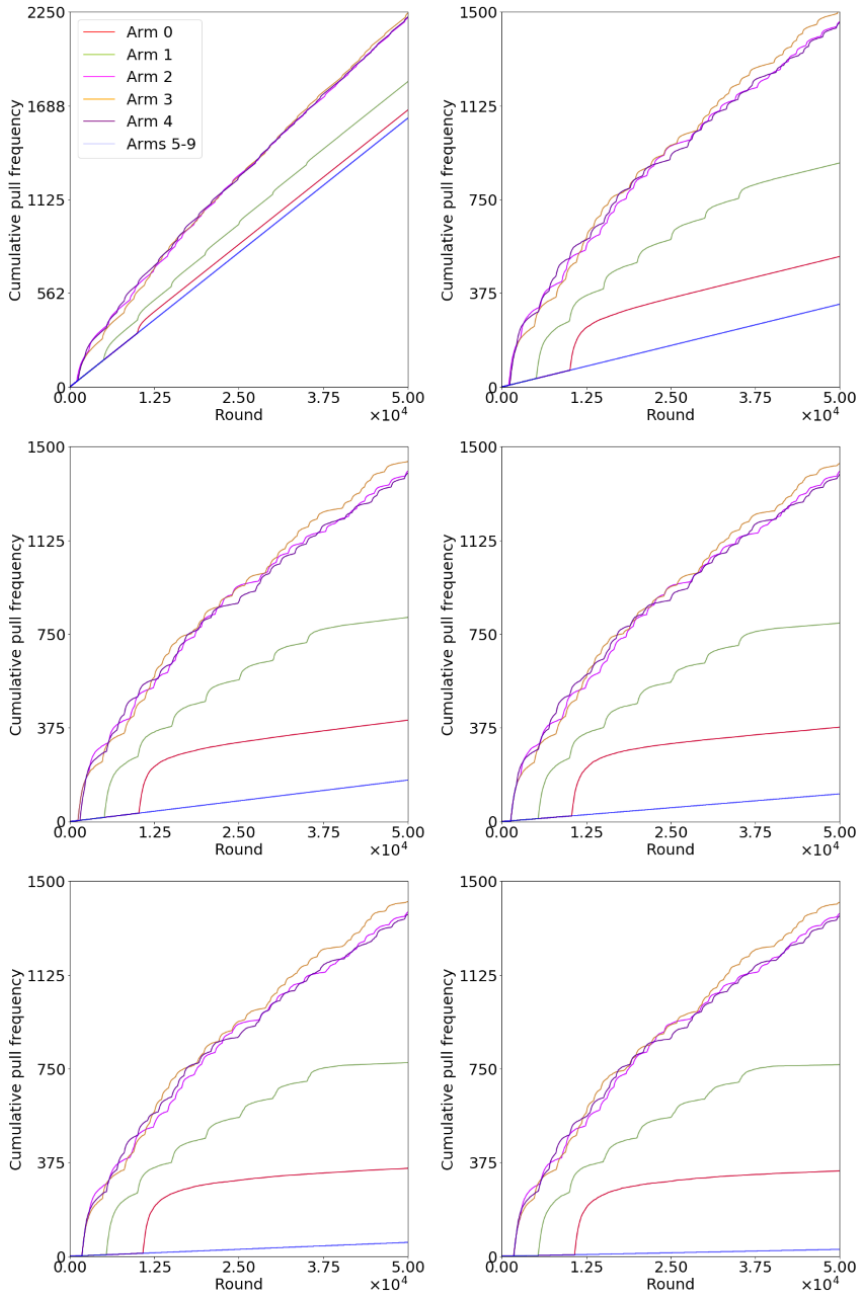


Figure 8.19: Average cumulative pull frequency for **AliceBandit** on the RUN3 simulated dataset, with $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (top left to bottom right). Arms $\{0, 1, 2, 3, 4\}$ shown by red, green, magenta, orange and purple lines, remaining arms given by light blue lines.

The average cumulative pull frequencies of **AliceBandit**, run on the ALICE RUN3 simulated dataset is compare over $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (see Figure 8.19). Arm 0 presents a single deviation by more than 3 µA, in round $t = 10000$. Arm 1 periodically exceeds 3 µA, every 5000 rounds, starting and ending below 3 µA (consisting of eight deviation of more than the deviation threshold). Arms 2, 3 and 4 simulate multiple deviation events and arms 5-9 consist of 50000 rounds of Anode current that does not deviate by more than 3 µA.

Detection of a deviation by more than the deviation threshold presents itself as a significant localised increase in cumulative pull frequency, above the baseline of pulling once every t_{max} rounds. If no further deviation is detected, the cumulative pull frequency returns to the t_{max} baseline rate (appearing parallel to other arms being pulled at the baseline rate). Subsequent deviations yield further localised increases in pull frequency and hence, arms with multiple deviations are identified by larger total cumulative pull frequencies.

As t_{max} increases, it is easier to distinguish localised increases in pull frequency, since each arm is likely to be pulled less frequently, unless unusual behaviour is detected. However, on average the time taken to detect a deviation (after it actually occurred), ranges from 45 seconds for $t_{max} = 1$ minute, to 4 minutes 26 seconds for $t_{max} = 5$ minutes, approximately 10 minutes for $t_{max} = 10, 15$ minutes and 27 minutes for $t_{max} = 30, 60$ minutes.

Table 8.7: Comparison of average false positive (α), negative (β) and discovery (FDR) rates for **AliceBandit** with maximum number of rounds between successive pulls $t_{max} = \{1, 5, 10, 15, 30, 60\}$ minutes, in the threshold interval setting, on the ALICE RUN3 simulated dataset.

t_{max}	α	β	FDR
1 min	0.036	0.959	0.929
5 mins	0.014	0.975	0.896
10 mins	0.012	0.976	0.881
15 mins	0.011	0.977	0.876
30 mins	0.010	0.977	0.869
60 mins	0.010	0.977	0.863

The average efficiency ratios from running **AliceBandit** in the threshold interval setting, on the RUN3 simulated dataset are compared for $t_{max} \in \{1, 5, 10, 15, 30, 60\}$ minutes (see Figure 8.7). The false positive and false negative rates appear to have a similar relationship to those given in Table 8.6, where **AliceBandit** is run on the ALICE TRD conditions dataset. As t_{max} increases, α decreases slightly and β increases slightly. However, the false discovery rates decrease at a higher rate on the RUN3 simulated dataset. This is potentially due to the simulated dataset having more defined periods of deviation.

8.6.4 Threshold interval summary

In this section, we review the empirical performance **LET-Exp3.M**, **dLET-Exp3.M** and **AliceBandit**, specifically for detecting deviations exceeding a deviation threshold (in the threshold interval setting), in Anode current channels using the ALICE TRD conditions dataset. We also summarise our findings from running **AliceBandit** on the RUN3 simulated dataset. The success of each algorithm is evaluated in relation to Research specifications 2 and 4.

Label efficient adversarial thresholding semi-bandits (including adversarial thresholding multi-armed bandits with multiple plays) were empirically analysed for a range of observation probabilities, $\epsilon \in \{1, 0.9, 0.8, 0.7\}$, where fewer arms are likely to be observed as ϵ decreases. Results show that the use of an importance-weighted distribution over arms is not sufficiently adaptive to detect multiple deviations exceeding a deviation threshold. Figure 8.12 suggests that performance is improved by making fewer observations, even when the optimal number of observations is known. Hence, we cannot say that **LET-Exp3.M** satisfies Research specification 2. The results given in Figure 8.13, indicate that no arm is observed significantly more often than any other. Hence, the algorithm is unable to detect any deviation exceeding the deviation threshold and Research specification 4 remains unsatisfied.

The empirical performance of dynamic label efficient adversarial thresholding semi-bandits in the threshold interval setting achieves comparable cumulative reward with **LET-Exp3.M**, when $\epsilon = 1$. However, reducing the maximum observation probability impacts both the ability to accumulate reward and identify an optimised number of arms to select for observation each round. Figures 8.14a - 8.14d, show that selecting

more than an optimal number of arms each round gives the algorithm an opportunity to select but not observe sub-optimal arms and achieve higher reward. However, selecting too many arms increases the chance of selecting and observing more sub-optimal arms.

The construction of a threshold list for **dLET-Exp3.M** has minimal effect on performance with the ALICE TRD conditions dataset due to its stepped characteristic. The results presented in Figure 8.15 suggest that the cumulative pull frequencies of each arm can be used to identify when arms have feedback that deviates by more than a given deviation threshold. Arms that deviate more often are observed more often. Preventative measures can then be taken on channels with Anode current below the ITRIPLimit and in particular, before a trip event. Hence, **dLET-Exp3.M** satisfies Research specification 2 and Research specification 4, with an appropriately sized k.

The **AliceBandit** algorithm achieves approximately five times more cumulative reward than label efficient or dynamic label efficient adversarial thresholding semi-bandits, without the need to maintain an importance-weighted distribution over all arms. Analysis on ALICE TRD conditions data and RUN3 simulations show that deviations of Anode current exceeding deviation threshold are detected within the maximum number of rounds between consecutive observations. This suggests that the algorithm is learning the behaviour of the Anode current. The evolution of cumulative pull frequencies demonstrates an effective method for identifying when the Anode current from some channel behaves anomalously.

Throughout our analysis of adversarial thresholding semi-bandit algorithms in the threshold interval setting, we have identified a consistent relationship in terms of average efficiency ratios. Regardless of subset size, maximum observation probability or maximum number of rounds between consecutive pulls, average false positive rates (α), remain small and both average false negative rates (β), and average false discovery rates (FDR), remain large. Since α defines the average ratio between the number of arms observed with deviations within threshold and the total number of arms with deviations within threshold then small α implies that all the algorithms developed in this thesis decide to not pull arms when they shouldn't, much more often than when they do decide to pull when they shouldn't. This suggests efficient behaviour. The average false negative rate defines the ratio between the number of arms not observed when their deviations exceed threshold and the total number of arms with deviations

exceeding threshold. Large β means many more arms are not pulled when deviations exceed threshold than are pulled when exceeding. The average false discovery rate is defined as the ratio between the number of arms pulled within threshold and the total number of arms pulled. Large FDR suggests that many more arms are pulled with deviations within threshold than are pulled exceeding threshold. Hence, large β and FDR are not desirable results in terms of efficiency. However, the statistical definitions of β and FDR necessitate that a true positive result occurs from pulling an arm precisely in the round when feedback deviates from the previous round by more than the threshold. The stepped nature of the datasets used in our analysis (see Figures 8.3 and 8.4) and the threshold interval setting, cause adversarial thresholding semi-bandits to perform poorly when measured against the efficiency ratios β and FDR.

8.7 Summary

In this chapter, an empirical study of the proposed algorithms has been performed. The monitoring of detector control channels in heavy-ion physics experiments was introduced as an application of adversarial thresholding semi-bandits, with the goal of efficiently observing control parameters. Ultimately, observing a parameter sufficiently means that a controller can prevent channel trips causing losses in data-taking opportunities and critically, prevent overcurrent scenarios from damaging sensitive components on the experiment.

Two approaches to the empirical study have been formalised to demonstrate how each algorithm addresses the requirements of our Research specifications (given in Section 1.3). Within each approach, ATSBP cumulative regret was used as the main metric for empirical performance, with cumulative pull frequency and efficiency ratios also considered for each algorithm, in terms of satisfying Research specifications 2 (empirical performance) and 4 (deviation detection). Consequently, we recognised the cumulative pull frequency of adversarial thresholding semi-bandits as a useful metric for experimental controllers, to decide when and if preventative measures are required during the operation of the heavy-ion experiment, at the LHC. We also discovered that our efficiency ratios were poor measures of ATSBP in the threshold interval setting.

In the fixed threshold setting, adversarial thresholding semi-bandits present similar results across all three measures of cumulative regret, cumulative pull frequency and efficiency ratios, with the exception of **AliceBandit** achieving of order ten times smaller regret. As such, we consider exponentially-weighted adversarial thresholding semi-bandits to partially satisfy Research specification 2 and **AliceBandit** to fully satisfy our empirical performance requirement. In terms of deviation detection, we measure ATSBP algorithms by cumulative reward, pull frequency and efficiency ratio. We conclude that label efficient adversarial thresholding semi-bandits fail to satisfy Research specification 4, **dLET-Exp3.M** achieves partial satisfaction and **AliceBandit** fully satisfies Research specification 4, due to achieving cumulative reward of order ten times larger than **dLET-Exp3.M**.

In addition to the empirical analysis in this chapter, we also note that all our algorithms required approximately the same time to complete their respective computations for each round (10^{-3} s), whilst running experiments on a Lenovo ThinkPad laptop and with the given software version specifications (see Section 8.4). This is due to the relatively small values of K and k used in our analysis. **AliceBandit** is more scalable (linear) in the total number of arms than exponentially-weighted adversarial thresholding semi-bandits (linearithmic), for ATSBP. Dynamic label efficient adversarial thresholding semi-bandits would theoretically approach the same scalability of **AliceBandit**, where k is small compared with K, as K grows large. However, we have already shown that even thought the subset size can be chosen, empirical performance of **dLET-Exp3.M** is affected if k is too small.

Finally, we have shown that **AliceBandit** outperforms **LET-Exp3.M** and **dLET-Exp3.M**, in both fixed threshold and threshold interval settings. We also demonstrate that **AliceBandit** in the threshold interval setting can identify deviations in Anode current channels from the ALICE Transition Radiation Detector several hours before a trip event occurs, providing a DCS operator with an opportunity to take preventative actions, such as power-cycling certain channels, without significantly impacting on data-taking during detector operation. This improves on the current system of automatically switching off (and then recovering) a channel, once the current has exceeded a given threshold (ITRIPLimit) and reduces the need to continuously observe all channels.

Chapter 9

Conclusions

In this chapter, we review the research presented in previous chapters, introducing adversarial thresholding semi-bandit problems as a novel abstraction of the multi-armed bandit model. In Section 9.1, we summarise the contributions made in Chapters 3, 4, 5, 6 and 7. Finally, in Section 9.2, we identify where further developments to adversarial thresholding semi-bandits can be made.

9.1 Summary of results

Multi-armed bandits are a classical approach for modelling the exploration-exploitation trade-off in sequential decision-making problems under uncertainty. Current bandit research involving the discovery of arms with feedback exceeding a given threshold, known as thresholding bandits (see Section 2.6 for details), only relate to stochastic sequences of feedback and are predominantly associated with pure exploration bandit paradigms. Concurrently, there is a demand from real-world applications for non-stochastic bandit models that determine which arms to pull, where feedback deviates by more than some deviation threshold, from the most recent observation for that arm. The online monitoring of multiple device power channels, controlling a heavy-ion collider experiment on the LHC, requires the efficient balancing of decisions to continue monitoring channels that have recently been observed behaving outside nominal operating parameters and deciding to observe channels that may have not been monitored for some time, due to nominal historic observations.

To address this gap in the literature, we introduce a novel abstraction of the bandit model, the adversarial thresholding semi-bandit, where the goal is to only pull those arms exceeding a given threshold (or deviating by more than a given deviation threshold, from the most recent pull), balancing the need to continue pulling arms if observations satisfy some threshold criteria, with the need to not completely forget about arms observed not meeting the threshold criteria, due to the adversarial nature of the feedback.

Following the introduction of the adversarial thresholding semi-bandit problem, we develop several bandit algorithms and measure their ability to address ATSBP through a set of Research specifications. These specifications require that ATSBP algorithms: (i) possess theoretically provable upper bounds on their performance (Research specification 1), (ii) demonstrate empirical performance on detecting feedback sequences exceeding a given threshold, from repeated experimentation (Research specification 2), (iii) have low computational complexity (Research specification 3) and (iv) demonstrate empirical performance on detecting deviations in feedback sequences, from repeated experimentation (Research specification 4).

After ATSBP is formally introduced in Chapter 3, we extended the adversarial multi-armed bandit with multiple plays to thresholding bandits, in Chapter 4. Under certain assumptions, Research specification 1 is satisfied and specification 3 is partially satisfied, achieving linearithmic computational complexity. However, Research specifications 2 and 4, fail to be satisfied. The algorithm developed in Chapter 5 generalises that of Chapter 4 and still satisfies Research specification 1, under restrictions on the number of observations the player can make (or the adversary is willing to reveal) over the entire game. Specification 3 is also partially satisfied and empirical performance decreases further as this restriction strengthens. The algorithm developed in Chapter 6, relaxes certain assumptions used in the development of algorithms in the previous two chapters. Specifications 1 and 3 are satisfied, but at a cost compared with those in Chapters 4 and 5. Research specifications 2 and 4 are partially satisfied. Our approach in Chapter 7, introduces a novel decomposition of the bandit pulling process into selection and observation decisions. The algorithm takes elements of the previous methods, significantly improving both empirical performance (Research specification 2) and computational complexity (Research specification 3). After minor modifications,

the algorithm developed in Chapter 7 also addresses the challenge of efficiently monitoring the behaviour of multiple device power channels on ALICE, in real-time, by analysing the frequency of observing each channel. In the following we outline each chapter:

- **Adversarial thresholding exponentially weighted exploration and exploitation with multiple plays (Chapter 4):** The first ATSBP algorithm is called **T-Exp3.M**. This addresses Research specification 1 and partially addresses Research specification 3, under the assumption that the number of arms exceeding threshold each round is fixed and known. The algorithm then pulls exactly this number of arms each round. However, which arms exceed the threshold remains unknown.

  In particular, **T-Exp3.M** combines multi-play adversarial bandits with the constraint that the player only receives reward when feedback from a pulled arm exceeds a given threshold and is penalised when pulling an arm within threshold.

  We have proven an upper bound on the theoretical performance of **T-Exp3.M** of $\mathcal{O}\left(\sqrt{\frac{kKT}{1-\theta}\log\left(\frac{K}{k}\right)}\right)$, where θ, k, K and T denote the threshold, number of arms pulled each round, the total number of arms and the total number of rounds played, respectively. This generalises state-of-the-art adversarial bandit algorithms, with and without multiple plays, satisfying specification 1.

  We show that the computational complexity of **T-Exp3.M** remains the same as that of the non-thresholding version, $\mathcal{O}(K(\log(k)+1))$. This is caused by the importance-weighting aspect of the algorithm and we consider it as partial fulfilment of specification 3. However, **T-Exp3.M** has no provision for the total number of arm pulls being constrained.

- **Label efficient adversarial thresholding exponentially weighted exploration and exploitation with multiple plays (Chapter 5):** We develop and algorithm that extends ATSBP to scenarios where the total expected number of arm pulls is constrained.

Under the assumption that the number of arms exceeding threshold each round is known, the *label efficient* **T-Exp3.M** (**LET-Exp3.M**) algorithm pulls no more than the same fixed number of arms each round and only decides to pull an arm based on the realisation of a Bernoulli random variable.

We have proven an upper bound of $\mathcal{O}\left(\sqrt{\frac{kKT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{K}{\delta}\right)}\right)$ on the theoretical performance of **LET-Exp3.M** with probability at least $1-\delta$, where ϵ is the Bernoulli parameter associated with deciding to observe a pulled arm. The computational complexity matches that of **T-Exp3.M**, satisfying specification 1 and partially satisfies specification 3, although the storage requirements of **LET-Exp3.M** are reduced by a factor of ϵ, compared with **T-Exp3.M**.

- **Dynamic label efficient exponentially weighted exploration and exploitation with multiple plays (Chapter 6):** The **dLET-Exp3.M** algorithm developed in this chapter is not restricted by needing to know how many arms will exceed threshold each round. The algorithm proceeds by maintaining two sets of arms. A fixed but tunable set of arms are chosen according to the same distribution as the algorithms presented previously. The decision to pull each arm is now based on the importance-weighted distribution over arms and the realisation of a dynamic Bernoulli parameter that decreases each time an arm is pulled within threshold. The Bernoulli (observation probability) parameter associated with each arm, is set to a maximum observation probability, if the arm is observed exceeding threshold. A second set of arms is updated each round, consisting only of arms pulled and observed exceeding threshold in the previous round. Irrespective of their Bernoulli parameter, these arms are definitely pulled. This feature of the algorithm is based on the argument that arms observed exceeding threshold in the previous round are highly likely to exceed threshold in the next round.

  We proved an upper bound of $\mathcal{O}\left(\sqrt{\frac{(k+K)kKT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{K}{\delta}\right)}\right)$ on the regret of **dLET-Exp3.M** with high probability. This indicates that theoretical performance suffers by a factor of order $\sqrt{k+K}$ compared with **LET-Exp3.M**. We consider this partial satisfaction of specification 1.

Empirical performance is significantly improved over previous algorithms, particularly in environments where arms switch between within threshold and exceeding threshold, satisfying specification 2. Computational complexity remains that of the previous two algorithms, again due to the importance weighting procedure and thus only partially satisfying specification 3.

- **AliceBandit (Chapter 7):** Finally, we combine the successful elements from the previous three chapters to develop an algorithm called **AliceBandit**. We introduce a novel decomposition of the bandit pulling process, into a selection and query procedure. A list of arms queried and observed exceeding threshold in the previous round, is pulled in the current round and arms are queried based on realisations of Bernoulli random variables with dynamically updated parameters. Parameters are reduced every time an arm is queried and observed within threshold, helping the algorithm to learn not to query arms that are less likely to exceed threshold. The Bernoulli parameter for an arm observed exceeding threshold is reset, enabling the algorithm to adapt in adversarial environments. A tunable parameter preventing any arm from being totally forgotten ensures that the algorithm can achieve low regret even in the most variable environments.

  We proved an upper bound of $\mathcal{O}\left(\sqrt{KH_\theta T}\right)$ on the regret of **AliceBandit**, where ATSBP hardness is defined by H_θ. This infers that theoretical performance of **AliceBandit** asymptotically scales with the number of arms, game length and the adversarial nature of feedback generated by the environment. For H_θ, not a function of T, **AliceBandit** satisfies Research specification 1.

  Whilst the parameter update rules ensure that arms unlikely to exceed threshold are queried infrequently, the arms observed exceeding threshold are queried with certainty, until they no longer exceed threshold. When parameters becomes so small that the arm is not queried for more than a given number of rounds, the player pulls that arm to check there has been no change in feedback, since no statistical assumptions have been placed on how feedback was generated. The tunability of **AliceBandit** and the domain expert knowledge from the ALICE DCS team enables the algorithm to achieve low regret and satisfy specification 2.

Since the arm selection procedure is separated from the query procedure, **AliceBandit** does not need to rely on the importance weighted bottleneck of the previous algorithms. A randomised round robin sub-routine enables the computational complexity of **AliceBandit** to be reduced to $\mathcal{O}(K)$. This satisfies Research specification 3. Research specification 4 is achieved by setting how the algorithm tests for exceedance as an interval, which is a function of the arm's most recently queried feedback.

The success of each algorithm developed in Chapters 4, 5, 6 and 7, is summarised in terms of satisfying Research specifications 1, 2, 3 and 4 (see Table 9.1). The contributions made in this thesis advance the body of knowledge in thresholding bandits to adversarial thresholding semi-bandits and the online monitoring of ALICE detector controls at the Large Hadron Collider. However, there remain many areas of interesting research to pursue.

9.2 Future directions

As shown in Chapter 8 and summarised in Table 9.1, the adversarial thresholding semi-bandit algorithms developed in this thesis satisfy almost all Research specifications and two algorithms satisfy all requirements. However, in this section, we identify several avenues that remain unexplored.

Theoretical upper bounds on the performance of each algorithm have been proved, which all converge asymptotically to proven lower bounds on non-thresholding adver-

Table 9.1: Summary of adversarial thresholding semi-bandit algorithms satisfying Research specifications. The symbols $\star$ and ★ denote partial and full satisfaction of Research specifications 1 and 3.

ATSBP algorithm	Theoretical upper bounds	Empirical performance (fixed threshold)	Computational complexity	Deviation detection
T-Exp3.M	$\mathcal{O}\left(\sqrt{\frac{kKT}{1-\theta}\log\left(\frac{K}{k}\right)}\right)$	$\star$	$\mathcal{O}(K(\log(k)+1))$	
LET-Exp3.M	$\mathcal{O}\left(\sqrt{\frac{kKT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{K}{\delta}\right)}\right)$	$\star$	$\mathcal{O}(K(\log(k)+1))$	
dLET-Exp3.M	$\mathcal{O}\left(\sqrt{\frac{(k+K)kKT}{\epsilon(1-\theta)}\log\left(\frac{K}{k}\right)\log\left(\frac{K}{\delta}\right)}\right)$	$\star$	$\mathcal{O}(K(\log(k)+1))$	$\star$
AliceBandit	$\mathcal{O}\left(\sqrt{KH_\theta T}\right)$	★	$\mathcal{O}(K)$	★

sarial bandits, either in the multi-play or semi-bandit setting. However, a theoretically proven lower bound on the performance of the adversarial thresholding semi-bandit problem remains open. Until an ATSBP lower bound has been established, the adversarial thresholding semi-bandit algorithms developed in this thesis cannot be considered optimal.

Many other multi-armed bandit models can be combined with the algorithms developed in this thesis to extend ATSBP. These include, (i) budget-limited bandits (see [71]), where label efficiency could involve a budgeted cost for pulling arms, (ii) bandits with delayed rewards (see [70]), where feedback from an arm may not be revealed to the player immediately and (iii) Graph-based thresholding bandits (see [52]), where arms may be associated with a particular location on a heavy-ion collider and feedback from an arm may reveal information regarding the feedback from other local arms.

The analysis performed on the algorithm developed in Chapter 6 does not take advantage of the dynamically updated observation probabilities, giving scope for improving on the theoretical upper bound for **dLET-Exp3.M**.

The randomised round robin arm-selection procedure, presented in Chapter 7, is computationally efficient and **AliceBandit** demonstrated empirical performance superior to other adversarial thresholding semi-bandits. However, an interesting area to research may involve efficiently adapting the selected subset-size, where the player learns how many arms to select for observation as they receive information, whilst retaining low computational complexity. There may also be the potential to paralellise certain aspects of **AliceBandit**.

Pursuing these additional avenues for research may extend the literature relating to the work in this thesis even further and reveal adversarial thresholding semi-bandits as applicable to many other interesting real-world problems.

Bibliography

[1] Agrawal, R., 1995. Sample mean based index policies by o (log n) regret for the multi-armed bandit problem. In *Advances in Applied Probability*, 27(4), pp.1054-1078.

[2] Alatur, P., Levy, K.Y. and Krause, A., 2019. Multi-Player Bandits: The Adversarial Case. In *arXiv preprint arXiv:1902.08036*.

[3] ALICE Collaboration, 2001. *ALICE Technical Design Report Transition Radiation Detector* (TRD). CERN-LHCC-2001-021.

[4] Alice Collaboration, 2018. The ALICE Transition Radiation Detector: construction, operation, and performance. In *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 881, pp.88-127.

[5] Allesiardo, R., Féraud, R. and Maillard, O.A., 2017. The non-stationary stochastic multi-armed bandit problem. In *International Journal of Data Science and Analytics*, 3(4), pp.267-283.

[6] Anjum, A., McClatchey, R., Ali, A. and Willers, I., 2006. Bulk scheduling with the DIANA scheduler. *IEEE Transactions on Nuclear Science*, 53(6), pp.3818-3829.

[7] Audibert, J.Y., Bubeck, S. and Munos, R., 2010. Best Arm Identification in Multi-Armed Bandits. In *COLT 2010*, p.41.

[8] Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R.E., 1995, October. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science* pp. 322-331.

[9] Auer, P., Cesa-Bianchi, N. and Fischer, P., 2002. Finite-time analysis of the multi-armed bandit problem. In *Machine learning*, 47(2-3), pp.235-256.

[10] Auer, P., Cesa-Bianchi, N., Freund, Y. and Schapire, R.E., 2002. The nonstochastic multiarmed bandit problem. In *SIAM journal on computing*, 32(1), pp.48-77.

[11] Auer, P. and Ortner, R., 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. In *Periodica Mathematica Hungarica*, 61(1-2), pp.55-65.

[12] Bartòk, G., 2013, June. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *Conference on Learning Theory* pp. 696-710.

[13] Bartòk, G., Foster, D.P., Pál, D., Rakhlin, A. and Szepesvári, C., 2014. Partial monitoring—classification, regret bounds, and algorithms. In *Mathematics of Operations Research*, 39(4), pp.967-997.

[14] Bernstein, S., 1946. The Theory of Probabilities. *Gastehizdat Publishing House*, Moscow.

[15] Besenyei, Á., 2018. Picard's weighty proof of Chebyshev's sum inequality. In *Mathematics Magazine*, 91(5), pp.366-371.

[16] Blackwell, D., 1954, January. Controlled random walks. In *Proceedings of the International Congress of Mathematicians* Vol. 3, pp. 336-338.

[17] Bubeck, S., Munos, R. and Stoltz, G., 2009, October. Pure exploration in multi-armed bandit problems. In *International conference on Algorithmic learning theory* pp. 23-37. Springer, Berlin, Heidelberg.

[18] Bubeck, S., Munos, R. and Stoltz, G., 2011. Pure exploration in finitely-armed and continuous-armed bandits. In *Theoretical Computer Science*, 412(19), pp.1832-1852.

[19] Bubeck, S. and Cesa-Bianchi, N., 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. In *Foundations and Trends in Machine Learning*, 5(1), pp.1-122.

[20] Bubeck, S., Wang, T. and Viswanathan, N., 2013, February. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning* pp. 258-265.

[21] Buncic, P., Krzewicki, M. and Vande Vyvre, P., 2015. Technical design report for the upgrade of the online-offline computing system (*No. CERN-LHCC-2015-006*).

[22] Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D.P., Schapire, R.E. and Warmuth, M.K., 1997. How to use expert advice. In *Journal of the ACM* (JACM), 44(3), pp.427-485.

[23] Cesa-Bianchi, N., Lugosi, G. and Stoltz, G., 2004, July. Minimizing regret with label efficient prediction. In *International Conference on Computational Learning Theory* pp. 77-92. Springer, Berlin, Heidelberg.

[24] Cesa-Bianchi, N., Lugosi, G. and Stoltz, G., 2006. Regret minimization under partial monitoring. In *Mathematics of Operations Research*, 31(3), pp.562-580.

[25] Cesa-Bianchi, N. and Lugosi, G., 2012. Combinatorial bandits. In *Journal of Computer and System Sciences*, 78(5), pp.1404-1422.

[26] Chen, W., Wang, Y. and Yuan, Y., 2013, February. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning* pp. 151-159.

[27] Chen, S., Lin, T., King, I., Lyu, M.R. and Chen, W., 2014. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems* pp. 379-387.

[28] Chernoff, H., 1952. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. In *The Annals of Mathematical Statistics*, 23(4), pp.493-507.

[29] Chung, F. and Lu, L., 2006. Old and new concentration inequalities. In *Complex Graphs and Networks*, 107, pp.23-56.

[30] Claure, H., Chen, Y., Modi, J., Jung, M. and Nikolaidis, S., 2020, March. Multi-Armed Bandits with Fairness Constraints for Distributing Resources to Human Teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* pp. 299-308.

[31] Even-Dar, E., Mannor, S. and Mansour, Y., 2002, July. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory* pp. 255-270. Springer, Berlin, Heidelberg.

[32] Even-Dar, E., Mannor, S. and Mansour, Y., 2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. In *Journal of machine learning research*, 7(Jun), pp.1079-1105.

[33] Gandhi, R., Khuller, S., Parthasarathy, S. and Srinivasan, A., 2006. Dependent rounding and its applications to approximation algorithms. In *Journal of the ACM* (JACM), 53(3), pp.324-360.

[34] Gittins, J.C., 1979. Bandit processes and dynamic allocation indices. In *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), pp.148-164.

[35] György, A. and Ottucsák, G., 2006. Adaptive routing using expert advice. In *The Computer Journal*, 49(2), pp.180-189.

[36] Habib, I., Anjum, A., Mcclatchey, R. and Rana, O., 2013. Adapting scientific workflow structures using multi-objective optimization strategies. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 8(1), pp.1-21.

[37] Hannan, J., 1957. Approximation to Bayes risk in repeated play. In *Contributions to the Theory of Games*, 3, pp.97-139.

[38] Hasham, K., Peris, A.D., Anjum, A., Evans, D., Gowdy, S., Hernandez, J.M., Huedo, E., Hufnagel, D., van Lingen, F., McClatchey, R. and Metson, S., 2011. CMS workflow execution using intelligent job scheduling and data access strategies. *IEEE Transactions on Nuclear Science*, 58(3), pp.1221-1232.

[39] Helmbold, D. and Panizza, S., 1997, July. Some label efficient learning results. In *Proceedings of the tenth annual conference on Computational learning theory* pp. 218-230. ACM.

[40] Hill, R., Devitt, J., Anjum, A. and Ali, M., 2017, June. Towards in-transit analytics for industry 4.0. *In 2017 IEEE International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 810-817). IEEE.

[41] Jamieson, K. and Talwalkar, A., 2016, May. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics* pp. 240-248.

[42] Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A., 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems* pp. 325-333.

[43] Kano, H., Honda, J., Sakamaki, K., Matsuura, K., Nakamura, A. and Sugiyama, M., 2019. Good arm identification via bandit feedback. In *Machine Learning*, 108(5), pp.721-745.

[44] Katehakis, M.N. and Robbins, H., 1995. Sequential choice from several populations. In *Proceedings of the National Academy of Sciences of the United States of America*, 92(19), p.8584.

[45] Katz-Samuels, J. and Scott, C., 2018, July. Feasible arm identification. In *International Conference on Machine Learning* pp. 2540-2548.

[46] Kleinberg, R. and Leighton, T., 2003, October. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* pp. 594-605.

[47] Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S. and Ivanov, P., 2016, May. Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* pp. 87-90.

[48] Kurepin, A., Augustinus, A., Bond, P.M., Chochula, P., Lang, J.L., Lechman, M., Pinazza, O. and Salas, K.C., 2018. ALICE DCS preparation for run 3. In *International Conference Distributed Computing and Grid-technologies in Science and Education* pp. 65-69. Rheinisch-Westfaelische Technische Hochschule Aachen.

[49] Lai, T.L. and Robbins, H., 1985. Asymptotically efficient adaptive allocation rules. In *Advances in applied mathematics*, 6(1), pp.4-22.

[50] Lattimore, T. and Szepesvàri, C., 2018. Bandit Algorithms. *Cambridge University Press.*

[51] Lattimore, T. and Szepesvàri, C., 2019, March. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory* pp. 529-556.

[52] LeJeune, D., Dasarathy, G. and Baraniuk, R., 2020, June. Thresholding Graph Bandits with GrAPL. In *International Conference on Artificial Intelligence and Statistics* pp. 2476-2485. PMLR.

[53] Lepri, B., Oliver, N., Letouzé, E., Pentland, A. and Vinck, P., 2018. Fair, transparent, and accountable algorithmic decision-making processes. In *Philosophy & Technology*, 31(4), pp.611-627.

[54] van Lingen, F., Steenberg, C., Thomas, M., Anjum, A., Azim, T., Khan, F., Newman, H., Ali, A., Bunn, J. and Legrand, I., 2005, June. The Clarens Web service framework for distributed scientific analysis in grid projects. In *2005 International Conference on Parallel Processing Workshops (ICPPW'05)* (pp. 45-52). IEEE.

[55] Locatelli, A., Gutzeit, M. and Carpentier, A., 2016, June. An optimal algorithm for the Thresholding Bandit Problem. In *International Conference on Machine Learning* pp. 1690-1698.

[56] Mannor, S. and Tsitsiklis, J.N., 2004. The sample complexity of exploration in the multi-armed bandit problem. In *Journal of Machine Learning Research*, 5(Jun), pp.623-648.

[57] Mantzaridis, P., Markouizos, A., Mitseas, P., Petridis, A., Potirakis, S., Tsilis, M. and Vassiliou, M., 2008. A High Voltage Distribution System for the ALICE Transition Radiation Detector. *ALICE-INT-2008-006 version 1.0*.

[58] McClatchey, R., Habib, I., Anjum, A., Munir, K., Branson, A., Bloodsworth, P., Kiani, S.L. and neuGRID Consortium, 2013. Intelligent grid enabled services for neuroimaging analysis. *Neurocomputing, 122*, pp.88-99.

[59] McClatchey, R., Branson, A., Anjum, A., Bloodsworth, P., Habib, I., Munir, K., Shamdasani, J., Soomro, K. and neuGRID Consortium, 2013. Providing traceability for neuroimaging analyses. *International journal of medical informatics*, 82(9), pp.882-894.

[60] Mukherjee, S., Purushothama, N.K., Sudarsanam, N. and Ravindran, B., 2017, August. Thresholding bandits with augmented UCB. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* pp. 2515-2521. AAAI Press.

[61] Neu, G. and Bartók, G., 2013, October. An efficient algorithm for learning with semi-bandit feedback. In *International Conference on Algorithmic Learning Theory* pp. 234-248. Springer, Berlin, Heidelberg.

[62] Neu, G., 2015, June. First-order regret bounds for combinatorial semi-bandits. In *Conference on Learning Theory* pp. 1360-1375.

[63] Palomo, L.V., 2017, October. The ALICE experiment upgrades for LHC Run 3 and beyond: contributions from mexican groups. In *Journal of Physics: Conference Series* Vol. 912, No. 012023, pp. 10-1088.

[64] Raginsky, M., Willett, R.M., Horn, C., Silva, J. and Marcia, R.F., 2012. Sequential anomaly detection in the presence of noise and limited feedback. In *IEEE Transactions on Information Theory*, 58(8), pp.5544-5562.

[65] Robbins, H., 1952. Some aspects of the sequential design of experiments. In *Bulletin of the American Mathematical Society*, 58(5), pp.527-535.

[66] Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J. and Auer, P., 2012. PAC-Bayesian inequalities for martingales. In *IEEE Transactions on Information Theory*, 58(12), pp.7086-7093.

[67] Steenberg, C., Bunn, J., Legrand, I., Newman, H., Thomas, M., Van Lingen, F., Anjum, A. and Azim, T., 2004, September. The clarens grid-enabled web services framework: Services and implementation. In *Proceedings of the CHEP* (Vol. 4).

[68] Streeter, M.J. and Smith, S.F., 2006. Selecting among heuristics by solving thresholded k-armed bandit problems. In *ICAPS 2006*, p.123.

[69] Thompson, W.R., 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. In *Biometrika*, 25(3/4), pp.285-294.

[70] Thune, T.S., Cesa-Bianchi, N. and Seldin, Y., 2019. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems* pp. 6541-6550.

[71] Tran-Thanh, L., Chapman, A., Rogers, A. and Jennings, N.R., 2012, January. Knapsack based optimal policies for budget-limited multi-armed bandits. In *Proceedings of the National Conference on Artificial Intelligence* Vol. 2, pp. 1134-1140.

[72] Uchiya, T., Nakamura, A. and Kudo, M., 2010, October. Algorithms for adversarial bandit problems with multiple plays. In *International Conference on Algorithmic Learning Theory* pp. 375-389. Springer, Berlin, Heidelberg.

[73] Valiant, L.G., 1984, December. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing* pp. 436-445. ACM.

[74] Vu, V.H., 2002. Chernoff type bounds for sum of dependent random variables and applications in additive number theory. In *Number theory for the millennium*, 3, pp.341-356.

[75] Wang, S. and Chen, W., 2018, July. Thompson Sampling for Combinatorial Semi-Bandits. In *International Conference on Machine Learning* pp. 5101-5109.

[76] Wang, J., Luo, J., Liu, X., Li, Y., Liu, S., Zhu, R. and Anjum, A., 2019. Improved Kalman filter based differentially private streaming data release in cognitive computing. *Future Generation Computer Systems, 98*, pp.541-549.

[77] Xie, Y., Huang, J. and Willett, R., 2012. Change-point detection for high-dimensional time series with missing data. In *IEEE Journal of Selected Topics in Signal Processing*, 7(1), pp.12-27.

[78] Yaseen, M.U., Anjum, A., Rana, O. and Antonopoulos, N., 2018. Deep learning hyper-parameter optimization for video analytics in clouds. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(1), pp.253-264.

[79] Zhong, J., Huang, Y. and Liu, J., 2017. Asynchronous parallel empirical variance guided algorithms for the thresholding bandit problem. In *arXiv preprint arXiv:1704.04567*.

[80] Zhou, D.P. and Tomlin, C.J., 2018, April. Budget-constrained multi-armed bandits with multiple plays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[81] Zhuang, H., Wang, C. and Wang, Y., 2017. Identifying outlier arms in multi-armed bandit. In *Advances in Neural Information Processing Systems* pp. 5204-5213.