Università degli Studi di Milano

Ph.D. Program in Computer Science

(XXXVII Cycle)

Department of Computer Science

A thesis submitted for the degree of

Doctor of Philosophy

Online Learning, Uniform Convergence, and a Theory of Interpretability

Subject Area: INF/01

Author
Emmanuel Esposito

Supervisor
Prof. Nicolò Cesa-Bianchi

Co-Supervisor
Prof. Massimiliano Pontil

PhD Coordinator
Prof. Roberto Sassi

Academic Year 2023–2024

To my parents and my brother

Abstract

This doctoral thesis covers various aspects of theoretical machine learning relative to two of its most fundamental paradigms: batch learning and online learning. In particular, we address the role of feedback models for multiple online learning problems, the sample complexity for uniform convergence, and a learning-theoretic approach to interpretable machine learning. First, we focus on online learning and investigate variants of the multi-armed bandit problem, including settings with feedback graphs, expert advice, and delayed feedback. We improve bounds on the minimax regret for undirected, strongly observable feedback graphs and develop nearly optimal algorithms for directed, stochastic feedback graphs without prior information on the distribution of the graphs. Additionally, we derive improved regret bounds for bandits with expert advice and explore the impact of intermediate observations in the delayed feedback setting, designing a meta-algorithm to achieve near-optimal regret which shows a reduced effect of the total delay. Second, we study the uniform convergence property of real-valued function classes with finite fat-shattering dimension. We provide an improved bound on the sample complexity of uniform convergence, closing the gap with existing lower bounds. Finally, regarding interpretability, we establish a taxonomy for approximating complex binary concepts with interpretable models such as shallow decision trees. Leveraging uniform convergence for Vapnik-Chervonenkis classes and von Neumann's minimax theorem, we achieve a surprising trichotomy for interpretable concepts while revealing connections between interpretable approximations and boosting.

Sommario

Questa tesi di dottorato affronta vari aspetti della teoria del machine learning relativamente a due fra i suoi paradigmi più importanti: batch learning e online learning. In particolar modo, gli studi contenuti in questa tesi riguardano il ruolo di vari modelli di feedback in problemi di online learning, la complessità campionaria della convergenza uniforme, e un approccio teorico e formale al machine learning interpretabile. Inizialmente, ci concentriamo sul modello di online learning e analizziamo varianti del famoso problema del multi-armed bandit, fra cui bandit con grafo di feedback, con raccomandazioni da esperti e con feedback posticipato. Miglioriamo le garanzie sul minimax regret per grafi di feedback non orientati e fortemente osservabili. Sviluppiamo inoltre algoritmi quasi-ottimali per il caso di grafi di feedback diretti e stocastici, senza alcuna informazione a priori sulla distribuzione dei grafi. Forniamo garanzie migliori anche per il minimax regret del problema di multi-armed bandit con suggerimenti da esperti, e analizziamo l'impatto di osservazioni intermedie nel caso di feedback con ritardo tramite un meta-algoritmo che ottiene regret quasi-ottimale con una riduzione dell'effetto negativo causato dal ritardo cumulativo totale. Successivamente, studiamo la proprietà di convergenza uniforme per classi di funzioni a valori reali con fat-shattering dimension finita. Deriviamo un miglioramento sul limite superiore alla complessità campionaria della convergenza uniforme, chiudendo il divario di garanzie precedenti rispetto ai limiti inferiori noti. Infine, riguardo il problema dell'interpretabilità, stabiliamo una tassonomia per l'approssimazione di concetti binari complessi tramite modelli intepretabili come, ad esempio, alberi decisionali poco profondi. Sfuttando stumenti quali la convergenza uniforme per classi Vapnik-Chervonenkis di funzioni e il teorema minimax di von Neumann, otteniamo una sorprendente tricotomia per concetti interpretabili, rivelando al contempo delle connessioni sufficientemente profonde tra approssimazioni interpretabili di concetti e il framework classico di boosting.

Acknowledgments

I would like to start by expressing my gratitude to my supervisors Nicolò and Massimiliano. Nicolò, I want to thank you for your unwavering support, for guiding me through the endeavors of academic research, and for being an endless source of inspiration. I do believe that I would have taken a different path without discovering my love and interest in theoretical machine learning had I not met you as a lecturer during my undergraduate studies. Massi, thank you for being so available to me, especially when I was still trying to figure out my way at the beginning of my PhD, and for allowing me to remotely engage with the wonderful people in Genova.

I also want to thank my mom Massima, my dad Giovanni, and my brother Francesco for always supporting me in any possible way, for being proud of my achievements, and for giving me the strength to pursue my dreams. I would not be the person I am today had you not been by my side. I thank my grandparents Antonietta and Carmine, and my late grandfather Nicola; even if our distance (be it physical or not) keeps us apart, you are always on my mind and in my heart.

I thank my splendid and exceptional coauthors Andrea, Antoine, Dirk, Federico, Hao, Julia, Khaled, Marco, Max, Nataly, Rob, Saeed, Shay, Tom, Yevgeny, and Yishay for all the stimulating brainstorming sessions and interactions, and for making research immeasurably more fun. I thank all my other beautiful colleagues (former and not) from LAILA, Alberto, Andrew, Giulia, Juliette, Kyoungseok, Luigi, Lukas, Matilde, and Pierre, for making the lab such a gezellig (Dirk, I hope I used it in the right way) and lovable place. I will always keep all the moments we shared in my heart.

I thank my friends from my undergraduate studies. A special thank you goes to Alessia, Andrea, Carlo, Dario, Davide, Elvis, Ettore, Francesca, Francesco, Gabriele, Gianluca, Luca F., Luca G., and Marco for being there since the very beginning of my adventure and still bearing with me to this day. The same goes for Andrea, Alessandro, Federico, Marco, Paolo, and Roberto, for supporting each other and for getting back in touch even when life felt so uncertain; in addition to this, Alessandro, thank you for every single invite to your place in Rota d'Imagna, as it is because of you that I met Arianna, Luca, Luiz, Maddalena, Mattia, Paola, Paolo, Samuele, and Sara, to whom I am grateful for appreciating me and warmly welcoming me as if we had known each other for ages.

Furthermore, I thank Tomer Koren, Haipeng Luo, and Panayotis Mertikopoulos for kindly accepting to be reviewers of this doctoral thesis and for providing useful and detailed comments that significantly helped to improve the presentation of the contents of this manuscript.

Finally, I want to thank all the bright colleagues, senior researchers, and friends I had great interactions with during various occasions, such as conferences, summer schools, workshops, research visits, or simply personal trips. There are probably too many to list here, but I want to thank you regardless for making me feel part of a wonderful community.

Ringraziamenti

Vorrei iniziare esprimendo la mia gratitudine ai miei supervisori Nicolò e Massimiliano. Nicolò, voglio ringraziarti per il tuo incrollabile supporto, per avermi guidato attraverso la mia avventura nella ricerca accademica e per essere una fonte infinita di ispirazione. Sono convinto che avrei intrapreso una strada diversa senza scoprire il mio amore per la ricerca se non ti avessi incontrato come docente durante i miei studi. Massi, grazie per essere stato così disponibile con me, soprattutto quando stavo ancora cercando di capire la mia strada all'inizio del dottorato, e per avermi permesso di interagire con le meravigliose persone del gruppo di Genova.

Voglio ringraziare mia mamma Massima, mio papà Giovanni e mio fratello Francesco per avermi sempre supportato in ogni modo possibile, per essere orgogliosi dei miei risultati e per avermi dato la forza di perseguire i miei sogni. Non sarei la persona che sono oggi se non foste stati al mio fianco. Ringrazio i miei nonni Antonietta e Carmine e il mio defunto nonno Nicola; anche se la nostra distanza (fisica o meno) ci tiene separati, siete costantemente nei miei pensieri e nel mio cuore.

Ringrazio i miei eccezionali e insostituibili collaboratori Andrea, Antoine, Dirk, Federico, Hao, Julia, Khaled, Marco, Max, Nataly, Rob, Saeed, Shay, Tom, Yevgeny e Yishay per tutte le stimolanti sessioni di brainstorming e interazioni, e per aver reso la ricerca incommensurabilmente più divertente. Ringrazio tutti gli altri miei stupendi colleghi (ex e non) di LAILA, Alberto, Andrew, Giulia, Juliette, Kyoungseok, Luigi, Lukas, Matilde e Pierre, per aver reso il laboratorio un posto così gezellig (Dirk, spero di averlo usato nel modo giusto) e accogliente. Terrò sempre nel mio cuore tutti i momenti che abbiamo condiviso.

Ringrazio i miei amici dei miei studi universitari. Un ringraziamento speciale va ad Alessia, Andrea, Carlo, Dario, Davide, Elvis, Ettore, Francesca, Francesco, Gabriele, Gianluca, Luca F., Luca G. e Marco per essere stati presenti fin dall'inizio della mia avventura e per avermi sopportato fino ad oggi. Lo stesso vale per Andrea, Alessandro, Federico, Marco, Paolo e Roberto, per esserci supportati a vicenda e per esserci rimessi in contatto anche quando la vita sembrava così incerta; oltre a questo, Alessandro, grazie per ogni singolo invito a casa tua a Rota d'Imagna, poiché è grazie a te che ho conosciuto Arianna, Luca, Luiz, Maddalena, Mattia, Paola, Paolo, Samuele e Sara, a cui sono grato per avermi apprezzato e accolto calorosamente come se ci conoscessimo da sempre.

Ringrazio inoltre Tomer Koren, Haipeng Luo e Panayotis Mertikopoulos per aver gentilmente accettato di essere revisori di questa tesi di dottorato e per aver fornito commenti preziosi e dettagliati che hanno contribuito in modo significativo a migliorare la presentazione dei suoi contenuti.

Infine, desidero ringraziare tutti i brillanti colleghi, i ricercatori senior e gli amici con cui ho avuto grandi interazioni in varie occasioni, come conferenze, scuole di dottorato, workshop, visite di ricerca o semplicemente viaggi personali. Probabilmente siete troppi da elencare qui, ma desidero ringraziarvi comunque per avermi fatto sentire parte di una meravigliosa comunità internazionale.

Preface

The contents of this dissertation are based on the following published papers or manuscripts under review, emerged from different collaborations with the respective co-authors during my PhD activities. Here is a comprehensive list divided into chapters:

- Chapter 3 is based on the published conference paper:

  > Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 46122–46133. Curran Associates, Inc., 2023b.

  In this paper, I have first co-authorship, and I had a main contribution to all the results therein.

- Chapter 4 is based on the under-review manuscript:

  > Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *arXiv preprint*, arXiv:2406.16802, 2024. URL `https://arxiv.org/abs/2406.16802`. Under review at *Journal of Artificial Intelligence Research*.

  In this work, authors are in alphabetical order, and I contributed to the design of the proposed algorithms, central results for its analysis, and the regret lower bound.

- Chapter 5 is based on the published conference paper:

  > Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 34776–34788, 2022.

  In this paper, I have first co-authorship, and I contributed to the design and the analysis of the first main algorithm, the definition and corresponding properties of the refined graph parameters required by the second algorithm, and the regret lower bounds.

- Chapter 6 is based on the published conference paper:

  > Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9374–9395. PMLR, 2023.

  In this paper, I have first co-authorship, and I had a main contribution to all the results therein.

- Chapter 8 is based on the accepted journal paper:

  > Roberto Colomboni, Emmanuel Esposito, and Andrea Paudice. An improved uniform convergence bound with fat-shattering dimension. *Information Processing Letters*, 188, 2025.

  In this work, authors are in alphabetical order, and I had a main contribution to all the results therein.

- Chapter 9 is based on the published conference paper:

  > Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 648–668. PMLR, 2024a.

  In this work, authors are in alphabetical order, and I contributed to the definition of the main model, the algebraic characterization of approximable and interpretable concepts, and the boosting results.

The following published conference paper has also been the result of my activities during the PhD program:

> Marco Bressan, Emmanuel Esposito, and Maximilian Thiessen. Efficient algorithms for learning monophonic halfspaces in graphs. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 669–696. PMLR, 2024b.

It has not been included in this thesis because, while related to both batch and sequential learning aspects, it specifically focuses on properties of hypothesis classes determined by special subsets of vertices in graphs. It thus felt somewhat tangential to the main narrative of this dissertation.

Contents

Chapter 1

Introduction

In the last few decades, machine learning has become a prominent field with a plethora of impactful real-world applications, gradually improving on the resolution of increasingly complex and structured problems at a fast pace. The core idea of machine learning is to automatically extract information from data by learning patterns in order to perform accurate decisions or predictions on unseen data. Machine learning algorithms have been fundamental tools in solving tasks such as clinical trials, image recognition, natural language processing, and sequential decision-making in dynamical environments. The advantage of adopting machine learning approaches is provided by their remarkable adaptability to learn and generalize from previously gathered data to unobserved scenarios. It is therefore of the utmost importance nowadays to deepen our understanding relative to how and why some of these approaches perform remarkably well in practice, so as to develop novel machine learning algorithms with improved performance and trustworthiness. This is the central motivation for defining formal models that allow us to study machine learning problems.

The most traditional framework for the mathematical analysis of learning algorithms is that of statistical learning theory, consisting of a statistical approach whose roots go back to the foundational work of Vapnik and Chervonenkis (1971),[*] and Valiant (1984). Their research laid the groundwork for statistical learning theory and, more precisely, for what is known as *batch learning.* Batch learning is a traditional machine learning paradigm where learning algorithms are trained on a fixed set of data collected, and possibly labeled, beforehand. Its utility lies in the possibility to rigorously quantify how well an algorithm generalizes, that is, performs well on new inputs. In particular, Vapnik and Chervonenkis (1971) characterized learnability within a statistical learning paradigm in terms of training set size, whereas Valiant (1984) later tackled the computational aspects of batch learning problems via the famous Probably Approximately Correct (PAC) learning framework. These theoretical models and their related insights have driven a vast portion of the research in machine learning.

While batch learning has been successful in a variety of domains, we need to consider situations where data is continuously generated and arrives sequentially, or whenever decisions must be performed in real time, often under uncertainty. This is indeed the case for real-world settings such as digital markets, online advertising, recommender systems, and adaptive clinical trials, which span a multitude of relevant fields. We therefore need a different machine learning paradigm that captures the sequential nature of these scenarios, which becomes especially crucial as technology

[*] An English translation by Seckler for the original article in Russian is: Vapnik and Chervonenkis (2015).

permeates everyday life more and more. *Online learning* addresses these challenges of dynamical environments by moving towards learning models that can be updated as soon as they observe new data, a property that is in stark contrast with the previous train-test procedure. This leads to the design and study of online learning algorithms that, unlike batch learning where all the training data is available upfront, are provided with one instance at a time. The pioneering work of Littlestone and Warmuth (Littlestone, 1990, Littlestone and Warmuth, 1994), and independently by Vovk (1990), introduced the online learning framework, allowing the development and the analysis of efficient and effective sequential decision-making algorithms. This milestone was the first step towards successful areas of research in machine learning such as reinforcement learning.

It is clear that both batch and online learning are equally important for the advancement of the overall field of machine learning. In this thesis, we delve into these two research areas individually and we provide relevant contributions to both.

1.1 Thesis Outline

The structure of this thesis consists of two main parts, each revolving around a different main topic.

Part I: The Role of Feedback in Online Learning. The first part of this dissertation concerns online learning problems. We provide novel results that further advance the understanding for some of the most important feedback models. After an introductory overview on the foundations of online learning in Chapter 2, we begin with our novel results on the minimax regret for online learning with feedback graphs in Chapter 3. In Chapter 4, we utilize similar techniques to derive improved regret guarantees for the multi-armed bandit problem with experts advice, whereas we extend the graph feedback model to consider probabilistic feedback via stochastic feedback graphs and obtain near-optimal guarantees in Chapter 5. Finally, in Chapter 6 we move to the online learning model with delayed bandit feedback and study a variation including intermediate observations, proving high-probability regret guarantees and nearly matching lower bounds.

Part II: Uniform Convergence and a Theory of Interpretability. The second part of this dissertation discusses topics related to statistical learning theory from a batch learning perspective, differently from the first part of this thesis. Basics concepts and definitions, together with foundational results of statistical learning, are provided in Chapter 7. Chapter 8 investigates the sample complexity for the uniform convergence property of real-valued function classes, a fundamental concept in learning theory and beyond. There we manage to close the gap in previously known results. Then, we move our focus to Chapter 9, where we propose a formal learning-theoretic model for the interpretability of binary concepts and derive a taxonomy of interpretable approximations.

1.2 Definitions and Notations

We denote by $\mathbb{N}$ the set of natural numbers $\{0, 1, 2, \dots\}$, by $\mathbb{N}^+ := \mathbb{N} \setminus \{0\}$ the set of positive natural numbers, by $\mathbb{R}$ the set of real numbers, and by $\mathbb{R}_{\geq 0}$ the set of non-negative real numbers (and so on). If A is any finite set, the number of elements in A is denoted by $|A|$. Fix any positive integer $n \in \mathbb{N}^+$. We denote by $[n] := \{1, \dots, n\}$ the set containing the first n positive integers. For notational convenience, we freely identify n-dimensional real-valued vectors with real-valued functions having $[n]$ as domain (i.e., $\mathbb{R}^n \cong \mathbb{R}^{[n]}$ via $x \mapsto (i \mapsto x_i)$). More generally, we will oftentimes treat a real-valued function $f\colon A \to \mathbb{R}$ with a finite domain A of size $|A| = n$ as an n-uple $(f(a))_{a\in A} \in \mathbb{R}^n$ with components possibly ordered according to some natural order over A, if any, and vice versa. The inner product, or dot product, between two Euclidean vectors $x, y \in \mathbb{R}^n$ is defined as $\langle x, y\rangle \coloneqq \sum_{i=1}^n x_i y_i$ or, alternatively, as $x^\top y$. For any $p \in (0,\infty)$, we denote the p-norm of a vector $x \in \mathbb{R}^n$ as $\|x\|_p \coloneqq \bigl(\sum_{i=1}^n |x_i|^p\bigr)^{1/p}$; whenever the subscript is unspecified, $\|\cdot\| = \|\cdot\|_2$ is considered as the Euclidean norm. We let Δ_n be the simplex $\{p \in [0,1]^n : \|p\|_1 = 1\}$, and we analogously extend its definition so that $\Delta_A \coloneqq \bigl\{p \in [0,1]^A : \|p\|_1 = 1\bigr\}$ denotes the family of probability distributions over any finite domain A. The indicator function $\mathbb{I}\{E\} \in \{0,1\}$ for some event E is defined as $\mathbb{I}\{E\} \coloneqq 1$ if event E occurs, and $\mathbb{I}\{E\} \coloneqq 0$ otherwise. Moreover, we define positive thresholding by $[x]_+ \coloneqq \max\{x, 0\}$ for any $x \in \mathbb{R}$.

Throughout this manuscript, we adopt the Bachmann-Landau symbols with the following meaning: for any two functions $f, g\colon \mathbb{R} \to \mathbb{R}$ of the same variable $x \in \mathbb{R}$, $f(x) = \mathcal{O}(g(x))$ denotes an upper bound $f(x) \lesssim g(x)$ ignoring positive multiplicative constants, $f(x) = \Omega(g(x))$ is the analogue for the lower bound $f(x) \gtrsim g(x)$, and $f(x) = \Theta(g(x))$ is the analogue for both upper and lower bounds $f(x) \approx g(x)$. We use $\widetilde{\mathcal{O}}(\cdot)$, $\widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ with a tilde to ignore up-to-polylogarithmic factors while preserving the same overall meaning of the notation. Furthermore, assuming we are interested in the limit as $x \to \infty$, $f(x) = o(g(x))$ means that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$ and, similarly, $f(x) = \omega(g(x))$ means $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \infty$.

Part I

The Role of Feedback in Online Learning

Chapter 2

Online Learning

This introductory chapter provides an overview on the foundations of online learning. Its purpose is to concisely define the main concepts related to prediction with expert advice and the multi-armed bandit problem, with related well-studied algorithmic techniques, and to introduce variations of their feedback models that generalize to a broader spectrum of sequential decision-making problems. We refer the reader to Cesa-Bianchi and Lugosi (2006), Hazan (2016), Orabona (2019), Lattimore and Szepesvári (2020), Slivkins (2019, 2024) for further material on topics related to the content of this chapter.

2.1 Foundations of Online Learning

The first part of this thesis has its main focus on *online learning*, an important field of machine learning involving sequential decision-making problems where an agent has to perform predictions, or decisions, in a sequential manner while updating its strategy based on the information gathered over time. Unlike traditional machine learning problems—typically addressed as batch learning—where the learner is given full access to a large-enough dataset before being asked to make predictions, in online learning tasks the decision-maker must learn incrementally from a stream of data. Online learning clearly has a wide range of applications where data is being generated continuously by sources such as financial markets, sensors, and user interactions. Real-world domains that require, or would benefit from, the versatility of online learning algorithms thus include finance and marketing, as well as healthcare and medicine, just to name a few.

The online learning framework formalizes the sequential decision-making problem as a repeated game between a learner and an adversary. The learner is required to perform decisions over $T \in \mathbb{N}^+$ rounds, selecting an action I_t at each round $t \in [T]$ from a specific non-empty *action set* V. The adversary, sometimes referred to as nature or the environment, then reveals the loss function $\ell_t \in [0,1]^V$ (or some information about it) to the learner, who in turn suffers loss $\ell_t(I_t)$ for its decision. Throughout this manuscript, we consider the case when the adversary is *oblivious*, meaning that it chooses the entire sequence of loss functions $\ell_1, \dots, \ell_T$ before the beginning of the game, in the worst way possible for the specific learner facing it.

2.1.1 Prediction with Expert Advice

The most notable setting of such a problem is that of *prediction with expert advice* (Cesa-Bianchi, Freund, Helmbold, Haussler, Schapire, and Warmuth, 1993, Littlestone and Warmuth, 1994, Freund and Schapire, 1997, Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997), where the learner observes the prediction of $K \geq 2$ fixed experts at each round t and selects one expert I_t possibly at random. In particular, if we denote the experts as the first K positive integers, the action set consists of the finite set $V \coloneqq [K]$. The performance of the learner is measured as the *expected regret*

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i \in V} \sum_{t=1}^{T} \ell_t(i) \;, \tag{2.1}$$

defined as the difference—in expectation with respect to the random draw of $I_1, \dots, I_T$—of the cumulative loss of the learner and the cumulative loss of the best fixed expert in hindsight. A summary of the protocol for the prediction with expert advice is provided in Online Protocol 2.1.

One may equivalently define the same problem by considering a decision-maker that selects a distribution p_t over experts from the probability simplex Δ_K over the K experts and, hence, the regret becomes $R_T = \sum_{t=1}^{T} \langle \ell_t, p_t \rangle - \min_{p \in \Delta_K} \sum_{t=1}^{T} \langle \ell_t, p \rangle$. Indeed, the notion of regret becomes equivalent because the loss of the learner at round t is $\langle \ell_t, p_t \rangle = \mathbb{E}\big[\ell_t(I_t)\big]$, where $I_t \sim p_t$, and the cumulative loss of the comparator also is $\min_{p \in \Delta_K} \sum_{t=1}^{T} \langle \ell_t, p \rangle = \min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i)$. We nevertheless consider the first formalization of this problem by convention as it better relates to the definition of the other problems considered in this thesis.

Online Protocol 2.1: Prediction with expert advice

```
environment: losses $\ell_t \in [0,1]^V$ for all $t \in [T]$
for $t = 1, \dots, T$ do
    The learner selects an expert $I_t \in V$ (possibly at random)
    The learner suffers loss $\ell_t(I_t)$
    The learner observes losses $(\ell_t(i))_{i \in V}$
```

Our primary goal for this type of problems is the design of online learning algorithms whose regret grows slower than the time horizon T, a concept that is captured by the following definition. We say that a learning algorithm for an online learning problem within the above framework is *no-regret* if $R_T = o(T)$ for $T \to \infty$. Considering the problem of prediction with expert advice, the most immediate strategy for a learner would consist of selecting an action that minimizes the cumulative loss up to the current round, that is,

$$I_t \in \arg\min_{i \in V} \sum_{s=1}^{t-1} \ell_s(i)$$

for $t \geq 2$ whereas the first action I_1 is chosen arbitrarily. This strategy takes the name of Follow The Leader (FTL) and, while performing well in stochastic settings where losses are generated as an i.i.d. sample according to some fixed distribution, and even being minimax optimal when losses are Bernoulli random variables in particular (Kotłowski, 2018), it is known to perform poorly in adversarial environments as it can achieve regret $R_T = \Theta(T)$—see, e.g., Cesa-Bianchi and Lugosi (2006, Section 4.3) or De Rooij, Van Erven, Grünwald, and Koolen (2014, Section 5.1).

To overcome the shortcomings of FTL, we can introduce a controlled amount of randomness in the choice of the experts. A more clever strategy consists of keeping a weight $w_t(i)$ over each expert $i \in V$ at every round $t \in [T]$. These weights are initially equal across all experts in the absence of any information about their loss, and are updated over time depending on the performance of each expert according to the observed losses. In particular, at the end of each round t the update of the weight $w_t(i)$ for expert i is multiplicative with factor $\exp(-\eta\ell_t(i))$, where $\eta > 0$ is the learning rate that can be appropriately tuned. We can then use these weights to define a probability distribution p_t such that $p_t(i) \propto w_t(i) = \exp\bigl(-\eta \sum_{s=1}^{t-1} \ell_s(i)\bigr)$, which we can use to sample the expert I_t. This popular algorithm is known as Hedge (whose pseudocode is provided in Algorithm 2.1) or, alternatively, with the names of Exponential Weights algorithm or Exponentially Weighted Average Forecaster. The last name particularly derives from the fact that Hedge is equivalent to the instantiation of the Weighted Average Forecaster with the exponential potential function $\Phi_\eta(\mathbf{x}) \coloneqq \frac{1}{\eta}\ln\bigl(\sum_{i\in[K]} \exp(\eta x_i)\bigr)$ for $\mathbf{x} \coloneqq (x_1, \dots, x_K)^\top \in \mathbb{R}^K$ and the same learning rate $\eta > 0$ (e.g., see Cesa-Bianchi and Lugosi (2006, Section 2.1)). The regret of Hedge is $R_T = \Theta(\sqrt{T \ln K})$ for the problem of prediction with expert advice and it is known to be minimax optimal.

Algorithm 2.1: Hedge

```
input: learning rate $\eta > 0$
$w_1(i) \leftarrow 1$ for all $i \in V$
for $t = 1, \dots, T$ do
    Let $W_t \leftarrow \sum_{i\in V} w_t(i)$
    Let $p_t(i) \leftarrow w_t(i)/W_t$ for all $i \in V$
    Select $I_t \sim p_t$
    Incur loss $\ell_t(I_t)$ and observe $(\ell_t(i))_{i\in V}$
    Let $w_{t+1}(i) \leftarrow w_t(i)\exp(-\eta\ell_t(i))$ for all $i \in V$
```

2.1.2 The Multi-Armed Bandit Problem

It is often the case that full information is unavailable and we regardless expect a good agent to sequentially learn. More precisely, many real-world scenarios allow the learner to observe only the loss $\ell_t(I_t)$ of the selected action I_t at round t. This new online learning protocol (Online Protocol 2.2) defines the *multi-armed bandit* problem (Auer, Cesa-Bianchi, Freund, and Schapire, 1995, 2002b, Auer, Cesa-Bianchi, and Fischer, 2002a) and this type of feedback intuitively takes the name of bandit feedback. For instance, some applications with bandit feedback are clinical trials, where the decision-maker selects one among different treatments to administer to patients and can only observe the outcome of the selected one, or online advertising, where a platform chooses ads to show to its users and can track their interactions (e.g., clicks) over the displayed advertisements only.

Online Protocol 2.2: Multi-armed bandit

```
environment: losses $\ell_t \in [0,1]^V$ for all $t \in [T]$
for $t = 1, \dots, T$ do
    The learner selects an action $I_t \in V$ (possibly at random)
    The learner suffers loss $\ell_t(I_t)$
    The learner observes loss $\ell_t(I_t)$
```

The multi-armed bandit problem clearly has a vast range of applications and it is a harder problem than prediction with expert advice because of its limited feedback. Nevertheless, it is possible to design online learning algorithms that achieve a regret guarantee with the same $\sqrt{T}$ dependence on the time horizon T. This result can be obtained by a modification of Hedge. The main difference consists of the change from full information to bandit feedback. Since Hedge necessitates the losses of all actions to run, we replace the true loss $\ell_t(i)$ with an unbiased estimate for it. The most common and famous unbiased estimate is the *importance-weighted estimator* $\widehat{\ell}_t$, defined as

$$\widehat{\ell}_t(i) \coloneqq \frac{\ell_t(i)}{p_t(i)}\mathbb{I}\{I_t = i\} \tag{2.2}$$

for each $i \in V$; note that computing it only requires the knowledge of the observed loss $\ell_t(I_t)$. If we let $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$ be the expectation conditioned on the past actions $I_1, \dots, I_{t-1}$, we can observe that $\widehat{\ell}_t(i)$ is a conditionally unbiased estimator for the loss $\ell_t(i)$ of $i \in V$ because $\mathbb{E}_t[\widehat{\ell}_t(i)] = \ell_t(i)$. Furthermore, we observe that $\mathbb{E}_t[\widehat{\ell}_t(i)^2] = \frac{\ell_t(i)^2}{p_t(i)} \leq \frac{1}{p_t(i)}$, meaning that the importance-weighted estimator for $\ell_t(i)$ has a conditional second moment bounded from above by the inverse of the probability $p_t(i)$ of selecting action i at round t. The algorithm obtained by executing Hedge over the sequence of importance-weighted estimates $\widehat{\ell}_1, \dots, \widehat{\ell}_T$, instead of the true losses, is the Exp3 algorithm (Exponential-weight algorithm for Exploration and Exploitation). Given the understanding of the behavior of Hedge and the properties of importance-weighting, the regret of Exp3 can be shown to be $R_T = \mathcal{O}(\sqrt{KT \ln K})$ for the multi-armed bandit problem, which is only a $\sqrt{K}$ factor worse than the minimax regret[*] in the full-information setting.

2.2 Feedback Models

We now introduce extensions and generalizations of prediction with expert advice and the multi-armed bandit problem. The following online learning frameworks consider more general feedback models that capture naturally occurring scenarios in many common sequential decision-making tasks.

2.2.1 Partial Feedback and Feedback Graphs

The first extension we consider is a more general setting with possibly partial feedback. For any $t \in [T]$, consider a map $S_t \colon V \to 2^V$ that determines the set $S_t(i) \subseteq V$ of actions whose loss is revealed to the learner if it selects action $I_t = i \in V$ at round t. The map S_t can vary over time and is also revealed to the learner. Depending on when the learner receives the information about feedback structure S_t, we can distinguish two settings: the *informed setting*, which happens when the learner knows S_t before performing its decision at round t, and the *uninformed setting*, whenever the learner sees S_t only after selecting the action at round t. Online Protocol 2.3 shows a summary of the interaction protocol for the uninformed partial feedback setting. Observe that this extension generalizes both full information, obtained when $S_1(i) = \cdots = S_T(i) = V$ for every i, and bandit feedback, given by $S_1(i) = \cdots = S_T(i) = \{i\}$ for each i.

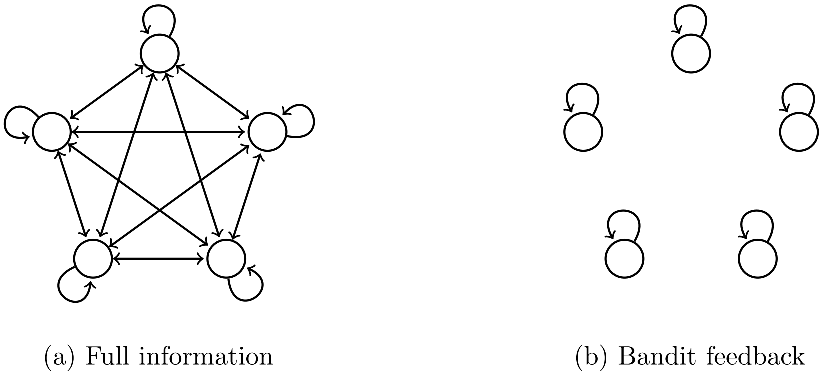
Observe that we can equivalently model this setting by defining directed graphs $G_t = (V, E_t)$ over V, that is, where vertices correspond to the actions V and the edges $E_t \subseteq V \times V$ may change over time. To be exact, we have an edge $(i, j) \in E_t$ for $i, j \in V$ whenever $j \in S_t(i)$. We therefore have that $E_t \coloneqq \{(i, j) \in V \times V : j \in S_t(i)\}$ given S_t. Further notice that the feedback set $S_t(i)$ for $i \in V$

[*] The best achievable worst-case regret guarantee.

Online Protocol 2.3: Online learning with partial feedback

```
environment: losses $\ell_t \in [0,1]^V$ and feedback set $S_t\colon V \to 2^V$, for all $t \in [T]$
for $t = 1, \dots, T$ do
    The learner selects an action $I_t \in V$ (possibly at random)
    The learner incurs loss $\ell_t(I_t)$
    The learner observes losses $\big\{\big(i, \ell_t(i)\big) : i \in S_t\,(I_t)\big\}$ and $S_t$
```

corresponds to its out-neighborhood $N^{\text{out}}_{G_t}(i)$ in the graph G_t, where $N^{\text{out}}_G(i) \coloneqq \{j \in V : (i,j) \in E\}$ for any graph $G = (V,E)$. Going back to the standard settings of full information and bandit feedback, the respective resulting graphs are shown in Figure 2.1.


(a) Full information

(b) Bandit feedback

Figure 2.1: Examples of feedback graphs for the full feedback and the bandit feedback settings.

This way of modeling the problem with partial feedback setting is called *online learning with feedback graphs*, introduced by Mannor and Shamir (2011) and extensively studied by subsequent work (Alon, Cesa-Bianchi, Gentile, and Mansour, 2013, Cohen, Hazan, and Koren, 2016, Alon, Cesa-Bianchi, Gentile, Mannor, Mansour, and Shamir, 2017, Chen, Huang, Li, and Zhang, 2021). We particularly mention Alon, Cesa-Bianchi, Dekel, and Koren (2015) who determined a trichotomy of regret regimes depending solely on structural properties of feedback graphs. Assume for simplicity that the feedback graph $G = (V,E)$ is fixed and given to the learner. Alon et al. (2015) define different notions of observability of actions depending on their in-neighborhood $N^{\text{in}}_G(i) \coloneqq \{j \in V : (j,i) \in E\}$, i.e., the set of actions that can observe the loss of $i \in V$.

Definition 2.1 (Observability)**.** *A vertex $i \in V$ is* observable *in G if $N^{\text{in}}_G(i) \neq \emptyset$. An observable vertex i is* strongly observable *when $i \in N^{\text{in}}_G(i)$ (i.e., G contains a self-loop over i) or $V \setminus \{i\} \subseteq N^{\text{in}}_G(i)$, and it is* weakly observable *otherwise.*

We can extend the same notions to the entire feedback graph G as follows.

Definition 2.2 (Graph observability)**.** *A graph $G = (V,E)$ is observable if every vertex in V is observable. An observable graph G is strongly observable if every vertex in V is strongly observable, and it is weakly observable otherwise.*

What Alon et al. (2015) demonstrate is that the overall dependence of the minimax regret on the number of rounds T, modulo logarithmic factors, essentially boils down to the structure of G. The authors even state a more granular characterization of the regret: the two classes of graphs in Definition 2.2 are both learnable and depend on different graph-theoretic parameters. The graph-theoretic quantity relating to strongly observable graphs is the independence number.

Definition 2.3 (Independence). *A subset $S \subseteq V$ of vertices in a graph $G = (V, E)$ is an* independent set *if $(i, j) \notin E$ for any two $i, j \in S$ such that $i \neq j$. The* independence number $\alpha(G)$ *of G is the size of a largest independent set in G.*

The graph parameter for weakly observable graphs is instead the weak domination number.

Definition 2.4 (Weak domination). *A subset $D \subseteq V$ of vertices in a graph $G = (V, E)$ is a* dominating set *for $U \subseteq V$, for any $i \in U$, there exists $j \in D$ such that $i \in N_G^{\text{out}}(j)$; alternatively, it satisfies $U \subseteq \bigcup_{i \in D} N_G^{\text{out}}(i)$. The* weak domination number $\delta(G)$ *of G is the size of a smallest dominating set in G for the set $W \subseteq V$ of weakly observable vertices.*

With these notions in mind, the above-mentioned trichotomy for the minimax regret of online learning with feedback graphs states what follows.

Proposition 2.1 (Alon et al. (2015, Theorem 1)). *Let $G = (V, E)$ be a feedback graph with $|V| \geq 2$. If T is sufficiently large, the minimax regret for online learning with feedback graph G is*

- $R_T = \widetilde{\Theta}\big(\sqrt{\alpha(G)T}\big)$ *if G is strongly observable;*
- $R_T = \widetilde{\Theta}\left(\delta(G)^{1/3}T^{2/3}\right)$ *if G is weakly observable;*
- $R_T = \Theta(T)$ *if G is not observable.*

Not only does the observability structure of the feedback graph impact the dependence of the regret on the time horizon T, but a specific graph parameter impacts it as well and it differs between the two learnable settings. We also remark that the more general partial monitoring problem (e.g., see Cesa-Bianchi and Lugosi (2006, Section 6.4)) has a trichotomy that resembles the one stated above and depends on observability properties too (Antos, Bartók, Pál, and Szepesvári, 2013, Bartók, Foster, Pál, Rakhlin, and Szepesvári, 2014). It is interesting to observe that the above characterization evinces a similar aspect in the simpler problem of multi-armed bandits with feedback graphs.

In Chapter 3, we further investigate the minimax regret for the problem of online learning with undirected strongly observable feedback graphs, providing improved upper and lower regret bounds. Moreover, in Chapter 5, we consider a further extension of the feedback graphs model to the case of stochastic feedback graphs, i.e., random graphs where, at every round, each edge realizes independently with some fixed but unknown probability.

2.2.2 Bandit Feedback with Expert Advice

Compared to standard multi-armed bandits, there are situations in which the learner obtains additional information, possibly related to the losses, by external sources. For example, in recommendation systems we may employ multiple recommendation policies, each using contextual information in different ways, as multiple experts and use their predictions to pick a single action. One may think of portfolio management as another example, where we might have many investment strategies at our disposal and we can leverage this side information to better select assets to allocate. In both cases, there are many experts that provide some advice to the learner, who then selects an action and only observes its loss.

This problem is a variant of bandits named *multi-armed bandit with expert advice* (Auer et al., 1995, 2002b). Compared to standard bandits, there are N experts providing advice to the learner at the beginning of each round. The advice at round t is formally defined as a collection $(\theta_t^j)_{j\in[N]}$ of distributions $\theta_t^j \in \Delta_V$ over actions in V, each associated to some expert $j \in [N]$. The learner then selects action I_t by exploiting the expert advice and receives bandit feedback, meaning that it only observes $\ell_t(I_t)$ as in the standard version of the problem. Online Protocol 2.4 provides a summary of the interaction protocol for this problem.

Online Protocol 2.4: Online learning with bandit feedback and expert advice

```
**environment:** losses $\ell_t \in [0,1]^V$ and expert advice $\theta_t^j \in \Delta_V$, for all $t \in [T]$ and all $j \in [N]$
**for** $t = 1, \dots, T$ **do**
    The learner receives expert advice $(\theta_t^j)_{j\in[N]}$
    The learner selects an action $I_t \in V$ (possibly at random)
    The learner incurs loss $\ell_t(I_t)$
    The learner observes loss $\ell_t(I_t)$
```

The notion of regret differs relative to what we considered so far. The main change is due to comparing to the best fixed expert in hindsight rather than the best fixed action. In particular, we first observe that the expected loss of an expert $j \in [N]$ at round t corresponds to $\langle \ell_t, \theta_t^j \rangle$ because θ_t^j is essentially the distribution that expert j would use to select an action to play at round t. We therefore have that the expected regret for bandits with expert advice becomes

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \ell_t(I_t)\right] - \min_{j\in[N]} \sum_{t=1}^T \langle \ell_t, \theta_t^j \rangle \; . \tag{2.3}$$

An algorithm obtained as a variant of Exp3, called Exp4 (Auer et al., 1995, 2002b), that adapts to the presence of expert advice is known to achieve regret $R_T = \mathcal{O}\big(\sqrt{KT \ln N}\big)$ when $N > K$, while we can reduce to standard bandits to show that the minimax regret is $R_T = \Theta\big(\sqrt{NT}\big)$ when $N \le K$. Starting from its design, the Exp4 algorithm has become an important baseline or building block for addressing many related problems; for example, sleeping bandits (Kleinberg, Niculescu-Mizil, and Sharma, 2010), online multi-class classification (Daniely and Helbertal, 2013), online non-parametric learning (Cesa-Bianchi, Gaillard, Gentile, and Gerchinovitz, 2017), and non-stationary bandits (Luo, Wei, Agarwal, and Langford, 2018). Assuming the case $N > K$, a more tailored algorithm is shown to achieve an improved regret bound $R_T = \mathcal{O}\big(\sqrt{KT \ln(N/K)}\big)$ (Kale, 2014) and a lower bound of order $\Omega\big(\sqrt{KT(\ln N)/(\ln K)}\big)$ is known to hold (Seldin and Lugosi, 2016).

In Chapter 4, with a spirit analogous to that of Chapter 3 for the feedback graphs model, we derive improved regret upper bounds for bandits with expert advice by adopting similar techniques. We additionally demonstrate an improved lower bound on the minimax regret under a slightly more restricted feedback.

2.2.3 Delayed Bandit Feedback

In some applications, it is unrealistic to assume that the learner is able to receive the feedback immediately after performing an action. Hence, the decision-maker might have to repeatedly perform decisions without immediately observing the outcome of these choices. The loss $\ell_t(I_t)$ of the action I_t chosen at any round t is instead delayed to a subsequent round $t + d_t$ with a fixed but initially unknown delay $d_t \in \mathbb{N}$; without loss of generality, we may assume that $d_t \leq T - t$ for each $t \in [T]$. This additional extension of bandits takes the name of *multi-armed bandit with delayed feedback* (Joulani, György, and Szepesvári, 2013, Cesa-Bianchi, Gentile, Mansour, and Minora, 2016, Cesa-Bianchi, Gentile, and Mansour, 2019, Zimmert and Seldin, 2020, Masoudian, Zimmert, and Seldin, 2022, 2023)—see Online Protocol 2.5 for a summary—and it arises in many realistic scenarios, such as online advertising, medical trials, or financial investments. To be clear, in online advertising the outcome can consist of the click-through rate or the number of conversions, which are often delayed as they depend on user interactions. Even in medical trials we expect that the effects of an administered drug or a prescribed treatment, which can include the success of the treatment or possible side-effects, will be observed (or ruled out) after an initially unknown amount of time.

Online Protocol 2.5: Online learning with delayed bandit feedback

```
environment: losses $\ell_t \in [0,1]^V$ and delays $d_t \leq T - t$, for all $t \in [T]$
for $t = 1, \dots, T$ do
    The learner selects an action $I_t \in V$ (possibly at random)
    The learner incurs loss $\ell_t(I_t)$                    ▷ observed at round $t + d_t$
    The learner observes losses $\big\{\big(s, \ell_s(I_s)\big) : s \leq t, s + d_s = t\big\}$
```

It is known that we can design online learning algorithms achieving regret $R_T = \mathcal{O}\big(\sqrt{KT} + \sqrt{D \ln K}\big)$ where $D = \sum_{t=1}^{T} d_t$ is the total delay. In the particular setting with fixed delay, i.e., when $d_1 = \cdots = d_T = d$ for some fixed d, we even know that the minimax regret (Cesa-Bianchi et al., 2016, 2019, Zimmert and Seldin, 2020) has rate $R_T = \Theta\big(\sqrt{KT} + \sqrt{dT \ln K}\big)$. These results immediately illustrate that the presence of delays negatively affects the regret in an unavoidable way. Although there exist techniques to avoid the significant impact of the few largest delays in the case of non-uniform delays (Thune, Cesa-Bianchi, and Seldin, 2019), a relevant fraction of the total delay will nonetheless influence the performance of any online learner in the worst case.

In Chapter 6, we study a variant of the delayed bandit feedback model in which the learner is immediately provided an intermediate observation between their decision at round t and the arrival of the loss at round $t + d_t$. The intermediate observation is a signal associated to each action at the current round, and the loss is a function of the type of signal only. The main objective is to understand when intermediate observations help reduce the effect of the total delay on the regret.

Chapter 3

On the Minimax Regret for Online Learning with Feedback Graphs

We consider the problem of online learning with undirected strongly observable feedback graphs. The best known upper bound for the minimax regret in this problem has order $\sqrt{\alpha T \ln K}$, where K is the number of actions, α is the independence number of the graph, and T is the time horizon. The $\sqrt{\ln K}$ factor is known to be necessary for prediction with experts advice ($\alpha = 1$). On the other hand, in multi-armed bandits ($\alpha = K$) the minimax rate is known to be $\sqrt{KT}$, and a lower bound of $\sqrt{\alpha T}$ is known to hold for any α. We provide an improved regret upper bound of $\sqrt{\alpha T(1 + \ln(K/\alpha))}$ for any α, which matches the lower bounds for bandits and experts, interpolating intermediate cases at the same time. We complement our regret guarantee with an improved $\sqrt{\alpha T(\ln K)/(\ln \alpha)}$ lower bound for all $\alpha > 1$, which shows that a logarithmic factor is necessary as soon as $\alpha < K$.

3.1 Introduction

Feedback graphs provide an elegant interpolation between two popular online learning models: multi-armed bandit and prediction with expert advice. When learning with an undirected feedback graph G over K actions, the online algorithm observes not only the loss of the action chosen in each round, but also the loss of the actions that are adjacent to it in the graph. As already discussed in Chapter 2, the two aforementioned models are special cases: prediction with expert advice is equivalent to the case when G is a clique, whereas K-armed bandits is equivalently modeled by a feedback graph G containing only self-loops. When losses are generated adversarially, the regret in the feedback graph setting with strong observability (see Definition 2.2 from Chapter 2 for a formal description) has been shown to scale with the independence number $\alpha \coloneqq \alpha(G)$ of G—see Proposition 2.1 from Chapter 2, which more generally considers arbitrary directed graphs. Intuitively, denser graphs, which correspond to smaller independence numbers, provide more feedback to the learner, thus enabling a better control on regret. More specifically, the best known upper and lower bounds on the regret after T rounds are $\mathcal{O}\big(\sqrt{\alpha T \ln K}\big)$ and $\Omega\big(\sqrt{\alpha T}\big)$ (Alon et al., 2013, 2017). It has been known for three decades that this upper bound is tight for $\alpha = 1$ (the experts case; see Cesa-Bianchi et al. (1993, 1997)). When $\alpha = K$ (the bandits case), the lower bound $\Omega\big(\sqrt{KT}\big)$—which has also been known for nearly three decades (Auer et al., 1995, 2002b)—was matched by a corresponding upper bound $\mathcal{O}\big(\sqrt{KT}\big)$ only by Audibert and Bubeck (2009). These results show that in feedback graphs, the logarithmic factor $\sqrt{\ln K}$ is necessary (at least) for the $\alpha = 1$ case, while it must vanish from the minimax regret as α grows from 1 to K, but the current bounds fail to capture this fact. In this chapter, we prove new upper and lower regret bounds that for the first time account for this vanishing logarithmic factor.

To prove our new upper bound, we use the standard Follow The Regularized Leader (FTRL) algorithm run with the q-Tsallis entropy regularizer (here q-FTRL for short). It is well-known (Abernethy, Lee, and Tewari, 2015) that for $q = \frac{1}{2}$ this algorithm (run with appropriate loss estimates) achieves regret $\mathcal{O}\big(\sqrt{KT}\big)$ when $\alpha = K$, while for $q \to 1^-$ the same algorithm (with full information of the losses) recovers the bound $\mathcal{O}\big(\sqrt{T \ln K}\big)$ when $\alpha = 1$. When G contains all self-loops, we show in Theorem 3.1 that, if q is chosen as a certain function $q(\alpha, K)$ of both α and K, then $q(\alpha, K)$-FTRL, run with standard importance-weighted loss estimates, achieves regret $\mathcal{O}\big(\sqrt{\alpha T(1 + \ln(K/\alpha))}\big)$. This is a strict improvement over the previous bound, and matches the lower bounds for bandits and experts while interpolating the intermediate cases. This interpolation is reflected by our choice of q, which goes from $\frac{1}{2}$ to 1 as α ranges from 1 to K. The main technical hurdle in proving this result is an extension to arbitrary values of $q \in \big[\frac{1}{2}, 1\big)$ of a standard result—see, e.g., Mannor and Shamir (2011, Lemma 3)—that bounds the variance term from the regret of q-FTRL in terms of α. In Theorem 3.2, using a modified loss estimate, this result is extended to any undirected strongly observable graph (Definition 2.2), a class of feedback graphs in which some of the actions might not reveal their loss when played. In Theorem 3.3, we show via a doubling trick that our new upper bound can also be obtained (up to constant factors) without the need of knowing (or computing) α. As the resulting algorithm is oblivious to α, our analysis also applies to arbitrary sequences of graphs $(G_t)_{t\in[T]}$, where K is constant but the independence number α_t of G_t can change over time, and the algorithm observes G_t only after choosing an action (the so-called uninformed case). In this setting, the analysis of the doubling trick is complicated by the non-trivial dependence of the regret on the sequence of α_t.

We also improve on the $\Omega\big(\sqrt{\alpha T}\big)$ lower bound by proving a new $\Omega\big(\sqrt{\alpha T \log_\alpha K}\big)$ lower bound for all $\alpha > 1$. This is the first result showing the necessity—outside the experts case—of a logarithmic factor in the minimax regret for all $\alpha < K$. Our proof uses a stochastic adversary generating both losses and feedback graphs via i.i.d. draws from a joint distribution. This sequence of losses and feedback graphs can be used to define a hard instance of the multitask bandits problem, a variant of the combinatorial bandits framework (Cesa-Bianchi and Lugosi, 2012). We then prove our result by adapting known lower bounding techniques for multitask bandits (Audibert, Bubeck, and Lugosi, 2014). Note that for values of α bounded away from 2 and K, the logarithmic factor $\log_\alpha K$ in the lower bound is smaller than the corresponding factor $1 + \ln(K/\alpha)$ in the upper bound—this is further discussed at the end of this chapter.

3.1.1 Related Work

Several previous works have used the q-Tsallis regularizer with q tuned to specific values other than $\frac{1}{2}$ and 1. For example, in Zimmert and Lattimore (2019, Section 4), q is chosen as a function of K to prove a regret bound of $\mathcal{O}\big(\sqrt{\alpha T \ln^3 K}\big)$ for any strongly observable directed feedback graph, which shaves off a $\ln T$ factor compared to previous works. This bound is worse than the corresponding bounds for undirected graphs because the directed setting is harder. Specific choices of q have been considered to improve the regret in settings of online learning with standard bandit feedback. For example, the choice $q = \frac{2}{3}$ was used by Rouyer and Seldin (2020) to improve the analysis of regret in bandits with decoupled exploration and exploitation. Regret bounds for arbitrary choices of q are derived by Zimmert and Seldin (2021), Jin, Liu, and Luo (2023) for a best-of-both-worlds analysis of bandits, though $q = \frac{1}{2}$ remains the optimal choice. The $\frac{1}{2}$-Tsallis entropy and the Shannon entropy ($q \approx 1$) regularizers have been combined before in different ways to obtain best-of-both-worlds guarantees for the graph feedback problem (Erez and Koren, 2021, Ito, Tsuchiya, and Honda, 2022). The idea of using values of $q \in (\frac{1}{2}, 1)$ for feedback graphs is quite natural and has been brought up before, e.g., in Rouyer, Van der Hoeven, Cesa-Bianchi, and Seldin (2022a), but achieving an improved dependence on the graph structure by picking a suitable value of q has not been, to the best of our knowledge, successfully pursued before.

On the other hand, the q-FTRL algorithm adopted in the current chapter is essentially equivalent to the PolyINF algorithm (Audibert, Bubeck, and Lugosi, 2011, Abernethy et al., 2015), which was used by Kale (2014) to achieve the best known worst-case upper bound. We indeed build on top of the intuition from Kale (2014), complementing it with our novel bound on the variance term of FTRL with the negative q-Tsallis entropy regularizer. An approach based on a similar use of the q-Tsallis regularizer has also been employed by Kwon and Perchet (2016) for the problem of multi-armed bandits with sparse losses to achieve a $\mathcal{O}\big(\sqrt{sT \ln(K/s)}\big)$ regret bound, where s is the maximum number of nonzero losses at any round. After deriving our results, a subsequent work by Ito (2024) applied the same principle to derive improved regret bounds for contextual linear bandits.

Our lower bound is reminiscent of the $\Omega\big(\sqrt{KT \log_K N}\big)$ lower bound proved in Seldin and Lugosi (2016) for the problem of bandits with expert advice (with $N \geq K$ being the number of experts) described in Chapter 2; for further details on this lower bound, see Eldowa, Cesa-Bianchi, Metelli, and Restelli (2023a) and Vural, Gokcesu, Gokcesu, and Kozat (2019). Although the two settings are different, the variant of the multitask bandit problem that our lower bound construction simulates is the same as the one used in the proof of Eldowa et al. (2023a, Theorem 7).

3.2 Problem Setting

We briefly recall the problem setting for online learning with feedback graphs from Section 2.2.1 of Chapter 2, here specialized to the case of undirected graphs. We consider the following game played over T rounds between a learner with action set $V = [K]$ and the environment. At the beginning of the game, the environment secretly selects a sequence of $[0,1]$-valued losses $(\ell_t)_{t\in[T]}$, and a sequence of undirected graphs $(G_t)_{t\in[T]}$ over the set of actions V, that is, $G_t \coloneqq (V, E_t)$. At any time t, the learner selects an action I_t (possibly at random), then pays loss $\ell_t(I_t)$ and observes the feedback graph G_t and all losses $\ell_t(i)$ of neighbouring actions $i \in N_{G_t}(I_t)$, where $N_{G_t}(i) \coloneqq \{j \in V : (i,j) \in E_t\}$—see Online Protocol 3.1, which is a reformulation of Online Protocol 2.3 for undirected feedback graphs. As already made clear in Chapter 2, the performance of the learner is measured by the regret, which we restate here for clarity:

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \min_{i\in[K]} \sum_{t=1}^{T} \ell_t(i) ,$$

where the expectation is over the internal randomization of the learner.

For simplicity, form now onwards we use N_t to denote the neighbourhood N_{G_t} in the graph G_t and we use α_t to denote the independence number $\alpha(G_t)$ of G_t at time t. In the current chapter, we

Online Protocol 3.1: Online learning with undirected feedback graphs

```
**environment:** losses $\ell_t \in [0,1]^V$ and undirected graphs $G_t = (V, E_t)$, for all $t \in [T]$
for $t = 1, \dots, T$ do
    The learner picks an action $I_t \in V$ (possibly at random)
    The learner incurs loss $\ell_t(I_t)$
    The learner observes losses $\big\{(i, \ell_t(i)) : i \in N_{G_t}(I_t)\big\}$ and graph $G_t$
```

only focus on strongly observable graphs, whereas in the subsequent Chapter 5 we shift towards arbitrary observability structures in a variation of the problem involving stochastic feedback graphs.

3.3 FTRL with Tsallis Entropy for Undirected Feedback Graphs

As a building block, in this section, we focus on the case when all the feedback graphs $G_1, \dots, G_T$ have the same independence number $\alpha_1 = \cdots = \alpha_T = \alpha$, whereas the general case is treated in the next section. For simplicity, we start with the assumption that all nodes have self-loops: $(i,i) \in E_t$ for all $i \in V$ and all t. We later lift this requirement and show that the regret guarantees that we provide can be extended to general undirected strongly observable feedback graphs, only at the cost of a constant multiplicative factor.

The algorithm we analyze is q-FTRL (described in Algorithm 3.1), which is an instance of the Follow The Regularized Leader (FTRL) framework—see, e.g., Orabona (2019, Chapter 7)—with the (negative) q-Tsallis entropy

$$\psi_q(x) \coloneqq \frac{1}{1-q}\left(1 - \sum_{i \in V} x(i)^q\right) \qquad \forall x \in \Delta_K \ ,$$

as the regularizer, whose parameter $q \in (0,1)$ can be tuned according to our needs. Since we do not observe all the losses in a given round, the algorithm makes use of unbiased estimates for the losses. When all self-loops are present, we define the estimated losses in the following standard manner. Let I_t be the action picked at round t, which is drawn from the distribution $p_t \in \Delta_K$ maintained by the algorithm; the loss estimate for an action $i \in V$ at round t is given by

$$\widehat{\ell}_t(i) \coloneqq \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\{I_t \in N_t(i)\} \ , \tag{3.1}$$

where $P_t(i) = \mathbb{P}\big(I_t \in N_t(i)\big) = \sum_{j \in N_t(i)} p_t(j)$ is the probability of observing action i. This estimate is an adaptation of the importance-weighted estimate (Equation (2.2) from Chapter 2) to the undirected feedback graphs case, which is also conditionally unbiased in the sense that $\mathbb{E}_t\big[\widehat{\ell}_t(i)\big] = \ell_t(i)$ for all $t \in [T]$ and all $i \in V$, where we denote $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$.

A key part of the standard regret analysis of q-FTRL (see, e.g., the proof of Lemma A.2 in Appendix A.1) is handling the variance term, which, with the choice of estimator given by Equation (3.1), takes the form

$$B_t(q) \coloneqq \sum_{i \in V} \frac{p_t(i)^{2-q}}{P_t(i)} \ . \tag{3.2}$$

Algorithm 3.1: q-FTRL for undirected feedback graphs

```
input: $q \in (0,1)$, $\eta > 0$
initialization: $p_1(i) \leftarrow 1/K$ for all $i = 1, \dots, K$
for $t = 1, \dots, T$ do
    Select action $I_t \sim p_t$ and incur loss $\ell_t(I_t)$
    Observe losses $\big\{\big(i, \ell_t(i)\big) : i \in N_t(I_t)\big\}$ and graph $G_t$
    Construct a loss estimate $\widehat{\ell}_t$ for $\ell_t$          $\triangleright$ e.g., (3.1) or (3.6)
    Let $p_{t+1} \leftarrow \arg\min_{p \in \Delta_K} \eta \langle \sum_{s=1}^{t} \widehat{\ell}_s, p \rangle + \psi_q(p)$
```

By Hölder's inequality, this term can be immediately upper bounded by

$$B_t(q) \le \sum_{i \in V} p_t(i)^{1-q} \le \left(\sum_{i \in V} p_t(i) \right)^{1-q} \left(\sum_{i \in V} 1^{1/q} \right)^{q} = K^q \ ,$$

while previous results on the regret analysis of multi-armed bandits with graph feedback (Mannor and Shamir, 2011, Alon et al., 2017) would give

$$B_t(q) \le \sum_{i \in V} \frac{p_t(i)}{P_t(i)} \le \alpha \ .$$

However, the former result would only recover a $\mathcal{O}\big(\sqrt{KT}\big)$ regret bound (regardless of α) with the best choice of $q = 1/2$, which could be trivially achieved by ignoring side-observations of the losses, whereas the latter bound would only manage to achieve a $\mathcal{O}\big(\sqrt{\alpha T \ln K}\big)$ regret bound, incurring the extra $\sqrt{\ln K}$ factor for all values of α. Other results in the literature (e.g., see Alon et al. (2015, 2013), Dann, Wei, and Zimmert (2023), Ito et al. (2022), Kocák, Neu, Valko, and Munos (2014), Rouyer et al. (2022a), Zimmert and Lattimore (2019)) do not bring an improvement in this setting when bounding the $B_t(q)$ term and, hence, do not suffice for achieving the desired regret bound. The following lemma provides a novel and improved bound on quantities of the same form as $B_t(q)$ in terms of the independence number $\alpha_t = \alpha$ of the undirected graph G_t.

Lemma 3.1. *Let $G = (V, E)$ be any undirected graph with $|V| = K$ vertices and independence number $\alpha(G) = \alpha$. Let $b \in [0, 1]$, $p \in \Delta_K$, and consider any nonempty subset $U \subseteq \big\{v \in V : v \in N_G(v)\big\}$. Then,*

$$\sum_{v \in U} \frac{p(v)^{1+b}}{\sum_{u \in N_G(v)} p(u)} \le \alpha^{1-b} \ .$$

Proof. First of all, observe that we can restrict ourselves to the subgraph $G[U]$ induced by U, i.e., the graph $G[U] \coloneqq (U, E \cap (U \times U))$. This is because the neighbourhoods in this graph are such that $N_{G[U]}(v) \subseteq N_G(v)$ for all $v \in U$, and its independence number is $\alpha(G[U]) \le \alpha(G)$. Hence, it suffices to prove the claimed inequality for any undirected graph $G = (V, E)$ with all self-loops, any $p \in [0, 1]^K$ such that $\|p\|_1 \le 1$, and the choice $U = V$. We assume this in what follows without loss of generality.

For any subgraph $H \subseteq G$ with vertices $V(H) \subseteq V$, denote the quantity we want to bound from above as

$$Q(H) \coloneqq \sum_{v \in V(H)} \frac{p(v)^{1+b}}{\sum_{u \in N_G(v)} p(u)} \ .$$

Our aim is thus to provide an upper bound to $Q(G)$.

Consider a greedy algorithm that incrementally constructs a subset of vertices in the following way: at each step, it selects a vertex v that maximizes $p(v)^b/\bigl(\sum_{u\in N_G(v)} p(u)\bigr)$, it adds v to the solution, and it removes v from G together with its neighbourhood $N_G(v)$. This step is iterated on the remaining graph until no vertex is left.

Let $S \coloneqq \{v_1, \dots, v_s\} \subseteq V$ be the solution returned by the above greedy algorithm on G. Also let $G_1, \dots, G_{s+1}$ be the sequence of graphs induced by the operations of the algorithm, where $G_1 = G$ and G_{s+1} is the empty graph, and let $N_r(v) = N_{G_r}(v)$ for $v \in V(G_r)$. At every step $r \in [s]$ of the greedy algorithm, the contribution to $Q(G)$ of the removed vertices $N_r(v_r)$ amounts to

$$\begin{aligned} Q(G_r) - Q(G_{r+1}) = \sum_{v\in N_r(v_r)} \frac{p(v)^{1+b}}{\sum_{u\in N_1(v)} p(u)} &\le \sum_{v\in N_r(v_r)} p(v) \frac{p(v_r)^b}{\sum_{u\in N_1(v_r)} p(u)} \\ &\le \frac{\sum_{v\in N_1(v_r)} p(v)}{\sum_{u\in N_1(v_r)} p(u)} p(v_r)^b = p(v_r)^b \ , \end{aligned}$$

where the last inequality is due to the fact that $N_i(v) \subseteq N_j(v)$ for all $i \ge j$ and $v \in V_i$. Therefore, we can observe that

$$Q(G) = \sum_{r=1}^{s} \bigl(Q(G_r) - Q(G_{r+1})\bigr) \le \sum_{v\in S} p(v)^b \ .$$

The solution S is an independent set of G by construction. Consider now any independent set $A \subseteq V$ of G. We have that

$$\begin{aligned} \sum_{v\in A} p(v)^b \le \max_{x\in\Delta_K} \sum_{v\in A} x(v)^b &= |A| \max_{x\in\Delta_K} \sum_{v\in A} \frac{x(v)^b}{|A|} \\ &\le |A| \max_{x\in\Delta_K} \left(\frac{1}{|A|} \sum_{v\in A} x(v) \right)^b \le |A|^{1-b} \le \alpha^{1-b} \ , \end{aligned} \tag{3.3}$$

where the second inequality follows by Jensen's inequality and the fact that $b \in [0, 1]$. □

Observe that this upper bound is tight for general probability distributions $p \in \Delta_K$ over the vertices V of any undirected strongly observable graph G (containing at least one self-loop), as it is exactly achieved by the distribution $p^\star \in \Delta_K$ defined as $p^\star(i) \coloneqq \frac{1}{|S|}\mathbb{I}\{i \in S\}$ for some maximum independent set $S \subseteq V$ of G. Using this lemma, the following theorem provides our improved upper bound under the simplifying assumptions we made thus far.

Theorem 3.1. *Let $G_1, \dots, G_T$ be a sequence of undirected feedback graphs, where each G_t contains all self-loops and has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 3.1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln(K/\alpha)^2 + 4} + 2}\right) \in [1/2, 1) \qquad \text{and} \qquad \eta = \sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}} \ ,$$

and loss estimates (3.1)*, then its regret satisfies $R_T \le 2\sqrt{e\alpha T\,(2 + \ln(K/\alpha))}$.*

Proof. One can verify that for any $i \in V$, the loss estimate $\widehat{\ell}_t(i)$ defined in Equation (3.1) satisfies $\mathbb{E}_t\bigl[\widehat{\ell}_t(i)^2\bigr] \le 1/P_t(i)$. Hence, using also the fact that $\mathbb{E}_t\bigl[\widehat{\ell}_t(i)\bigr] = \ell_t(i)$, Lemma A.1 in Appendix A.1

implies that

$$R_T \le \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{i\in V}\frac{p_t(i)^{2-q}}{P_t(i)}\right] \tag{3.4}$$

$$\le \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q}\alpha^q T\,, \tag{3.5}$$

where the second inequality follows by Lemma 3.1 with $b = 1-q$ since all actions $i \in V$ are such that $i \in N_G(i)$. Our choices for q and η allow us to further upper bound the right-hand side of Equation (3.5) by

$$\begin{aligned}\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}T} &= \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\alpha K)-\frac{1}{2}\sqrt{\ln^2(K/\alpha)+4}\right)\left(2+\sqrt{\ln^2(K/\alpha)+4}\right)}\\ &\le \sqrt{2e\alpha T\left(2+\sqrt{\ln^2(K/\alpha)+4}\right)}\\ &\le 2\sqrt{e\alpha T\sqrt{\ln^2(K/\alpha)+4}}\\ &\le 2\sqrt{e\alpha T\,(2+\ln(K/\alpha))}\,.\end{aligned}$$

□

The regret bound achieved in the above theorem achieves the optimal regret bound for the experts setting (i.e., $\alpha = 1$) and the bandits setting (i.e., $\alpha = K$) simultaneously. Moreover, it interpolates the intermediate cases for α ranging between 1 and K, introducing the multiplicative logarithmic factor only for graphs with independence number strictly smaller than K. We remark that the chosen values of q and η do in fact minimize the right-hand side of Equation (3.5). Note that we relied on the knowledge of α to tune the parameter q. This is undesirable in general because computing α is NP-hard and it is even hard to approximate. We will show how to lift this requirement in Section 3.4. The same comment applies to Theorem 3.2, below.

We now illustrate how to achieve the improved regret bound of Theorem 3.1 in the case of undirected strongly observable feedback graphs where some self-loops may be missing; i.e., there may be actions $i \in V$ such that $i \notin N_G(i)$. Using the loss estimator defined in Equation (3.1) may lead to a large variance term due to the presence of actions without self-loops. One approach to deal with this—see, e.g., Zimmert and Lattimore (2019) or Luo, Tong, Zhang, and Zhang (2023)—is to suitably alter the loss estimates of these actions.

Define $S_t \coloneqq \{i \in V : i \notin N_t(i)\}$ as the subset of actions without self-loops in the feedback graph G_t at each time step $t \in [T]$. The idea is that we need to carefully handle some action $i \in S_t$ only in the case when the probability $p_t(i)$ of choosing i at round t is sufficiently large, say, larger than $1/2$. Define the set of such actions as $J_t \coloneqq \{i \in S_t : p_t(i) > 1/2\}$ and observe that $|J_t| \le 1$. Similarly to Zimmert and Lattimore (2019), define new loss estimates

$$\widehat{\ell}_t(i) \coloneqq \begin{cases}\frac{\ell_t(i)}{P_t(i)}\mathbb{I}\{I_t \in N_t(i)\} & \text{if } i \in V\setminus J_t\\ \frac{\ell_t(i)-1}{P_t(i)}\mathbb{I}\{I_t \in N_t(i)\}+1 & \text{if } i \in J_t\end{cases} \tag{3.6}$$

for which it still holds that $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$ and that $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$ for all $i \notin J_t$. This change, along with the use of Lemma 3.1 for the actions in $V \setminus S_t$, suffices in order to prove the following regret bound (see Appendix A.2 for the proof) when the feedback graphs do not necessarily contain self-loops for all actions.

Theorem 3.2. *Let $G_1, \dots, G_T$ be a sequence of undirected strongly observable feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 3.1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln^2(K/\alpha) + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}} \,,$$

and loss estimates (3.6)*, then its regret satisfies $R_T \le 6\sqrt{e\alpha T\,(2 + \ln(K/\alpha))}$.*

3.4 Adapting to Arbitrary Sequences of Graphs

In the previous section, we assumed for simplicity that all the graphs have the same independence number. This independence number was then used to tune q, the parameter of the Tsallis entropy regularizer used by the algorithm. In this section, we show how to extend our approach to the case when the independence numbers of the graphs are neither the same nor known a-priori by the learner. Had these independence numbers been known a-priori, one approach is to set q as in Theorem 3.2, but instead using the average independence number

$$\bar{\alpha}_T \coloneqq \frac{1}{T}\sum_{t=1}^{T} \alpha_t \,.$$

Doing so would allow us to achieve a $\mathcal{O}\Big(\sqrt{\sum_{t=1}^{T} \alpha_t (1 + \ln(K/\bar{\alpha}_T))}\Big)$ regret bound. We now show that we can still recover a bound of the same order without prior knowledge of $\bar{\alpha}_T$. For round t and any fixed $q \in [0,1]$, define

$$H_t(q) \coloneqq \sum_{i \in V \setminus S_t} \frac{p_t(i)^{2-q}}{P_t(i)} \,.$$

We know from Lemma 3.1 that $H_t(q) \le \alpha_t^q$. Thus, we can leverage these observations and use a doubling trick (similar in principle to Alon et al. (2017)) to guess the value of $\bar{\alpha}_T$. This approach is outlined in Algorithm 3.2. Starting with $r = 0$ and $T_r = 1$, the idea is to instantiate Algorithm 3.1 at time-step T_r with q and η set as in Theorem 3.2 but with 2^r replacing the independence number. Then, at $t \ge T_r$, we increment r and restart Algorithm 3.1 only if

$$\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1},$$

since (again thanks to Lemma 3.1) the left-hand side of the above inequality is always bounded form above by $\bar{\alpha}_T$. The following theorem shows that this approach essentially enjoys the same regret bound of Theorem 3.2 up to an additive $\log_2 \bar{\alpha}_T$ term.

Algorithm 3.2: q-FTRL for an arbitrary sequence of undirected strongly observable graphs

```
input: Time horizon $T$
define: For each $r \in \{0, \dots, \lfloor \log_2 K \rfloor\}$,
    $q_r = \frac{1}{2}\left(1 + \frac{\ln(K/2^r)}{\sqrt{\ln^2(K/2^r) + 4} + 2}\right)$   and   $\eta_r = \sqrt{\frac{2 q_r K^{1-q_r}}{11T(1-q_r)\left(2^r\right)^{q_r}}}$
initialization: $T_0 \leftarrow 1$, $r \leftarrow 0$, instantiate Algorithm 3.1 with $q = q_0$, $\eta = \eta_0$, and loss estimates (3.6)
for $t = 1, \dots, T$ do
    Perform one step of the current instance of Algorithm 3.1
    if $\frac{1}{T}\sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} > 2^{r+1}$ then
        $r \leftarrow r + 1$
        $T_r \leftarrow t + 1$
        Restart Algorithm 3.1 with $q = q_r$, $\eta = \eta_r$, and loss estimates (3.6)
```

Theorem 3.3. *Let $C := 4\sqrt{6}e^{\frac{\sqrt{\pi}+\sqrt{4-2\ln 2}}{\ln 2}}$. Then, the regret of Algorithm 3.2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T} \alpha_t\left(2 + \ln\left(\frac{K}{\bar{\alpha}_T}\right)\right)} + \log_2 \bar{\alpha}_T \ .$$

Proof sketch. For simplicity, we sketch here the proof for the case when in every round t, all the nodes have self-loops; hence, $H_t(q) = B_t(q)$. See the full proof in Appendix A.3, which treats the general case in a similar manner. Let $n := \lceil \log_2 \bar{\alpha}_T \rceil$ and assume without loss of generality that $\bar{\alpha}_T > 1$. Since Lemma 3.1 implies that for any r and t, $B_t(q_r) \le \alpha_t^{q_r}$, we have as a consequence that for any $t \ge T_r$,

$$\frac{1}{T}\sum_{s=T_r}^{t} B_s(q_r)^{1/q_r} \le \frac{1}{T}\sum_{s=T_r}^{t} \alpha_s \le \bar{\alpha}_T \le 2^n \ .$$

Hence, the maximum value of r that the algorithm can reach is $n-1$. In doing so, we will execute n instances of Algorithm 3.1, each corresponding to a value of $r \in \{0, \dots, n-1\}$. For every such r, we upper bound the instantaneous regret at step $T_{r+1} - 1$ (the step when the restarting condition is satisfied) by 1, hence the added $\log_2 \bar{\alpha}_T$ term in the regret bound. For the rest of the interval—namely, for $t \in [T_r, T_{r+1} - 2]$—we have via Equation (3.4) that the regret of Algorithm 3.1 is bounded by

$$\frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r)\right] \ . \tag{3.7}$$

Define $T_{r:r+1} := T_{r+1} - T_r - 1$, and notice that

$$\begin{aligned}
\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r) &\le T_{r:r+1}\left(\frac{1}{T_{r:r+1}}\sum_{t=T_r}^{T_{r+1}-2} B_t(q_r)^{1/q_r}\right)^{q_r} \\
&\le T_{r:r+1}\left(\frac{T}{T_{r:r+1}}2^{r+1}\right)^{q_r} \\
&\le 2T\left(2^r\right)^{q_r} ,
\end{aligned}$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0, 1)$, and the second follows from the restarting condition of Algorithm 3.2. After, plugging this back into Equation (3.7), we can simply use the definitions of η_r and q_r and bound the resulting expression in a similar manner to the proof of Theorem 3.1. Overall, we get that

$$R_T \le 4\sqrt{3eT} \sum_{r=0}^{n-1} \sqrt{2^r \ln\bigl(e^2 K 2^{-r}\bigr)} + \log_2 \bar{\alpha}_T \ ,$$

from which the theorem follows by using Lemma A.3 in Appendix A.1, which shows, roughly speaking, that the sum on the right-hand side is of the same order as its last term. □

Although Algorithm 3.2 requires knowledge of the time horizon, this can be dealt with by applying a standard doubling trick on T at the cost of a larger constant factor. It is also noteworthy that the bound we obtained is of the form $\sqrt{T\bar{\alpha}_T(1 + \ln(K/\bar{\alpha}_T))}$ and not $\sqrt{\sum_t \alpha_t(1 + \ln(K/\alpha_t))}$. Even if both coincide with the bound of Theorem 3.2 when α_t is the same for all time steps, the latter is smaller by the concavity of $x(1 + \ln(K/x))$ in x. It is not clear, however, whether there is a tuning of $q \in (0, 1)$ that can achieve the second bound (even with prior knowledge of the entire sequence $\alpha_1, \dots, \alpha_T$ of independence numbers).

3.5 An Improved Lower Bound via Multitask Learning

In this section, we provide a novel lower bound on the minimax regret showing that, apart from the bandits case, a logarithmic factor is indeed necessary in general. When the graph is fixed over time, it is known that a lower bound of order $\sqrt{\alpha T}$ holds for any value of α (Mannor and Shamir, 2011, Alon et al., 2017). Whereas for the experts case ($\alpha = 1$), the minimax regret is of order[*] $\sqrt{T \ln K}$ (Cesa-Bianchi et al., 1997). The following theorem provides, for the first time, a lower bound that interpolates between the two aforementioned bounds for the intermediate values of α.

Theorem 3.4. *Pick any $K \ge 2$ and any α such that $2 \le \alpha \le K$. Then, for any algorithm and for all $T \ge \frac{\alpha \log_\alpha K}{4 \log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1, \dots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t \in [T]$ and*

$$R_T \ge \frac{1}{18\sqrt{2}} \sqrt{\alpha T \log_\alpha K}.$$

In essence, the proof of this theorem (see Appendix A.4) constructs a sequence of feedback graphs and losses that is equivalent to a hard instance of the multitask bandit problem (MTB, see Cesa-Bianchi and Lugosi (2012)), an important special case of combinatorial bandits with a convenient structure for proving lower bounds (Audibert et al., 2014, Cohen, Hazan, and Koren, 2017, Ito, Hatano, Sumita, Takemura, Fukunaga, Kakimura, and Kawarabayashi, 2019). We consider a variant of MTB in which, at the beginning of each round, the decision-maker selects an arm to play in each one of M stochastic bandit games. Subsequently, the decision-maker only observes (and suffers) the loss of the arm played in a single randomly selected game. For proving the lower bound, we use a class of stationary stochastic adversaries (i.e., environments), each generating graphs and losses in a manner that simulates an MTB instance.

[*] As a lower bound, this is commonly known to hold asymptotically as K and T grow. However, it can also be shown to hold non-asymptotically (though with worse leading constants); see Haussler, Kivinen, and Warmuth (1998, Theorem 3.22) or Cesa-Bianchi and Lugosi (2006, Theorem 3.6).

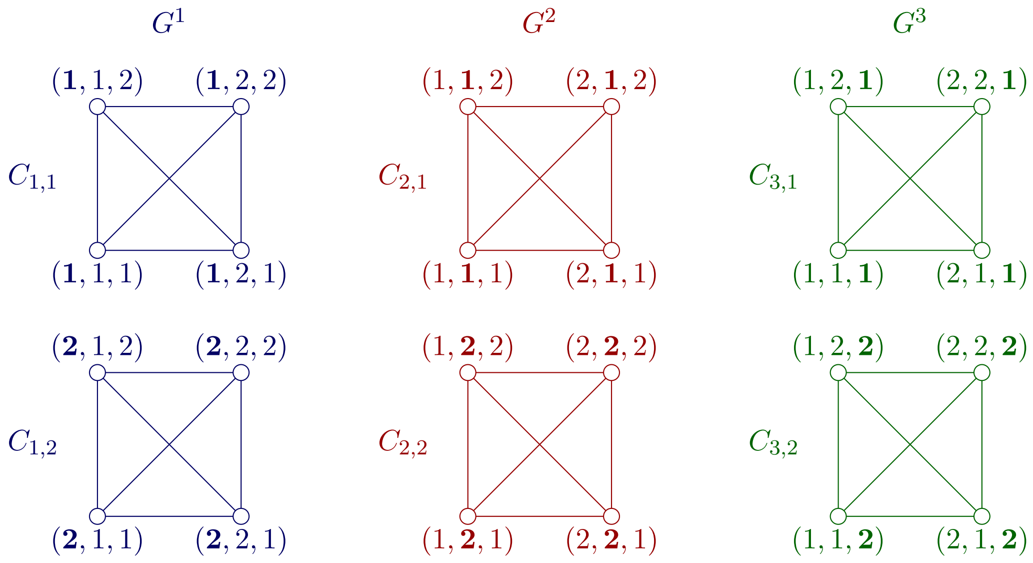


Figure 3.1: This figure shows an example of the multi-task bandit construction used to prove the lower bound. Here, $K = 8$ and $\alpha = 2$; thus, the number of games is $M = 3$. Each action is identified by a tuple of three numbers, each corresponding to a choice of one out of a pair of "base actions" in each game. Each of the three graphs in the figure corresponds to a game, such that two actions share an edge if and only if they choose the same base action in the corresponding game. At every round, a graph is randomly drawn, and all actions belonging to the same clique suffer the same loss.

Fix $2 \le \alpha \le K = |V|$ and assume for simplicity that $M = \log_\alpha K$ is an integer. We now construct an instance of online learning with time-varying feedback graphs $G_t = (V, E_t)$ with $\alpha(G_t) = \alpha$ that is equivalent to an MTB instance with M bandit games each containing α "base actions". Since $K = \alpha^M$, we can uniquely identify each action in V with a vector $a = \big(a(1), \dots, a(M)\big)$ in $[\alpha]^M$. The action $a_t \in V$ chosen by the learner at round t is equivalent to a choice of base actions $a_t(1), \dots, a_t(M)$ in the M games. The feedback graph at every round is sampled uniformly at random from a set of M undirected graphs $\{G^i\}_{i=1}^M$, where $G^i = (V, E^i)$ is such that $(a, a') \in E^i$ if and only if $a(i) = a'(i)$. This means (see Figure 3.1) that each graph G^i consists of α isolated cliques $\{C_{i,j}\}_{j=1}^\alpha$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. Clearly, the independence number of any such graph is α. Drawing feedback graph $G_t = G^i$ corresponds to the activation of game i in the MTB instance. Hence, choosing $a_t \in V$ with feedback graph $G_t = G^i$ is equivalent to playing base action $a_t(i)$ in game i in the MTB. As for the losses, we enforce that, given a feedback graph G_t, all actions that belong to the same clique of the feedback graph are assigned the same loss. Namely, if $G_t = G^i$ and $a(i) = a'(i) = j$, then $\ell_t(a) = \ell_t(a')$, which can be seen as the loss $\ell_t(j)$ assigned to base action j in game G^i. To choose the distribution of the losses for the base actions, we apply the classic needle-in-a-haystack approach of Auer et al. (1995) over the M games. More precisely, we construct a different environment for each action $a \in V$ in such a way that the distribution of the losses in each MTB game slightly favors (with a difference of a small $\varepsilon > 0$) the base action corresponding to a in that game. The proof then proceeds similarly to, for example, the proof of Theorem 5 in Audibert et al. (2014) or Theorem 7 in Eldowa et al. (2023a).

While both our upper and lower regret bounds achieve the desired goal of interpolating between the minimax rates of experts and bandits, the logarithmic factors in the two bounds are not exactly matching. In particular, if we compare $1 + \log_2(K/\alpha)$ and $\log_\alpha K$, we can see that although they coincide at $\alpha = 2$ and $\alpha = K$, the former is larger for intermediate values. It is reasonable to believe that the upper bound is of the correct order, seeing as it arose naturally as a result of choosing the best parameter for the Tsallis entropy regularizer, whereas achieving the extra logarithmic term in the lower bound required a somewhat contrived construction.

3.6 Conclusions

We have demonstrated that a proper tuning of the q-FTRL algorithm allows one to achieve a $\mathcal{O}\big(\sqrt{\alpha T(1 + \ln(K/\alpha))}\big)$ regret for the problem of online learning with undirected strongly observable feedback graphs. Our bound interpolates between the minimax regret rates of the bandits and the experts problems, the two extremes of the strongly observable graph feedback spectrum. Furthermore, we have shown how to achieve an analogous bound when the graphs vary over time, and without requiring any prior knowledge of the graphs. These results are complemented by our new regret lower bound of $\Omega\big(\sqrt{\alpha T(\ln K)/(\ln \alpha)}\big)$, which holds for $\alpha \geq 2$ and shows the necessity of a logarithmic factor in the minimax regret except for the bandits case. While our results provided the tightest characterization of the minimax rate for this setting, a subsequent work (Chen, He, and Zhang, 2024) showed a lower bound for fixed feedback graphs composed of disjoint cliques that implies worst-case optimality (up to constant factors) of our proposed algorithm for each pair of K and α—see Appendix A.5 for a more detailed comparison with results therein. Extending our results to the case of directed strongly observable feedback graphs is a considerably harder task—see Appendix A.6 for a preliminary discussion. Better understanding this more general setting is an interesting future direction.

Chapter 4

Improved Regret Bounds for Bandits with Expert Advice

In this chapter, we revisit the multi-armed bandit problem with expert advice. Under a restricted feedback model, we prove a novel lower bound of order $\sqrt{KT \ln(N/K)}$ for the worst-case regret by reduction to the multi-armed bandit problem with feedback graphs, where K is the number of actions, $N > K$ is the number of experts, and T is the time horizon. This matches a previously known upper bound of the same order and improves upon the best available lower bound of order $\sqrt{KT(\ln N)/(\ln K)}$. For the standard feedback model with expert advice, we prove a new instance-based regret bound that depends on a disagreement measure between the experts and provides a logarithmic improvement compared to prior results.

4.1 Introduction

The problem of bandits with expert advice provides a simple and general framework for incorporating contextual information into the multi-armed bandit problem. This framework, already introduced in Chapter 2, can be summarized as follows: compared to standard bandits, the learner additionally receives in every round a recommendation, in the form of a probability distribution over the actions, from each expert in a given set. This set of experts can be seen as a set of strategies, each mapping an unobserved context to a (randomized) action choice. The goal of the learner is to minimize their expected regret with respect to the best expert in hindsight (see its definition in Equation (2.3)); that is, the difference between their expected cumulative loss and that of the best fixed expert, which differs from competing against the best fixed action as per the classical notion of regret from plain bandits. We briefly recall that this problem was first formulated by Auer et al. (1995, 2002b), who proposed the Exp4 algorithm as a solution strategy. Auer et al. (2002b) proved a regret bound of order $\sqrt{KT \ln N}$ on the expected regret incurred by the Exp4 strategy, where T denotes the number of rounds, K the number of actions, and N the number of experts. This result is of a worst-case nature, in that it holds for any sequence of losses assigned to the actions and for any sequence of expert recommendations.

The appealing feature of the bound of Auer et al. (2002b) is that it exhibits only a logarithmic dependence on the number of experts, in addition to the $\sqrt{K}$ dependence on the number of actions known to be unavoidable in the classical bandit problem. While the minimax regret in the latter problem has been shown to be of order $\sqrt{KT}$ (Audibert and Bubeck, 2009) modulo constant factors, a similar exact characterization remains missing for the expert advice problem. Kale (2014) studied a generalized version of the bandits with expert advice problem—originally proposed by Seldin, Crammer, and Bartlett (2013)—where the learner is only allowed to query the advice of $M \leq N$ experts. When $M = N$, the results of Kale (2014) imply an upper bound of order $\sqrt{\min\{K,N\}T\bigl(1+\ln(N/\min\{K,N\})\bigr)}$ on the minimax regret, improving upon the bound of Auer et al. (2002b). Unlike the latter, the logarithmic factor in Kale (2014) bound diminishes as K increases with respect to N, leading to a bound of order $\sqrt{NT}$ when $N \leq K$, which is tight in general as the experts in that case can be made to emulate an N-armed bandit problem. This improved bound was achieved via the PolyINF algorithm (Audibert and Bubeck, 2009, 2010) played on the expert set utilizing the importance-weighted loss estimators of Exp4. Later, Seldin and Lugosi (2016) proved a lower bound of order $\sqrt{KT(\ln N)/(\ln K)}$ for $N \geq K$.

As these upper and lower bounds on the regret still preserve an open gap, the correct minimax rate remains unclear. In this chapter, we take a step towards resolving this issue by showing that the upper bound is not improvable in general under a restricted feedback model in which the importance-weighted loss estimators used by Exp4 or PolyINF remain implementable. In this restricted model, without observing the experts' recommendations yet, the learner picks an expert (possibly at random) at the beginning of each round, and the environment subsequently samples the action to be executed from the chosen expert's distribution. Afterwards, the learner only observes the distributions of the experts that had assigned positive probability to the chosen action. As evinced in the previous paragraph, the phenomenon manifested in the regret bound by Kale (2014) is similar in spirit to the behaviour we observed in the previous Chapter 3 for the problem of multi-armed bandits with feedback graphs. We then leverage this (seemingly faint) connection and, via a reduction from the feedback graphs setting, we use the recent results of Chen et al. (2024) to obtain a lower bound of order $\sqrt{KT\ln(N/K)}$ for bandits with expert advice when $N > K$.

Departing from the worst-case results discussed thus far, a few works have obtained instance-dependent bounds for this problem. The dependence on the instance can be in terms of the assigned sequence of losses through small-loss bounds (Allen-Zhu, Bubeck, and Li, 2018), or in terms of the sequence of expert recommendations through bounds that reflect the similarity between the recommended expert distributions (McMahan and Streeter (2009), Lattimore and Szepesvári (2020, Theorem 18.3), Eldowa, Cesa-Bianchi, Metelli, and Restelli (2024)). Our focus here is on the latter case, where to the best of our knowledge the state of the art is a bound of order $\sqrt{\sum_{t=1}^{T}\mathcal{C}_t \ln N}$, shown in the recent work of Eldowa et al. (2024) for the Exp4 algorithm. Here, $\mathcal{C}_t$ is the (chi-squared) capacity of the recommended distributions at round t. This quantity measures the dissimilarity between the experts' recommendations and satisfies $0 \leq \mathcal{C}_t \leq \min\{K,N\}-1$. Improving upon this result, we illustrate that it is possible to achieve a bound of order $\sqrt{\sum_{t=1}^{T}\mathcal{C}_t\bigl(1+\ln(N/\max\{\bar{\mathcal{C}}_T,1\})\bigr)}$, where $\bar{\mathcal{C}}_T = \sum_{t=1}^{T}\mathcal{C}_t/T$ is the average capacity. This bound combines the best of the bound of Eldowa et al. (2024), i.e., its dependence on the agreement between the experts, and that of Kale (2014), i.e., its improved logarithmic factor, simultaneously outperforming both.

Roadmap. For better clarity, we recall the formal definition of the problem setting in the next section. In Section 4.3, as a preliminary building block, we present Algorithm 4.1, an instance of the FTRL algorithm with the (negative) q-Tsallis entropy as the regularizer, also previously adopted in Chapter 3 for the multi-armed bandit problem with feedback graphs. We then show in Section 4.4 that combining this algorithm with a doubling trick allows us to achieve the improved instance-based bound mentioned above. The lower bound for the restricted feedback setting is presented in Section 4.5. Finally, we provide some concluding remarks in Section 4.6.

4.2 Problem Setting and Notations

We present here the formal definition of the multi-armed bandit problem with expert advice, which has slightly different notation compared to its introduction in Chapter 2. This choice will later help depict in a more immediate way the connection to the feedback graphs setting. Furthermore, only within this chapter, we let $x \vee y \coloneqq \max\{x, y\}$ and $x \wedge y \coloneqq \min\{x, y\}$ for any $x, y \in \mathbb{R}$ to ease the presentation of the results and the overall readability.

Let $V \coloneqq [N]$ be a finite set of N experts and $\mathcal{A} \coloneqq [K]$ be a finite set of K actions. The environment is characterized by a fixed and unknown sequence of $[0,1]$-valued loss functions $(\ell_t)_{t\in[T]}$ over actions, and a fixed and unknown sequence of expert advice $(\theta_t^i)_{i\in V, t\in[T]}$, where $\theta_t^i \in \Delta_K$ is the distribution over actions recommended by expert i at round t. At the beginning of each round $t \in [T]$, the expert recommendations $(\theta_t^i)_{i\in V}$ are revealed to the learner, who then selects (possibly at random) an action $A_t \in \mathcal{A}$ and subsequently suffers and observes the loss $\ell_t(A_t)$. For any expert $i \in V$, we define its loss in round t as $y_t(i) \coloneqq \langle \ell_t, \theta_t^i \rangle = \sum_{a\in\mathcal{A}} \theta_t^i(a)\ell_t(a)$. The goal is to minimize the expected regret with respect to the best expert in hindsight:

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(A_t)\right] - \min_{i\in V} \sum_{t=1}^{T} y_t(i) \ ,$$

where the expectation is taken with respect to the randomization of the learner.

4.3 FTRL with Tsallis Entropy for Bandits with Expert Advice

The Exp4 algorithm can be seen as an instance of the FTRL framework (Orabona, 2019, Chapter 7), where a distribution p_t over the experts is maintained at each round t and updated as follows

$$p_{t+1} \leftarrow \operatorname*{arg\,min}_{p\in\Delta_N} \eta \left\langle \sum_{s=1}^{t} \widehat{y}_s, p \right\rangle + \sum_{i\in V} p(i) \ln p(i) \, ,$$

where $\eta > 0$ is the learning rate, the second term is the negative Shannon entropy of p, and $\widehat{y}_s(i)$ is an importance-weighted estimate of $y_s(i)$. The action A_t is then drawn from the mixture distribution $\sum_{i\in V} p_t(i)\theta_t^i(\cdot)$. Consider a more general algorithm (outlined in Algorithm 4.1) where the negative Shannon entropy is replaced with the negative q-Tsallis entropy, which for $q \in (0, 1)$, we recall, is given by

$$\psi_q(x) = \frac{1}{1-q}\left(1 - \sum_{i\in V} x(i)^q\right) \qquad \forall x \in \Delta_N \, .$$

In the limit when $q \to 1^-$, the negative Shannon entropy is recovered. The following theorem provides a regret bound for this proposed algorithm. This result is not novel, as a similar bound is implied by Theorem 2 in Kale (2014) for a closely related algorithm in a more general setting.

Algorithm 4.1: q-FTRL for bandits with expert advice

```
input: $q \in (0,1)$, $\eta > 0$
initialization: $p_1(i) \leftarrow 1/N$ for all $i \in V$
for $t = 1, \dots, T$ do
    receive expert advice $(\theta_t^i)_{i \in V}$
    draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
    construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) := \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
    let $p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta \langle \sum_{s=1}^t \widehat{y}_s, p \rangle + \psi_q(p)$
```

We provide a concise proof of the result for completeness. As mentioned before, when $N \leq K$, this bound is trivially tight in general. Otherwise, when $N > K$, we prove an order-wise matching minimax lower bound in Section 4.5 under additional restrictions on the received feedback.

Theorem 4.1. *Let $\xi := K \wedge N$. Algorithm 4.1 run with*

$$q = \frac{1}{2}\left(1 + \frac{\ln(N/\xi)}{\sqrt{\ln^2(N/\xi) + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \sqrt{\frac{2qN^{1-q}}{T(1-q)\xi^q}} ,$$

guarantees expected regret

$$R_T \leq 2\sqrt{e\xi T\big(2 + \ln(N/\xi)\big)} .$$

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^T y_t(i)$, and note that $R_T = \sum_{t=1}^T \mathbb{E}\big[y_t(I_t) - y_t(i^*)\big]$ as $\mathbb{E}[\ell_t(A_t)] = \mathbb{E}[y_t(I_t)]$. For any round $t \in [T]$, let $\mathcal{F}_t := \sigma(I_1, A_1, \dots, I_t, A_t)$ denote the σ-algebra generated by the random events up to the end of round t, and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_0$ being the trivial σ-algebra. For any action $a \in \mathcal{A}$, let $\phi_t(a) := \sum_{i \in V} p_t(i)\theta_t^i(a)$ and note that, conditioned on $\mathcal{F}_{t-1}$, A_t is distributed according to ϕ_t. As p_t is $\mathcal{F}_{t-1}$-measurable, it is then easy to verify that $\mathbb{E}_t[\widehat{y}_t] = y_t$. Hence, Lemma A.2 in Appendix A.1 (relative to the results from Chapter 3) implies that

$$R_T \leq \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^T \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{y}_t(i)^2\right] . \tag{4.1}$$

For any fixed $t \in [T]$ and $i \in V$, we have that

$$\mathbb{E}_t\left[\widehat{y}_t(i)^2\right] = \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\ell_t(A_t)^2\right] \leq \mathbb{E}_t\left[\frac{\theta_t^i(A_t)^2}{\phi_t(A_t)^2}\right] = \mathbb{E}_t\left[\sum_{a \in \mathcal{A}} \frac{\theta_t^i(a)^2}{\phi_t(a)^2}\mathbb{I}\{a = A_t\}\right] = \sum_{a \in \mathcal{A}} \frac{\theta_t^i(a)^2}{\phi_t(a)} , \tag{4.2}$$

where the inequality holds because $\ell_t(A_t) \in [0,1]$ and the final equality holds because $\mathbb{E}_t[\mathbb{I}\{a = A_t\}] = \mathbb{P}(a = A_t \mid \mathcal{F}_{t-1}) = \phi_t(a)$. Thus, it holds that

$$\begin{aligned}\mathbb{E}_t\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{y}_t(i)^2\right] &= \sum_{a \in \mathcal{A}} \frac{\sum_{i \in V} p_t(i)^{2-q}\theta_t^i(a)^2}{\phi_t(a)} \\ &\leq \sum_{a \in \mathcal{A}} \frac{\sum_{i \in V} p_t(i)^{2-q}\theta_t^i(a)^{2-q}}{\phi_t(a)} \max_{i \in V} \theta_t^i(a)^q\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{a\in\mathcal{A}} \frac{\left(\sum_{i\in V} p_t(i)\theta_t^i(a)\right)^{2-q}}{\phi_t(a)} \max_{i\in V} \theta_t^i(a)^q \\
&= \sum_{a\in\mathcal{A}} \phi_t(a)\left(\frac{\max_{i\in V}\theta_t^i(a)}{\phi_t(a)}\right)^q \\
&\le \left(\sum_{a\in\mathcal{A}} \max_{i\in V} \theta_t^i(a)\right)^q \le \xi^q \ ,
\end{aligned}
$$

where the second inequality follows from the superadditivity of x^{2-q} for $x \geq 0$ and $q \in (0,1)$, the third inequality follows from the concavity of x^q for $q \in (0,1)$ because of Jensen's inequality, and the last inequality holds since $\max_{i\in V}\theta_t^i(a) \le \min\bigl\{1, \sum_{i\in V}\theta_t^i(a)\bigr\}$. Substituting back into Equation (4.1) yields that

$$
R_T \le \frac{N^{1-q}}{(1-q)\eta} + \frac{\eta}{2q}\xi^q T \ .
$$

Finally, in a similar manner as the proof of Theorem 3.1 from Chapter 3, substituting the specified values of η and q allows us to conclude the proof by showing that the expected regret satisfies

$$
\begin{aligned}
R_T &\le \sqrt{\frac{2N^{1-q}\xi^q}{q(1-q)}T} \\
&= \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\xi N) - \frac{1}{2}\sqrt{\ln^2(N/\xi)+4}\right)\left(2+\sqrt{\ln^2(N/\xi)+4}\right)} \\
&\le \sqrt{2T\exp\left(1+\frac{1}{2}\ln(\xi N) - \frac{1}{2}\ln(N/\xi)\right)\left(2+\sqrt{\ln^2(N/\xi)+4}\right)} \\
&= \sqrt{2e\xi T\left(2+\sqrt{\ln^2(N/\xi)+4}\right)} \le 2\sqrt{e\xi T\sqrt{\ln^2(N/\xi)+4}} \\
&\le 2\sqrt{e\xi T\bigl(2+\ln(N/\xi)\bigr)} \ .
\end{aligned}
$$

□

4.4 An Improved Instance-Based Regret Bound

We now proceed to introduce some fundamental notions, also present in Eldowa et al. (2024), with the aim of deriving better instance-dependent regret guarantees. These concepts will allow us to obtain a more refined regret bound whose form is analogous to the bound of Theorem 4.1, except that it will depend on the dissimilarity between the experts' recommendations at each round, replacing $K \wedge N$ with an effective number of experts. Before discussing the algorithm, we introduce these relevant quantities: for any round $t \in [T]$ and any probability distribution $\tau \in \Delta_N$, define

$$
Q_t(\tau) \coloneqq \sum_{i\in V}\tau(i)\chi^2\bigl(\theta_t^i \,\|\, \textstyle\sum_{j\in V}\tau(j)\theta_t^j\bigr) = \displaystyle\sum_{a\in\mathcal{A}} \frac{\sum_{i\in V}\tau(i)\theta_t^i(a)^2}{\sum_{j\in V}\tau(j)\theta_t^j(a)} - 1 \ ,
$$

where $\chi^2(p\,\|\,q) \coloneqq \sum_{a\in\mathcal{A}} q(a)\bigl(p(a)/q(a)-1\bigr)^2 = \sum_{a\in\mathcal{A}} p(a)^2/q(a) - 1$ is the chi-squared divergence between distributions $p, q \in \Delta_K$. Additionally, let

$$
\mathcal{C}_t \coloneqq \sup_{\tau\in\Delta_N} Q_t(\tau) \qquad \text{and} \qquad \bar{\mathcal{C}}_T \coloneqq \frac{1}{T}\sum_{t=1}^T \mathcal{C}_t
$$

be the chi-squared capacity of the recommended distributions at round t and their average over the T rounds. As remarked before, $\mathcal{C}_t$ is never larger than $(K \wedge N) - 1$ and can be arbitrarily smaller depending on the agreement between the experts at round t. In particular, it vanishes when all recommendations are identical.

The idea of Algorithm 4.2 is to tune Algorithm 4.1 as done in Theorem 4.1, but with $\bar{\mathcal{C}}_T$ replacing $K \wedge N$. However, to avoid requiring prior knowledge of $\bar{\mathcal{C}}_T$, we rely on a doubling trick to adapt to its value. At any given time step t, we maintain a running instance of Algorithm 4.1 tuned with an estimate for $\bar{\mathcal{C}}_T$. Let m_t be the time step when the present execution of Algorithm 4.1 at round t had started. If the current estimate is found to be smaller than $\frac{1}{2T}\sum_{s=m_t}^{t} Q_s(p_s)$, the algorithm is restarted and the estimate is (at least) doubled. This quantity we test against is a simple lower bound for $\bar{\mathcal{C}}_T/2$ that can be constructed without computing the capacity at any round. As the value of $\bar{\mathcal{C}}_T$ can be arbitrarily close to zero, the initial guess (which ideally should be a lower bound) is left as a user-specified parameter for the algorithm, and appears in the first (and more general) bound of Theorem 4.2. The second statement of the theorem shows that choosing $\ln(e^2N)/T$ as the initial guess suffices to obtain a regret guarantee of order $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t\big(1 + \ln(N/\max\{\bar{\mathcal{C}}_T, 1\})\big)}$, up to an additive logarithmic term. This simultaneously outperforms the $\sqrt{\sum_{t=1}^{T} \mathcal{C}_t \ln N}$ bound of Eldowa et al. (2024) and the $\sqrt{(K \wedge N)T\big(1 + \ln(N/(K \wedge N))\big)}$ bound of Kale (2014).

The proof combines elements from the proof of Theorem 1 of Eldowa et al. (2024) and the proof of Theorem 3.3 from Chapter 3 which employs a similar algorithm to address online learning with time-varying feedback graphs. Compared to the techniques adopted in Chapter 3, we require a more refined analysis to account for the case when $\bar{\mathcal{C}}_T < 1$. This refinement is achieved in part via the use of Lemma B.1, which also allows adapting the analysis of Eldowa et al. (2024) to account for the fact that we use the negative q-Tsallis entropy as a regularizer in place of the negative Shannon entropy.

Theorem 4.2. *Assuming that $T \geq \ln(e^2N)$, Algorithm 4.2 run with input $J \in (0, N]$ satisfies*

$$R_T \leq 38e\sqrt{(\bar{\mathcal{C}}_T \vee J)T \ln\left(\frac{e^2N}{\bar{\mathcal{C}}_T \vee J \vee 1}\right)} + \left[\log_2\left(\frac{\bar{\mathcal{C}}_T}{J}\right)\right]_+ + \frac{18e}{5}\ln\big(e^2N\big)\left[\log_2\left(\frac{4\left(\big(JT \vee \bar{\mathcal{C}}_T T\big) \wedge \ln(e^2N)\right)}{JT}\right)\right]_+ + 1\,.$$

In particular, setting $J = \ln(e^2N)/T$ yields that

$$R_T \leq 38e\sqrt{\bar{\mathcal{C}}_T T \ln\left(\frac{e^2N}{\bar{\mathcal{C}}_T \vee 1}\right)} + \left[\log_2\left(\frac{\bar{\mathcal{C}}_T T}{\ln(e^2N)}\right)\right]_+ + 46e\ln\big(e^2N\big) + 1\,.$$

Proof. For brevity, we define $U := \bar{\mathcal{C}}_T \vee J$. Let $s := \lceil \log_2 J \rceil - 1$ and $n := \lceil \log_2 U \rceil - 1$, the latter of which is the largest value that r_t can take because, for any round t,

$$\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \leq \frac{1}{T}\sum_{s=1}^{T} Q_s(p_s) \leq \frac{1}{T}\sum_{s=1}^{T} \mathcal{C}_s \leq 2^{n+1}\,.$$

Without loss of generality, we assume that for any (integer) $r \in \{s, \ldots, n\}$, there are at least two rounds in which $r_t = r$, and we use T_r to refer to the index of the first such round. Additionally, we

Algorithm 4.2: q-FTRL with the doubling trick for bandits with expert advice

```
input: $J \in (0, N]$
initialization: $r_1 \leftarrow \lceil \log_2 J \rceil - 1$, $m_1 \leftarrow 1$, $p_1(i) \leftarrow 1/N$ for all $i \in V$
define: For each integer $r \in (-\infty, \log_2 N]$,
    $q_r := \frac{1}{2}\left(1 + \frac{\ln(N/2^r)}{\sqrt{\ln^2(N/2^r) + 4} + 2}\right)$
    $\eta_r := \min\left\{ \sqrt{\frac{q_r(N^{1-q_r} - 1)}{eT(1-q_r)\,(2^r)^{q_r}}}\,,\ \frac{q_r}{1-q_r}\left(1 - e^{\frac{q_r - 1}{2 - q_r}}\right) \right\}$
for $t = 1, \dots, T$ do
    receive expert advice $(\theta_t^i)_{i \in V}$
    draw expert $I_t \sim p_t$ and action $A_t \sim \theta_t^{I_t}$
    construct $\widehat{y}_t \in \mathbb{R}^N$ where $\widehat{y}_t(i) := \frac{\theta_t^i(A_t)}{\sum_{j \in V} p_t(j)\theta_t^j(A_t)} \ell_t(A_t)$ for all $i \in V$
    if $\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) > 2^{r_t + 1}$ then
        $p_{t+1}(i) \leftarrow 1/N$ for all $i \in V$
        $r_{t+1} \leftarrow \lceil \log_2\big(\frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s)\big) \rceil - 1$, $m_{t+1} \leftarrow t + 1$
    else
        $p_{t+1} \leftarrow \arg\min_{p \in \Delta_N} \eta_{r_t} \langle \sum_{s=m_t}^{t} \widehat{y}_s, p \rangle + \psi_{q_{r_t}}(p)$
        $r_{t+1} \leftarrow r_t$, $m_{t+1} \leftarrow m_t$
```

define $T_{n+1} := T + 2$. Note that for any r in this range, $q_r \in [1/2, 1)$. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^{T} y_t(i)$. We start by decomposing the regret over the intervals corresponding to fixed values of $r_t \in \{s, \dots, n\}$ and bounding the instantaneous regret at the last step of each but the last interval by 1:

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\le \mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] + n - s \\
&\le \mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] + \log_2\big(U/J\big) + 1\,.
\end{aligned}
\tag{4.3}
$$

Let $\mathbf{e}_{i^*} \in \mathbb{R}^N$ be the indicator vector for i^* and define $\widetilde{y}_t \in \mathbb{R}^N$ where $\widetilde{y}_t(i) := \widehat{y}_t(i) - \ell_t(A_t)$ for every $i \in V$. Similar to the proof of Theorem 3.3 from Chapter 3, for each $r \in \{s, \dots, n\}$ we note that

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\langle p_t - \mathbf{e}_{i^*}, \widehat{y}_t\rangle\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t\rangle\right]
\end{aligned}
$$

$$= \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t \rangle\right] ,$$

where (a) follows since $\mathbb{E}_t\big[y_t(I_t)\big] = \sum_{i\in V} p_t(i)y_t(i)$, $\mathbb{E}_t\big[\widehat{y}_t\big] = y_t$, and the fact that the indicator at round t is measurable with respect to $\mathcal{F}_{t-1}$ (where $\mathcal{F}_{t-1}$ and $\mathbb{E}_t$ are defined in the same way as in the proof of Theorem 4.1); and (b) follows since $p_t, \mathbf{e}_{i^*} \in \Delta_N$ and $\widehat{y}_t(i) - \widetilde{y}_t(i) = \ell_t(A_t)$ is identical for all $i \in V$. Similarly to the last argument, the fact that $\big\langle \widetilde{y}_s - \widehat{y}_s, p - q \big\rangle = 0$ holds for any $p, q \in \Delta_N$ at any round s implies that p_{t+1} can be equivalently defined as $\arg\min_{p\in\Delta_N} \eta_{r_t} \big\langle \sum_{s=m_t}^{t} \widetilde{y}_s, p \big\rangle + \psi_{q_{r_t}}(p)$. Hence, using that $\widetilde{y}_t(i) \geq -1$, we can invoke Lemma B.1 (with $b = 1$ and $c = e$) to obtain that

$$\sum_{t=T_r}^{T_{r+1}-2} \langle p_t - \mathbf{e}_{i^*}, \widetilde{y}_t \rangle \leq \frac{N^{1-q_r} - 1}{(1-q_r)\eta_r} + \frac{e\eta_r}{2q_r} \sum_{t=T_r}^{T_{r+1}-2} \sum_{i\in V} p_t(i)^{2-q_r} \widetilde{y}_t(i)^2 .$$

For any round $t \in [T]$ and action $a \in \mathcal{A}$, recall that ϕ_t is defined so that $\phi_t(a) = \sum_{i\in V} p_t(i)\theta_t^i(a)$. Similar to Equation (4.2) in the proof of Theorem 4.1, we have that

$$\begin{aligned}
\mathbb{E}_t\left[\widetilde{y}_t(i)^2\right] &= \mathbb{E}_t\left[\ell_t(A_t)^2 \frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&\leq \mathbb{E}_t\left[\frac{\big(\theta_t^i(A_t) - \phi_t(A_t)\big)^2}{\phi_t(A_t)^2}\right] \\
&= \sum_{a\in\mathcal{A}} \frac{\big(\theta_t^i(a) - \phi_t(a)\big)^2}{\phi_t(a)} = \sum_{a\in\mathcal{A}} \phi_t(a)\left(\frac{\theta_t^i(a)}{\phi_t(a)} - 1\right)^2 = \chi^2(\theta_t^i \,\|\, \phi_t) .
\end{aligned}$$

Therefore, for any round t and any $r \in \{s, \dots, n\}$, it holds that

$$\begin{aligned}
\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q_r} \widetilde{y}_t(i)^2\right] &\leq \sum_{i\in V} p_t(i)^{2-q_r} \chi^2(\theta_t^i \,\|\, \phi_t) \\
&= Q_t(p_t) \sum_{i\in V} \frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)} p_t(i)^{1-q_r} \\
&\leq Q_t(p_t) \left(\sum_{i\in V} \frac{p_t(i)\chi^2(\theta_t^i \,\|\, \phi_t)}{Q_t(p_t)} p_t(i)\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r} \left(\sum_{i\in V} p_t(i)^2 \chi^2(\theta_t^i \,\|\, \phi_t)\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r} \left(\sum_{i\in V} p_t(i)^2 \sum_{a\in\mathcal{A}} \frac{\theta_t^i(a)^2}{\phi_t(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r} \left(\sum_{a\in\mathcal{A}} \frac{\sum_{i\in V} p_t(i)^2 \theta_t^i(a)^2}{\sum_{j\in V} p_t(j)\theta_t^j(a)} - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&\leq Q_t(p_t)^{q_r} \left(\sum_{a\in\mathcal{A}} \sum_{i\in V} p_t(i)\theta_t^i(a) - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \\
&= Q_t(p_t)^{q_r} \left(1 - \sum_{i\in V} p_t(i)^2\right)^{1-q_r} \leq Q_t(p_t)^{q_r} ,
\end{aligned}$$

where the second inequality follows from the definition of $Q_t(p_t)$ and the fact that x^{1-q_r} is concave in $x \ge 0$, and the third inequality uses the superadditivity of x^2 for non-negative real numbers and the non-negativity of the quantity in brackets. If we define $T_{r:r+1} \coloneqq T_{r+1} - T_r - 1$, it then holds that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{y}_t(i)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=m_t}^{t} Q_s(p_s) \le 2^{r_t}\right\}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{y}_t(i)^2\right] \\
&\le \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)^{q_r}\right] \\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{1}{T_{r:r+1}}\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t)\right)^{q_r}\right] \\
&\le \mathbb{E}\left[T_{r:r+1}\left(\frac{T}{T_{r:r+1}}2^{r+1}\right)^{q_r}\right] \\
&\le 2T\left(2^r\right)^{q_r} ,
\end{aligned}$$

where the second inequality uses the concavity of x^{q_r} in $x \ge 0$ and the third inequality uses that $\frac{1}{T}\sum_{t=T_r}^{T_{r+1}-2} Q_t(p_t) \le 2^{r+1}$ since the algorithm is not reset in the interval $[T_r, T_{r+1}-2]$.

Overall, we have shown that

$$\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \le \frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T .$$

If $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} \le \frac{q_r}{1-q_r}\left(1 - e^{\frac{q_r-1}{2-q_r}}\right)$, then substituting the values of η_r and q_r gives that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T &= 2\sqrt{\frac{e(N^{1-q_r}-1)\left(2^r\right)^{q_r} T}{q_r(1-q_r)}} \\
&= 2\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{\frac{eN^{1-q_r}\left(2^r\right)^{q_r} T}{q_r(1-q_r)}} \\
&\le 2e\sqrt{2}\sqrt{\frac{N^{1-q_r}-1}{N^{1-q_r}}}\sqrt{2^r\left(2+\ln(N2^{-r})\right)T} \\
&\le 2e\sqrt{2}\left(\sqrt{\frac{\ln N}{\ln(N2^{-r})}} \wedge 1\right)\sqrt{2^r\left(2+\ln(N2^{-r})\right)T} \\
&= 2e\sqrt{2}\sqrt{2^r\ln\big(e^2N(2^{-r}\wedge 1)\big)T} ,
\end{aligned}$$

where the first inequality holds via the same arguments laid in the last passage of the proof of Theorem 4.1, and the second inequality holds because

$$\begin{aligned}
\frac{N^{1-q_r}-1}{N^{1-q_r}} &= 1 - \exp\left(-\ln\big(N^{1-q_r}\big)\right) \\
&\le (1-q_r)\ln N \\
&= \frac{1}{2}\left(1 - \frac{\ln(N/2^r)}{\sqrt{\ln(N/2^r)^2+4}+2}\right)\ln N
\end{aligned}$$

$$= \frac{\ln N}{2\ln(N/2^r)}\Big(2 + \ln(N/2^r) - \sqrt{\ln(N/2^r)^2 + 4}\Big) \le \frac{\ln N}{\ln(N/2^r)} \ ,$$

using also the fact that $1 - e^{-x} \le x$. Otherwise, if $\sqrt{\frac{q_r(N^{1-q_r}-1)}{eT(1-q_r)(2^r)^{q_r}}} > \frac{q_r}{1-q_r}\Big(1 - e^{\frac{q_r-1}{2-q_r}}\Big)$, then η_r takes the latter value and we obtain that

$$\begin{aligned}
\frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \frac{e\eta_r}{q_r}\left(2^r\right)^{q_r} T &\le \frac{N^{1-q_r}-1}{(1-q_r)\eta_r} + \eta_r \frac{N^{1-q_r}-1}{(1-q_r)}\left(\frac{1-q_r}{q_r\Big(1-e^{\frac{q_r-1}{2-q_r}}\Big)}\right)^2 \\
&= 2\frac{N^{1-q_r}-1}{q_r\Big(1-e^{\frac{q_r-1}{2-q_r}}\Big)} \\
&\le \frac{18\big(N^{1-q_r}-1\big)}{5q_r(1-q_r)} \\
&= \frac{18\left(2^r\right)^{-q_r}\big(N^{1-q_r}-1\big)\left(2^r\right)^{q_r}}{5q_r(1-q_r)} \\
&\le \frac{18e}{5}\left(2^r\right)^{1-q_r}\ln\big(e^2N(2^{-r}\wedge 1)\big) \\
&\le \frac{18e}{5}\big(1 \vee \sqrt{2^r}\big)\ln\big(e^2N(2^{-r}\wedge 1)\big) \ ,
\end{aligned}$$

where the last inequality holds because $q_r \ge 1/2$, and the second inequality holds since

$$1 - e^{\frac{q_r-1}{2-q_r}} \ge \frac{1-q_r}{2-q_r} - \frac{1}{2}\left(\frac{1-q_r}{2-q_r}\right)^2 = \frac{3-q_r}{2(2-q_r)^2}(1-q_r) \ge \frac{5}{9}(1-q_r)\ln\big(e^2N(2^{-r}\wedge 1)\big) \ ,$$

using that $e^{-x} \le 1 - x + x^2/2$ for $x \ge 0$ and that $q_r \ge 1/2$. Thus, the results above yield that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\qquad \le \max\left\{2e\sqrt{2}\sqrt{2^rT\ln\big(e^2N(2^{-r}\wedge 1)\big)},\ \frac{18e}{5}\big(1 \vee \sqrt{2^r}\big)\ln\big(e^2N(2^{-r}\wedge 1)\big)\right\} \ .
\end{aligned} \tag{4.4}$$

Let $M \coloneqq \ln(e^2N)/T$ and $m \coloneqq \log_2 M$, and note that $m \le 0$ (and $M \le 1$) holds by the assumption that $T \ge \ln(e^2N)$. In the case when $n \le 0$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r=s}^{n}\sum_{t=T_r}^{T_{r+1}-2}\big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\qquad \le \frac{18e}{5}\big[(n \wedge \lfloor m \rfloor) - s + 1\big]_+ \ln\big(e^2N\big) + 2e\sqrt{2}\sum_{r=n\wedge\lceil m \rceil}^{n}\sqrt{2^rT\ln\big(e^2N\big)} \\
&\qquad \le \frac{18e}{5}\big[\log_2\big(4(U \wedge M)/J\big)\big]_+ \ln\big(e^2N\big) + 8e\sqrt{2UT\ln\big(e^2N\big)} \ ,
\end{aligned}$$

where the second inequality uses that

$$\sum_{r=\alpha}^{n}\big(\sqrt{2}\big)^r = \big(\sqrt{2}\big)^\alpha\sum_{r=0}^{n-\alpha}\big(\sqrt{2}\big)^r = \big(\sqrt{2}\big)^\alpha\frac{\big(\sqrt{2}\big)^{n-\alpha+1} - 1}{\sqrt{2}-1} \le \frac{\sqrt{2}}{\sqrt{2}-1}\big(\sqrt{2}\big)^n \le 4\sqrt{U} \ ,$$

with $\alpha \coloneqq n \wedge \lceil m \rceil$. Otherwise, if $n > 0$, then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{r=s}^{n} \sum_{t=T_r}^{T_{r+1}-2} \big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\quad\le \frac{18e}{5}\big[\log_2(4M/J)\big]_+ \ln\!\big(e^2N\big) + 8e\sqrt{2T\ln\!\big(e^2N\big)} + \mathbb{E}\left[\sum_{r=s_+}^{n} \sum_{t=T_r}^{T_{r+1}-2} \big(y_t(I_t) - y_t(i^*)\big)\right] \\
&\quad\le \frac{18e}{5}\big[\log_2(4M/J)\big]_+ \ln\!\big(e^2N\big) + 8e\sqrt{2T\ln\!\big(e^2N\big)} + \frac{18e}{5}\sum_{r=0}^{n}\sqrt{2^r\ln\!\big(e^2N2^{-r}\big)T} \\
&\quad\le \frac{18e}{5}\big[\log_2(4M/J)\big]_+ \ln\!\big(e^2N\big) + 8e\sqrt{2T\ln\!\big(e^2N\big)} + 26e\sqrt{UT\ln\!\big(e^2N/U\big)} \\
&\quad\le \frac{18e}{5}\big[\log_2(4M/J)\big]_+ \ln\!\big(e^2N\big) + 38e\sqrt{UT\ln\!\big(e^2N/U\big)}\ ,
\end{aligned}$$

where the first inequality follows from the analysis of the first case with $n = 0$, the second inequality uses that $r \ge 0$ and the assumption that $T \ge \ln(e^2N)$, the third inequality uses Lemma A.3 from Appendix A.1 (relative to Chapter 3), and the fourth uses that $x\ln(e^2N/x)$ is increasing in $[0, eN]$ and that $U \ge 2$ in this case. The theorem then follows by combining the bounds provided for the two cases together with Equation (4.3). □

4.5 A Lower Bound for Restricted Advice via Feedback Graphs

In this section, we provide a novel lower bound on the minimax regret for a slightly harder formulation of the multi-armed bandit problem with expert advice. We consider a setting where the learner picks an expert I_t (possibly at random) instead of an action at the beginning of each round $t \in [T]$ without observing any of the experts' recommendations relative to the current round beforehand. The action A_t to be performed is subsequently drawn by the environment from the chosen expert's distribution, i.e., $A_t \sim \theta_t^{I_t}$. Afterwards, the learner observes A_t, the incurred loss $\ell_t(A_t)$, and only the advice θ_t^i of experts $i \in V$ that have the drawn action A_t in their support, i.e., $\theta_t^i(A_t) > 0$. For experts outside this set, the learner can only infer that, by definition, $\theta_t^i(A_t) = 0$. We will refer to this variation of the problem as the multi-armed bandit with *restricted* expert advice.[*] Observe that Algorithm 4.1 is still implementable in this scenario and guarantees a regret upper bound of order $\sqrt{\xi T\,(1 + \ln(N/\xi))}$ for $\xi \coloneqq K \wedge N$, as previously analyzed. Here we show that the regret of Algorithm 4.1 is the best regret we can hope for, up to constant factors, for any number K of actions and any number N of experts. While a $\Omega(\sqrt{NT})$ regret lower bound in the case $N \le K$ is immediate as mentioned before, the following theorem provides an $\Omega\big(\sqrt{KT\ln(N/K)}\big)$ lower bound when $N > K$, improving upon the $\Omega\big(\sqrt{KT(\ln N)/(\ln K)}\big)$ lower bound of Seldin and Lugosi (2016).

In what follows, we fix $N > K \ge 2$. We derive the lower bound relying on a reduction from the multi-armed bandit problem with feedback graphs. In the latter setting, we assume there exists a graph $G = (V, E)$ over a finite set $V = [N]$ of actions from which the learner selects one action $J_t \in V$ at each round $t \in [T]$. Then, the learner observes the losses of the neighbours of J_t in G. For the construction of the lower bound, it suffices to assume that G is undirected and contains all self-loops, i.e., $(i, i) \in E$ for each $i \in V$. Consequently, the graph G is strongly observable and the

[*] This differs from the limited expert advice model studied by Kale (2014).

learner always observes the loss of the selected action. We particularly focus on a specific family of graphs (also considered in the recent work of Chen et al. (2024)) where the N vertices are partitioned into disjoint cliques with self-loops. Precisely, we let $M := \lfloor K/2 \rfloor \geq 1$ be the number of disjoint cliques in G. For any $k \in [M]$, let C_k be the set of vertices of the k-th clique in G. Since each C_k is a clique with all self-loops, we have that $(i, j) \in E$ if and only if $i, j \in C_k$ for some $k \in [M]$, and thus $E := \bigcup_{k \in [M]} (C_k \times C_k)$. Additionally, for our purposes, we only consider the partition into cliques $C_k := \big\{i \in [N] : i \equiv k \mod M\big\}$ of roughly the same size $|C_k| \geq \lfloor N/M \rfloor \geq \lfloor 2N/K \rfloor \geq N/K$.

Hence, we will focus on the class of instances, denoted by Ξ_{FG}, of the multi-armed bandit problem with feedback graphs where the graph assumes the particular structure described above. In particular, any instance $\mathcal{I} \in \Xi_{\mathrm{FG}}$ is defined as a tuple $\mathcal{I} := (T, G, \mathcal{L})$ containing the number T of rounds, the feedback graph $G = (V, E)$ over $V = [N]$ composed of the disjoint cliques $C_1, \dots, C_M$ as defined above, and the sequence $\mathcal{L} := (\ell_t)_{t \in [T]}$ of binary loss functions $\ell_t \colon V \to \{0, 1\}$ over V. On the other hand, we let Ξ_{BEA} be the class of instances for the multi-armed bandit problem with restricted expert advice, with N experts and K actions. An instance $\mathcal{I} \in \Xi_{\mathrm{BEA}}$ is a tuple $\mathcal{I} := \big(T, V, \mathcal{A}, \Theta, \mathcal{L}\big)$ containing the number T of rounds, the set $V = [N]$ of experts, the set $\mathcal{A} = [K]$ of actions, the sequence $\Theta := (\theta_t^i)_{i \in V, t \in [T]}$ of expert advice where $\theta_t^i \in \Delta_K$, and the sequence $\mathcal{L} := (\ell_t)_{t \in [T]}$ of loss functions $\ell_t \colon \mathcal{A} \to \{0, 1\}$ over $\mathcal{A}$. The sought result is established by showing that the worst-case regret of any algorithm against a particular subset of instances in Ξ_{BEA} is order-wise at least as large as the minimax regret on Ξ_{FG}, combined with a lower bound on the latter quantity by Chen et al. (2024).

Theorem 4.3. *Let $\mathcal{B}$ be any possibly randomized algorithm for the multi-armed bandit problem with restricted expert advice for any number $K \geq 2$ of actions $\mathcal{A} = [K]$ and any number $N > K$ of experts $V = [N]$. Then, for a sufficiently large T, there exist a sequence $\ell_1, \dots, \ell_T \colon \mathcal{A} \to \{0, 1\}$ of binary loss functions and a sequence $(\theta_t^i)_{i \in V, t \in [T]}$ of expert advice such that the expected regret of $\mathcal{B}$ is $\Omega\big(\sqrt{KT \ln(N/K)}\big)$.*

Proof. We first describe a reduction from the multi-armed bandit problem with feedback graphs to the multi-armed bandit problem with restricted expert advice. We accomplish this by providing a mapping $\rho \colon \Xi_{\mathrm{FG}} \to \Xi_{\mathrm{BEA}}$ from the considered instance class Ξ_{FG} of the former problem to the instance class Ξ_{BEA} of the latter.

Consider any instance $\mathcal{I} := (T, G, \mathcal{L}) \in \Xi_{\mathrm{FG}}$ and recall that $G = (V, E)$ is a union of $M = \lfloor K/2 \rfloor$ disjoint cliques $C_1, \dots, C_M$ over $V = [N]$. The mapped instance $\rho(\mathcal{I}) := (T, V, \mathcal{A}, \Theta, \mathcal{L}') \in \Xi_{\mathrm{BEA}}$ is defined over the same number of rounds T and an experts set corresponding to the actions V in the original instance $\mathcal{I}$, whose sequence of recommendations is provided by $\Theta = (\theta_t^i)_{i \in V, t \in [T]}$. We first observe that the cardinality of the new action set $\mathcal{A} = [K]$ does relate to the number of cliques M. In particular, considering the partition of experts given by the cliques in G, we also partition the actions (in the expert advice instance $\rho(\mathcal{I})$) by associating 2 actions to each clique. Precisely, for any $k \in [M]$, we associate actions $\mathcal{A}_k := \{2k - 1, 2k\}$ to C_k. If K is even, this partitions the entire set of actions $\mathcal{A}$, while it leaves out action K otherwise. We can ignore the latter case and assume K is even without loss of generality, since we can otherwise leave action K outside of the support of any expert advice $\theta_t^i \in \Delta_K$ in the following construction (thus becoming a spurious action).

Second, we focus on the construction of the loss sequence $\mathcal{L}' := (\ell'_1, \dots, \ell'_T)$. For any $t \in [T]$, we define $\ell'_t \in \{0,1\}^{\mathcal{A}}$ as

$$\ell'_t(2k-1) \coloneqq 0 \qquad \text{and} \qquad \ell'_t(2k) \coloneqq 1 \qquad \forall k \in [M] \,.$$

Finally, we define the sequence of expert advice $(\theta^i_t)_{i\in V, t\in[T]}$ depending on the sequence of losses $\mathcal{L}$ of the starting instance $\mathcal{I}$. For any $t \in [T]$, any $k \in [M]$, and any $i \in C_k$, we define $\theta^i_t \in \Delta_K$ as

$$\theta^i_t \coloneqq \begin{cases} \delta_{2k-1} & \text{if } \ell_t(i) = 0 \\ \delta_{2k} & \text{if } \ell_t(i) = 1 \end{cases} ,$$

where $\delta_j \in \Delta_K$ is the Dirac delta at $j \in \mathcal{A}$. This ensures that the loss of expert i at round t, given by $y_t(i) = \sum_{a\in\mathcal{A}} \theta^i_t(a)\ell'_t(a)$ coincides with $\ell_t(i)$, the loss of action i in the original feedback graphs instance at the same round. Moreover, the knowledge of $\ell_t(i)$ suffices to infer θ^i_t.

At this point, given our instance mapping ρ and our algorithm $\mathcal{B}$, we design an algorithm $\mathcal{B}_\rho$ for the class Ξ_{FG}. Consider any instance $\mathcal{I} \in \Xi_{\text{FG}}$. Over the interaction period, the algorithm $\mathcal{B}_\rho$, without requiring prior knowledge of $\mathcal{I}$, maintains a running realization of $\mathcal{B}$ on instance $\rho(\mathcal{I})$. At any round $t \in [T]$, let I_t be the expert selected by algorithm $\mathcal{B}$ in $\rho(\mathcal{I})$, and let $k_t \in [M]$ be the index of the clique I_t belongs to, i.e., $I_t \in C_{k_t}$. Algorithm $\mathcal{B}_\rho$, interacting with the instance $\mathcal{I}$, executes action $J_t = I_t$ provided by $\mathcal{B}$ and observes the losses $(\ell_t(i))_{i\in C_{k_t}}$. Then, thanks to the design of the mapping ρ, $\mathcal{B}_\rho$ can construct and provide $\mathcal{B}$ the feedback it requires and which complies with instance $\rho(\mathcal{I})$. Namely, it determines that $A_t = 2k_t - 1$ if $\ell_t(J_t) = 0$ or else that $A_t = 2k_t$, then passes A_t, its loss $\ell'_t(A_t)$ (trivially determined), and the restricted advice $(\theta^i_t)_{i\in C_{k_t}}$ to $\mathcal{B}$. The last of which is a super-set of the recommended distributions having positive support on A_t since A_t is never picked by experts outside C_{k_t} by construction. Now, let

$$R^{\mathcal{B}}(\mathcal{I}') \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell'_t(A_t)\right] - \min_{i\in V} \sum_{t=1}^T \sum_{a\in\mathcal{A}} \theta^i_t(a)\ell'_t(a) = \mathbb{E}\left[\sum_{t=1}^T y_t(I_t)\right] - \min_{i\in V} \sum_{t=1}^T \sum_{a\in\mathcal{A}} \theta^i_t(a)\ell'_t(a)$$

be the expected regret of algorithm $\mathcal{B}$ on some instance $\mathcal{I}' = \big(T, V, \mathcal{A}, (\theta^i_t)_{i\in V, t\in[T]}, (\ell'_t)_{t\in[T]}\big) \in \Xi_{\text{BEA}}$. Similarly, let

$$R^{\mathcal{B}_\rho}(\mathcal{I}) \coloneqq \mathbb{E}\left[\sum_{t=1}^T \ell_t(J_t)\right] - \min_{i\in V} \sum_{t=1}^T \ell_t(i)$$

be the expected regret of algorithm $\mathcal{B}_\rho$ on some instance $\mathcal{I} = \big(T, G, (\ell_t)_{t\in[T]}\big) \in \Xi_{\text{FG}}$. Since $J_t = I_t$, we have that $y_t(I_t) = \ell_t(J_t)$ via the properties of ρ laid out before. Hence, we can conclude that $R^{\mathcal{B}}(\rho(\mathcal{I})) = R^{\mathcal{B}_\rho}(\mathcal{I})$ for any instance $\mathcal{I} \in \Xi_{\text{FG}}$. Define $\rho(\Xi_{\text{FG}}) \coloneqq \big\{\rho(\mathcal{I}) : \mathcal{I} \in \Xi_{\text{FG}}\big\} \subseteq \Xi_{\text{BEA}}$ as the subclass of instances in Ξ_{BEA} obtained from Ξ_{FG} via ρ. Then, it holds that

$$\sup_{\mathcal{I}\in\Xi_{\text{BEA}}} R^{\mathcal{B}}(\mathcal{I}) \geq \sup_{\mathcal{I}\in\rho(\Xi_{\text{FG}})} R^{\mathcal{B}}(\mathcal{I}) = \sup_{\mathcal{I}\in\Xi_{\text{FG}}} R^{\mathcal{B}}(\rho(\mathcal{I})) = \sup_{\mathcal{I}\in\Xi_{\text{FG}}} R^{\mathcal{B}_\rho}(\mathcal{I}) \,.$$

On the other hand, Lemma E.1 in Chen et al. (2024) implies that

$$\sup_{\mathcal{I}\in\Xi_{\text{FG}}} R^{\mathcal{B}_\rho}_T(\mathcal{I}) = \Omega\left(\sqrt{T \sum_{k\in[M]} \ln(1 + |C_k|)}\right) = \Omega\big(\sqrt{KT\ln(N/K)}\big)$$

for sufficiently large T since $\sum_{k\in[M]} \ln(1+|C_k|) \geq M\ln(N/M) \geq K\ln(2N/K)/4$, thus concluding the proof. □

4.6 Conclusions

As the lower bound of Theorem 4.3 was proved for a harder formulation of the problem, it remains to be shown whether the same impossibility result holds for the more standard setup. We conjecture it should be possible to prove such a lower bound. If it indeed holds, this would imply that the minimax regret in the two variants is of the same order; that is, as far as we are only concerned with the worst-case regret, the standard feedback setup would be shown to be essentially as hard as the restricted one.

Chapter 5

Online Learning with Stochastic Feedback Graphs

We consider once more the framework of online learning with feedback graphs. Here we study an extension where the graph directed and stochastic, following a distribution similar to the heterogeneous Erdős-Rényi model: at every round, each edge in the graph independently realizes with some unknown edge-specific probability. We prove nearly optimal regret bounds depending on graph-theoretic quantities measured on the support of the stochastic feedback graph with thresholded edge probabilities. Our results hold without any preliminary knowledge about the graphs distribution. When the learner is allowed to observe the entire realized graph, we derive a more efficient algorithm that exhibits improved bounds in some cases, featuring a dependence on weighted versions of the same graph parameters.

5.1 Introduction

In this chapter, we move back to the feedback graphs setting by considering a meaningful extension of its framework. If the aim of Chapter 3 was the achievement of improved (and actually worst-case optimal up to constants) regret guarantees, which also resulted in techniques and ideas that led to improved regret upper and lower bounds for bandits with expert advice in Chapter 4, here we introduce a more general and harder feedback model for partial loss observability. In this setting, the loss of any action in all decision rounds is preliminarily chosen by an oblivious adversary as usual, but the feedback is *probabilistically* received by the learner at the end of each round t. More precisely, the loss $\ell_t(i)$ of each action $i \in V$ (including the action I_t selected by the learner at round t) is independently observed with a certain probability $p(I_t, i)$, where the probabilities $p(i, j) \in [0, 1]$ for all pairs $i, j \in V$ are fixed but unknown.

This feedback model can be viewed as a probabilistic feedback version of the plain graph feedback model for online learning as we considered so far, where the feedback received by the learner is determined by a *deterministic* directed graph defined over the set of actions. In the standard model, the learner deterministically observes the losses of all the actions in the out-neighborhood of the one selected in a specific round. In certain applications, however, deterministic feedback is not realistic. Consider for instance a sensor network for monitoring the environment, where the learner can decide which sensor to probe in order to maximize some performance measure. Each probed sensor may also receive readings from other sensors, but whether a sensor successfully transmits information to another sensor depends on a number of uncontrollable environmental factors, which include the positions of the sensors, as well as their internal state (e.g., battery levels), the weather conditions, and so on. Due to the variability of some of these factors, the possibility of reading from another sensor can be naturally modeled as a stochastic event.

Online learning with adversarial losses and *stochastic feedback graphs* has been studied before, but under fairly restrictive assumptions on the probabilities $p(i,j)$. Let $\mathcal{G}$ be a stochastic feedback graph, represented by its probability matrix $\big[p(i,j)\big]_{i,j\in V} \in [0,1]^{V\times V}$ where $V := [K]$ is the action set (for $K \geq 2$). When $p(i,j) = \varepsilon$ for all distinct $i,j \in V$ and for some $\varepsilon > 0$, then $\mathcal{G}$ follows the Erdős-Rényi random graph model. Under the assumption that ε is known and $p(i,i) = 1$ for all $i \in V$ (all self-loops occur with probability 1), Alon et al. (2017) show that the optimal regret after T rounds is of order $\sqrt{T/\varepsilon}$, up to logarithmic factors. This result has been extended by Kocák, Neu, and Valko (2016a), who prove a regret bound of order $\sqrt{\sum_{t=1}^{T}(1/\varepsilon_t)}$ when the parameter ε_t of the random graph is unknown and allowed to change over time. However, their result holds only under rather strong assumptions on the sequence $(\varepsilon_t)_{t\in[T]}$ of probabilities. In a recent work, Ghari and Shen (2022, 2024) show a regret bound of order $(\alpha/\varepsilon)\sqrt{KT}$, ignoring logarithmic factors, when each (unknown) probability $p(i,j)$ in $\mathcal{G}$ is either zero or at least ε for some known $\varepsilon > 0$, while all self-loops (i,i) are guaranteed to have probability $p(i,i) \geq \varepsilon$. Here α is the independence number (computed ignoring edge orientations) of the support graph $\operatorname{supp}(\mathcal{G})$, i.e., the directed graph whose adjacency matrix $A \in \{0,1\}^{V\times V}$ is defined so that $A(i,j) := \mathbb{I}\{p(i,j) > 0\}$. Their bound holds under the assumption that $\operatorname{supp}(\mathcal{G})$ is preliminarily known to the learner.

Our analysis does away with a crucial assumption that was key to prove all previous results. Namely, we do not assume any special property of the matrix $\mathcal{G}$, and we do not require the learner to have any preliminary knowledge of this matrix. The fact that positive edge probabilities are not bounded away from zero implies that the learner must *adaptively choose* a threshold $\varepsilon \in (0,1]$ below which the edges are deemed to be too rare to be exploitable for learning. Indeed, waiting for the realization of rare edges slows down learning if ε is too small. On the other hand, when ε is too large, then the feedback becomes sparse and the regret increases.

To formalize the intuition behind rare edges, we introduce the notion of thresholded graph $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ for any $\varepsilon > 0$.[*] This is the directed graph with adjacency matrix A such that $A(i,j) := \mathbb{I}\{p(i,j) \geq \varepsilon\}$ for each $i,j \in V$. As the thresholded graph is a deterministic feedback graph G, we can refer to Proposition 2.1 from Chapter 2 (and more generally to Alon et al. (2015)) for a characterization of the minimax regret R_T based on whether G is not observable (R_T of order T), weakly observable (R_T of order $\delta^{1/3}T^{2/3}$), or strongly observable (R_T of order $\sqrt{\alpha T}$).[†] We remind the reader that α and δ in this context are, respectively, the independence and the weak domination number of G (see Definitions 2.3 and 2.4). Let α_ε and δ_ε respectively denote the independence number and the weak domination number of $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ for any $\varepsilon > 0$. As α_ε and δ_ε both grow when ε gets larger, the ratios $\alpha_\varepsilon/\varepsilon$ and $\delta_\varepsilon/\varepsilon$ capture the trade-off involved in choosing the threshold ε. We define the optimal values for ε as follows:

$$\varepsilon_s^* := \operatorname*{arg\,min}_{\varepsilon\in(0,1]} \left\{ \frac{\alpha_\varepsilon}{\varepsilon} \ : \ \operatorname{supp}([\mathcal{G}]_\varepsilon) \text{ is strongly observable} \right\} , \tag{5.1}$$

[*] This notation is more precisely defined in Section 5.2.

[†] All these rates ignore logarithmic factors.

$$\varepsilon_w^* := \operatorname*{arg\,min}_{\varepsilon \in (0,1]} \left\{ \frac{\delta_\varepsilon}{\varepsilon} \ : \ \operatorname{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ is observable} \right\} . \tag{5.2}$$

We adopt the convention that the minimum of an empty set is infinity and the relative arg min is set to 0. The arg min's are well defined: there are at most K^2 values of ε for which the support of $[\mathcal{G}]_\varepsilon$ varies, and the minimum of each ratio is attained in one of these values. For simplicity, we let $\alpha^* := \alpha_{\varepsilon_s^*}$ and $\delta^* := \delta_{\varepsilon_w^*}$. Our first result can be informally stated as follows.

Theorem 5.1 (Informal). *Consider the problem of online learning with an unknown stochastic feedback graph $\mathcal{G}$ on T rounds. If* $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ *is not observable for $\varepsilon = \widetilde{\Theta}(K^3/T)$, then any learning algorithm suffers regret linear in T. Otherwise, there exists an algorithm whose expected regret, ignoring logarithmic factors in K and T, satisfies*

$$R_T \lesssim \min\left\{ \sqrt{\frac{\alpha^*}{\varepsilon_s^*} T}, \left(\frac{\delta^*}{\varepsilon_w^*}\right)^{1/3} T^{2/3} \right\} .$$

This bound is tight (up to logarithmic factors).

This result shows that, without any preliminary knowledge of $\mathcal{G}$, we can obtain a bound that optimally trades off between the strongly observable rate $\sqrt{(\alpha^*/\varepsilon_s^*)T}$, for the best threshold ε for which $\operatorname{supp}\left([\mathcal{G}]_\varepsilon\right)$ is strongly observable, and the (weakly) observable rate $(\delta^*/\varepsilon_w^*)^{1/3}T^{2/3}$, for the best threshold ε for which $\operatorname{supp}\left([\mathcal{G}]_\varepsilon\right)$ is (weakly) observable. Note that this result improves on Ghari and Shen (2022, 2024) bound $(\alpha_\varepsilon/\varepsilon)\sqrt{KT}$, who additionally assume that $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ and ε (a lower bound on the self-loop probabilities) are both preliminarily available to the learner. On the other hand, the algorithm achieving the bound of Theorem 5.1 need not receive any information (neither prior nor during the learning process) besides the stochastic feedback.

We obtain positive results in Theorem 5.1 via an elaborate reduction to online learning with deterministic feedback graphs. Our algorithm works in two phases: first, it learns the edge probabilities in a round-robin procedure, then it commits to a carefully chosen estimate of the feedback graph and feeds its support to an algorithm for online learning with deterministic feedback graphs. There are two main technical challenges the algorithm faces: on the one hand, it needs to switch from the first to the second phase at the right time in order to achieve the near-optimal regret. On the other hand, in order for the reduction to work, it needs to simulate the behaviour of a deterministic feedback graph using only feedback from a stochastic feedback graph (with unknown edge probabilities). We complement the results in Theorem 5.1 with matching lower bounds that are achieved by a suitable modification of the hard instances in Alon et al. (2015, 2017) so as to consider stochastic feedback graphs.

Our final result in the current chapter is a second algorithm that, at the cost of an additional assumption on the feedback (i.e., the learner additionally observes the realization of the entire feedback graph at the end of each round), has regret which is never worse and may be considerably better than the regret of the algorithm in Theorem 5.1. While the bounds in Theorem 5.1 are tight up to logarithmic factors, we show that the factors α^*/ε_s^* and δ^*/ε_w^* can be improved for specific feedback graphs. Specifically, we design weighted versions of the independence and weak domination numbers, where the weights of a given node depend on the probabilities of seeing the loss of that node. On the technical side, we design a new importance-weighted estimator which uses a particular version of upper-confidence estimates of the edge probabilities $p(i, j)$, rather than the true edge probabilities, which are unknown. We prove that the cost of using this estimator is of the same order as the regret bound achievable had we known $p(i,j)$. Additionally, the algorithm that obtains these improved bounds is more efficient than the algorithm of Theorem 5.1. The improvement in efficiency comes from the following idea: we start with an optimistic algorithm that assumes that the support of $\mathcal{G}$ is strongly observable and only switches to the assumption that the support of $\mathcal{G}$ is (weakly) observable when it estimates that the regret under this second assumption is smaller. The algorithm learns which regime is better by keeping track of a bound on the regret of the optimistic algorithm while simultaneously estimating the regret in the (weakly) observable case, which it can do efficiently.

5.1.1 Related Work

The results of Alon et al. (2015) for adversarial online learning with feedback graphs—also based on prior work by Alon et al. (2013), Kocák et al. (2014)—have been more recently improved by Chen et al. (2021), with tighter graph-theoretic constants in the regret bound. Variants of the adversarial setting have been studied by Feng and Loh (2018), Arora, Marinov, and Mohri (2019), Rangi and Franceschetti (2019) and Van der Hoeven, Fusco, and Cesa-Bianchi (2021), who study online learning with feedback graphs and switching costs, and online multi-class classification with feedback graphs. There is also a considerable amount of work in stochastic enviromnents (Liu, Buccapatnam, and Shroff, 2018, Cortes, DeSalvo, Gentile, Mohri, and Yang, 2019, Li, Chen, Wen, and Leung, 2020). Finally, Rouyer, Van der Hoeven, Cesa-Bianchi, and Seldin (2022b) and Ito et al. (2022) independently designed different best-of-both-worlds learning algorithms achieving nearly optimal (up to logarithmic factors in T) regret bounds in the adversarial and stochastic settings. It is therefore clear that bandits with feedback graphs have been extended and investigated in multiple flavours in the past, even in very recent work.

As per the initial definition of the model that we introduced in Chapter 2 and considered at some point in Chapter 3 too, some lines of work focus on scenarios where the feedback graph is not fixed but changes over time, resulting in a sequence $G_1, \dots, G_T$ of feedback graphs. Specifically, Cohen et al. (2016) study a setting where the graphs are adversarially chosen and only the local structure of the feedback graph is observed. They show that, if the losses are generated by an adversary and all nodes always have a self-loop, one cannot do better than $\sqrt{KT}$ regret, and we might as well simply employ a standard multi-armed bandit algorithm. Furthermore, removing the guarantee on the self-loops induces an $\Omega(T)$ regret. In Section 5.3, we are in a similar situation, as we also observe only local information about the feedback graph and the losses are generated by an adversary. However, we show that if the graphs are stochastically generated with a strongly observable support for some threshold ε, there is a $\sqrt{\alpha T/\varepsilon}$ regret bound. As a consequence, for ε not too small, observing only the local information about the feedback graphs is in fact sufficient to obtain better results than in the bandit setting. Similarly, if there are no self-loops in the support but the support is weakly observable, then our regret bounds are sublinear rather than linear in T. Alon et al. (2013, 2017) and Kocák et al. (2014) also consider adversarially generated sequences $G_1, G_2, \dots$ of deterministic feedback graphs. In the case of directed feedback graphs, Alon et al. (2013) investigate a model in which G_t is revealed to the learner at the beginning of each round t. Alon et al. (2017) and Kocák et al. (2014) extend this analysis to the case when G_t is strongly observable and made available only at the end of each round t. These are precisely the informed and uninformed settings, as we briefly mentioned in Chapter 2. In comparison, in our setting the graphs (or the local information about the graph) revealed to the learner (at the end of each round) may not even be observable, let alone strongly observable. Despite this seemingly challenging setting for previous works, we nevertheless obtain sublinear regret bounds. Finally, Kocák, Neu, and Valko (2016b) study a feedback model where the losses of other actions in the out-neighborhood of the action played are observed with an edge-dependent noise. In their setting, the feedback graphs G_t are weighted and revealed at the beginning of each round. They introduce edge weights $s_t(i,j) \in [0,1]$ that determine the feedback according to the following additive noise model: $s_t(I_t,j)\ell_t(j) + (1 - s_t(I_t,j))\xi_t(j)$, where $\xi_t(j)$ is a zero-mean bounded random variable. Hence, if $s_t(i,j) = 1$, then $I_t = i$ allows to observe the loss of action j without any noise. If $s_t(i,j) = 0$, then only noise is observed. Note that they assume $s_t(i,i) = 1$ for every action i, implying strong observability. Although somewhat similar in spirit to our feedback model, our results do not directly compare with theirs.

Further work also takes into account a time-varying probability for the revelation of side-observations (Kocák et al., 2016a). While the idea of a general probabilistic feedback graph has been already considered in the stochastic setting (Li et al., 2020, Cortes, DeSalvo, Gentile, Mohri, and Zhang, 2020), the recent work by Ghari and Shen (2022, 2024) seems to be the first one in the adversarial setting that generalizes from the Erdős-Rényi model to a more flexible distribution where they allow "edge-specific" probabilities. We remark, however, that the assumptions of Ghari and Shen (2022, 2024) exclude some important instances of feedback graphs. For example, we cannot hope to employ their algorithm for efficiently solving the revealing action problem (see for example Alon et al. (2015)). In a spirit similar to ours, Resler and Mansour (2019) studied a version of the problem where the topology of the graph is fixed and known a priori, but the feedback received by the learner is perturbed when traversing edges.

5.2 Problem Setting and Notations

A feedback graph over a set $V := [K]$ of actions is any directed graph $G := (V,E)$, possibly with self-loops, contrarily to the more restricted case of undirected graphs assumed thus far. For any vertex $i \in V$, we adopt the already introduced notation for the in-neighborhood $N^{\mathrm{in}}_G(i) = \{j \in V : (j,i) \in E\}$ and the out-neighborhood $N^{\mathrm{out}}_G(i) = \{j \in V : (i,j) \in E\}$ of i; we may omit the subscript when the graph is clear from the context.

In the online learning problem with a stochastic feedback graph, an oblivious adversary privately chooses a stochastic feedback graph $\mathcal{G}$ (i.e., the distribution of the feedback graphs) and a sequence $\ell_1, \ell_2, \dots$ of loss functions $\ell_t \colon V \to [0,1]$. At each round $t = 1, 2, \dots$, the learner selects an action $I_t \in V$ to play and, independently, the adversary draws a feedback graph G_t from $\mathcal{G}$ (denoted by $G_t \sim \mathcal{G}$). The learner then incurs loss $\ell_t(I_t)$ and observes the feedback $\big\{(i, \ell_t(i)) : i \in N^{\mathrm{out}}_{G_t}(I_t)\big\}$. In some cases we consider a richer feedback, where at the end of each round t the learner also observes the entire realized graph G_t. The learner's performance is measured using the standard notion of regret which, we recall, is defined as

$$R_T = \max_{k \in V} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(I_t) - \ell_t(k)\big)\right] ,$$

where $I_1, \dots, I_T$ are the actions played by the learner, and the expectation is computed over both the

Algorithm 5.1: RoundRobin

```
environment: stochastic feedback graph $\mathcal{G}$, sequence of losses $\ell_1, \ell_2, \dots, \ell_T$
input: time horizon $T$, stopping function $\Phi$, actions $V = [K]$
$n_e \leftarrow 0$, for all $e \in V^2$
for each $\tau = 1, 2, \dots, \lfloor T/K \rfloor$ do
    for each $i = 1, 2, \dots K$ do
        play action $i$ and observe $N^{\text{out}}_{G_t}(i)$ from $G_t \sim \mathcal{G}$        ▷ $t$ is the time step
        $n_e \leftarrow n_e + 1$ for all $e \in N^{\text{out}}_{G_t}(i)$
    $\widehat{p}^{\tau}_e \leftarrow n_e/\tau$ for all edges $e \in V^2$
    $\varepsilon_\tau \leftarrow 60 \ln(KT)/\tau$
    $\widehat{\mathcal{G}}_\tau \leftarrow \big(V, \{e \in V^2 : \widehat{p}^{\tau}_e \geq \varepsilon_\tau\}\big)$ with weights $\widehat{p}^{\tau}_e$        ▷ estimated feedback graph
    if $\Phi(\widehat{\mathcal{G}}_\tau, T) \leq \tau K$ then
        output $\widehat{\mathcal{G}}_\tau$, $\varepsilon_\tau$
output $\widehat{\mathcal{G}}_\tau$, $\varepsilon_\tau$
```

sequence $G_1, \dots, G_T$ of feedback graphs drawn i.i.d. from $\mathcal{G}$ and the learner's internal randomization.

Fix any stochastic feedback graph $\mathcal{G} \coloneqq \{p(i,j) : i,j \in V\}$,[‡] implicitly described by edge probabilities $[p(i,j)]_{i,j \in V} \in [0,1]^{V \times V}$. We sometimes use e to denote a pair (i,j), in which case we write p_e to denote the probability $p(i,j)$. When $G_t \coloneqq (V, E_t)$ is drawn from $\mathcal{G}$, each pair $(i,j) \in V \times V$ independently becomes an edge (i.e., $(i,j) \in E_t$) with probability $p(i,j)$. For any threshold $\varepsilon > 0$, we define the *thresholding* $[\mathcal{G}]_\varepsilon$ of $\mathcal{G}$ as the stochastic graph represented by $\{p'(i,j) : i,j \in V\}$, where $p'(i,j) \coloneqq p(i,j)\mathbb{I}\{p(i,j) \geq \varepsilon\}$. We also define the *support graph* of $\mathcal{G}$ as the (deterministic) graph $\operatorname{supp}(\mathcal{G}) \coloneqq (V, E)$ having $E \coloneqq \{(i,j) \in V \times V : p(i,j) > 0\}$. To keep the notation tidy, we write $\alpha(\mathcal{G})$ instead of $\alpha(\operatorname{supp}(\mathcal{G}))$ and similarly for δ.

5.3 Block Decomposition Approach

In this section, we present an algorithm for online learning with stochastic feedback graphs via a reduction to online learning with deterministic feedback graphs. Our algorithm EdgeCatcher (Algorithm 5.3) has an initial exploration phase followed by a commit phase. In the exploration phase, the edge probabilities are learned online in a round-robin fashion. A carefully designed stopping criterion then triggers the commit phase, where we feed the support of the estimated stochastic feedback graph to an algorithm for online learning with (deterministic) feedback graphs.

5.3.1 Estimating the Edge Probabilities

As a first step we design a routine, RoundRobin (Algorithm 5.1), that sequentially estimates the stochastic feedback graph until a certain stopping criterion is met. The stopping criterion depends on a non-negative function Φ that takes as input a stochastic feedback graph $\mathcal{G}$ together with a time horizon. Let $\widehat{\tau} \leq T/K$ be the index of the last iteration of the outer for-loop in Algorithm 5.1. We want to make sure that, for all $\tau \leq \widehat{\tau}$, the stochastic feedback graphs $\widehat{\mathcal{G}}_\tau$ are valid estimates of the underlying $\mathcal{G}$ up to a $\Theta(\varepsilon_\tau)$ precision. To formalize this notion of approximation, we introduce the following definition.

[‡] From now on, we may slightly abuse the set notation to denote the collection of edge probabilities.

Definition 5.1 (ε-good approximation)**.** *A stochastic feedback graph $\widehat{\mathcal{G}} := \{\widehat{p}_e : e \in V^2\}$ is an ε-good approximation of $\mathcal{G} := \{p_e : e \in V^2\}$ for some $\varepsilon \in (0, 1]$, if the following holds:*

1. *all the edges $e \in \operatorname{supp}(\mathcal{G})$ with $p_e \geq 2\varepsilon$ belong to $\operatorname{supp}(\widehat{\mathcal{G}})$;*
2. *for all edges $e \in \operatorname{supp}(\widehat{\mathcal{G}})$ with $p_e \geq \varepsilon/2$ it holds that $|\widehat{p}_e - p_e| \leq p_e/2$;*
3. *no edge $e \in V^2$ with $p_e < \varepsilon/2$ belongs to $\operatorname{supp}(\widehat{\mathcal{G}})$.*

We can now state the following theorem; we defer the proof in Appendix C.2. The proof follows from an application of the multiplicative Chernoff bound on edge probabilities.

Theorem 5.2. *If* RoundRobin *(Algorithm 5.1) is run on the stochastic feedback graph $\mathcal{G}$, then, with probability at least $1 - 1/T$, the estimate $\widehat{\mathcal{G}}_\tau$ is an ε_τ-good approximation of $\mathcal{G}$ simultaneously for all $\tau \leq \widehat{\tau}$, where $\widehat{\tau} \leq T/K$ is the index of the last iteration of the outer for-loop in Algorithm 5.1.*

5.3.2 Reduction to Deterministic Feedback Graphs

As a second step, we present BlockReduction (Algorithm 5.2) which reduces the problem of online learning with stochastic feedback graph to the corresponding problem with deterministic feedback graph. Surprisingly enough, in order for this reduction to work, we do not need the exact edge probabilities: an ε-good approximation is sufficient for this purpose.

The intuition behind BlockReduction is simple: given that each edge e in $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ appears in G_t with probability $p_e \geq \varepsilon$ at each time step t, if we wait for $\Theta\big((1/\varepsilon)\ln T\big)$ time steps it will appear at least once with high probability. Applying a union bound over all edges, we can argue that considering $\Delta = \Theta\big((1/\varepsilon)\ln(KT)\big)$ realizations of the stochastic feedback graph, then all the edges in $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ are realized at least once with high probability.

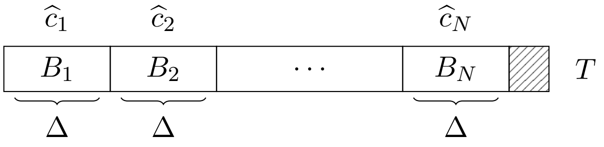


Figure 5.1: Illustration for the blocks reduction with blocks $B_1, \dots, B_N$ each of size Δ, with potential remainder rounds outside any block.

Imagine now to play a certain action a consistently during a block B_τ of Δ consecutive rounds. We want to reconstruct the average loss suffered by a' in B_τ:

$$c_\tau(a') := \sum_{t \in B_\tau} \frac{\ell_t(a')}{\Delta} \,, \tag{5.3}$$

and we want to do it for all actions a' in the out-neighborhood of a. Let $\Delta^\tau_{(a,a')}$ be the number of times that the loss of a' is observed by the learner within block B_τ; i.e., the number of times that (a, a') is realized in the Δ rounds. With this notation in mind, we can define the natural estimator $\widehat{c}_\tau(a')$:

$$\widehat{c}_\tau(a') := \sum_{t \in B_\tau} \ell_t(a') \frac{\mathbb{I}\{(a, a') \in E_t\}}{\Delta^\tau_{(a,a')}} \,. \tag{5.4}$$

Conditioning on the event $\mathcal{E}^\tau_{(a,a')}$ that the edge (a, a') in $\widehat{G}$ is observed at least once in block B_τ, we show in Lemma C.1 in Appendix C.2 that $\widehat{c}_\tau(a')$ is an unbiased estimator of $c_\tau(a')$. The overall idea

Algorithm 5.2: BlockReduction

```
environment: stochastic feedback graph $\mathcal{G}$, sequence of losses $\ell_1, \ell_2, \dots, \ell_T$
input: time horizon $T$, threshold $\varepsilon$, estimate $\widehat{\mathcal{G}}$ of $\mathcal{G}$, learning algorithm $\mathcal{A}$
$\Delta \leftarrow \lceil \frac{2}{\varepsilon} \ln(KT) \rceil$, $N \leftarrow \lfloor T/\Delta \rfloor$, $\widehat{G} \leftarrow \operatorname{supp}(\widehat{\mathcal{G}})$
initialize: $\mathcal{A}$ with time horizon $N$ and graph $\widehat{G}$
$B_\tau \leftarrow \{(\tau-1)\Delta+1, \dots, \tau\Delta\}$, for all $\tau = 1, \dots, N$
for each round $\tau = 1, 2, \dots, N$ do
    let $p_\tau$ be the probability distribution over actions output by $\mathcal{A}$
    draw action $a_\tau \sim p_\tau$
    for each round $t \in B_\tau$ do
        play action $a_\tau$ and observe the revealed feedback          ▷ $G_t \sim \mathcal{G}$
    for all $a' \in N^{\mathrm{out}}_{\widehat{G}}(a_\tau)$ do compute $\widehat{c}_\tau(a')$ as in (5.4), and feed them to $\mathcal{A}$
play arbitrarily in the remaining $T - \Delta N$ rounds
```

of the blocks reduction is briefly illustrated in Figure 5.1.

Therefore, we can construct conditionally unbiased estimators of the average losses over the blocks as if the stochastic feedback graph were deterministic. This allows us to reduce the original problem to that of online learning with deterministic feedback graph on the meta-instance given by the blocks. The details of BlockReduction are reported in Algorithm 5.2, while the theoretical properties are summarized in the next result, whose proof can be found in Appendix C.2.

Theorem 5.3. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$, and let $\widehat{\mathcal{G}}$ be an ε-good approximation of $\mathcal{G}$. Let $\mathcal{A}$ be an algorithm for online learning with arbitrary deterministic feedback graph G with regret bound $R^{\mathcal{A}}_N(G)$ over any sequence of N losses in $[0,1]$. Then, the regret of* BlockReduction *(Algorithm 5.2) run with input $(T, \varepsilon/2, \widehat{\mathcal{G}}, \mathcal{A})$ is at most $\Delta R^{\mathcal{A}}_N\big(\operatorname{supp}(\widehat{\mathcal{G}})\big) + \Delta$, where $N := \lfloor T/\Delta \rfloor$ and $\Delta := \lceil \frac{4}{\varepsilon} \ln(KT) \rceil$.*

For online learning with deterministic feedback graphs we use the variants of the well-known Exp3.G algorithm proposed by Alon et al. (2015). Together with Theorem 5.3, this gives the following corollary; the details of the proof are in Appendix C.2.

Corollary 5.1. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$, and let $\widehat{\mathcal{G}}$ be an ε-good approximation of $\mathcal{G}$ for $\varepsilon \geq 1/T$ and with support $\widehat{G}$. The following statements hold:*

- *If $\widehat{G}$ is strongly observable with independence number α, then the regret of* BlockReduction *run with parameter $\varepsilon/2$ using* Exp3.G *for strongly observable graphs as base algorithm $\mathcal{A}$ satisfies: $R_T \leq 4C_s\sqrt{(\alpha/\varepsilon)T} \cdot \ln^{3/2}(KT)$, where $C_s > 0$ is a constant in the regret bound of $\mathcal{A}$.*
- *If $\widehat{G}$ is (weakly) observable with weak domination number δ, then the regret of* BlockReduction *run with parameter $\varepsilon/2$ using* Exp3.G *for weakly observable graphs as base algorithm $\mathcal{A}$ satisfies: $R_T \leq 4C_w(\delta/\varepsilon)^{1/3}T^{2/3}\ln^{2/3}(KT)$, where $C_w > 0$ is a constant in the regret bound of $\mathcal{A}$.*

Note that we can explicitly compute valid constants $C_s = 12 + 2\sqrt{2}$ and $C_w = 8$ directly from the proofs of the main results by Alon et al. (2015).

5.3.3 Explore then Commit to a Graph

We are now ready to combine the two routines we presented, RoundRobin and BlockReduction, in our final online learning algorithm, EdgeCatcher (Algorithm 5.3). EdgeCatcher has two

Algorithm 5.3: EdgeCatcher

```
environment: stochastic feedback graph $\mathcal{G}$, sequence of losses $\ell_1, \ell_2, \dots, \ell_T$
input: time horizon $T$ and actions $V = [K]$
let $\Phi$ be defined as in Equation (5.5)
run RoundRobin$(T, \Phi, V)$ and obtain $\widehat{\mathcal{G}}$ and $\widehat{\varepsilon}$
compute $\widehat{\varepsilon}_s^*$ and $\widehat{\varepsilon}_w^*$ for graph $\widehat{\mathcal{G}}$ as in Equations (5.1) and (5.2)
let $\widehat{\varepsilon}^*$ be the best threshold as in Equation (5.5)
if $\widehat{\varepsilon}^* = \widehat{\varepsilon}_s^*$ then
    let $\mathcal{A}$ be Exp3.G for strongly observable feedback graph
else
    let $\mathcal{A}$ be Exp3.G for weakly observable feedback graph
let $T' := T - \widehat{\tau}K$ be the remaining time steps        ▷ $\widehat{\tau}$ as in RoundRobin
run BlockReduction$(T', \widehat{\varepsilon}^*/2, [\widehat{\mathcal{G}}]_{\widehat{\varepsilon}^*}, \mathcal{A})$
```

phases: in the first phase, RoundRobin is used to quickly obtain an ε-good approximation $\widehat{\mathcal{G}}$ of the underlying feedback graph $\mathcal{G}$, for a suitable ε. In the second phase, the algorithm commits to $\widehat{\mathcal{G}}$ and feeds it to BlockReduction. The crucial point is when to commit to a certain (estimated) stochastic feedback graph. If we commit too early, we might not observe a denser support graph, which implies missing out on a richer feedback. If we wait for too long, then the exploration phase ends up dominating the regret in a suboptimal way. To balance this trade-off, we use the stopping function Φ. This function takes as input a probabilistic feedback graph together with a time horizon and outputs the regret bound that BlockReduction would guarantee on this pair. It is defined as

$$\Phi(\mathcal{G}, T) := \min\left\{4C_s\sqrt{\frac{\alpha^*}{\varepsilon_s^*}T \cdot \ln^3(KT)}\,,\ 4C_w\left(\frac{\delta^*}{\varepsilon_w^*}\ln^2(KT)\right)^{1/3} T^{2/3}\right\} \tag{5.5}$$

for the specific choice of Exp3.G as the learning algorithm $\mathcal{A}$ adopted by BlockReduction. Note that the dependence of Φ on the feedback graph $\mathcal{G}$ is contained in the topological parameters α^* and δ^* and the corresponding thresholds ε_s^* and ε_w^*, defined in Equations (5.1) and (5.2); see Appendix C.1 for more details on their computation. If we apply Φ to a stochastic feedback graph that is observable w.p. zero, its value is conventionally set to infinity. Observe that, otherwise, the minimum is achieved for a specific ε^* and a specific $\mathcal{G}^* = [\mathcal{G}]_{\varepsilon^*}$.

In Appendix C.2, we provide a sequence of lemmas (Lemmas C.2 and C.3 in particular) showing that, if RoundRobin outputs an ε-good approximation of the graph, then the regret is bounded by a multiple of the stopping criterion evaluated at $\mathcal{G}$. Combined with Theorem 5.2, which tells us that RoundRobin does in fact output an ε-good approximation of the graph with high probability, this proves our main result for this section as stated in Theorem 5.4 below.

Theorem 5.4. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$ on T time steps. If* $\operatorname{supp}\left([\mathcal{G}]_{\varepsilon(K,T)}\right)$ *is observable for $\varepsilon(K,T) := CK^3(\ln(KT))^2/T$ for a given constant $C > 0$, then there exists an algorithm whose regret R_T, ignoring logarithmic factors in K and T, satisfies* $R_T \lesssim \min\left\{\sqrt{(\alpha^*/\varepsilon_s^*)T}, \left(\delta^*/\varepsilon_w^*\right)^{1/3}T^{2/3}\right\}$.

5.4 Lower Bounds

In this section, we provide lower bounds that match the regret bound guaranteed by EdgeCatcher, shown in Theorem 5.4, up to logarithmic factors in K and T. These lower bounds are valid even if the learner is allowed to observe the realization of the entire feedback graph at every time step, and knows a priori the "correct" threshold ε to work with. Theorem 5.5 summarizes the lower bounds whose proofs can be found in Appendix C.3.

Theorem 5.5 (Informal). *Let $\mathcal{A}$ be a possibly randomized algorithm for the online learning problem with stochastic feedback graphs. Consider any directed graph $G = (V, E)$ with $|V| \geq 2$ and any $\varepsilon \in (0, 1]$. There exists a stochastic feedback graph $\mathcal{G}$ with* $\operatorname{supp}(\mathcal{G}) = G$ *and, for a sufficiently large time horizon T, there is a sequence $\ell_1, \dots, \ell_T$ of loss functions on which the expected regret of $\mathcal{A}$ with respect to the stochastic generation of $G_1, \dots, G_T \sim \mathcal{G}$ is*

- $\Omega(\sqrt{(\alpha_\varepsilon/\varepsilon)T})$ *if G is strongly observable,*
- $\widetilde{\Omega}((\delta_\varepsilon/\varepsilon)^{1/3}T^{2/3})$ *if G is weakly observable,*
- $\Omega(T)$ *if G is not observable,*

where $\alpha_\varepsilon := \alpha([\mathcal{G}]_\varepsilon)$ and $\delta_\varepsilon := \delta([\mathcal{G}]_\varepsilon)$.

The lower bound in the non-observable case is the same as Alon et al. (2015, Theorem 6) with a deterministic feedback graph. The remaining lower bounds are nontrivial adaptations of the corresponding bounds for the deterministic case by Alon et al. (2015, 2017). The main technical hurdle is due to the stochastic nature of the feedback graph, which needs to be taken into account in the proofs. The rationale behind the constructions used for proving the lower bounds is as follows: since each edge is realized only with probability ε, any algorithm requires $1/\varepsilon$ rounds in expectation in order to observe the loss of an action in the out-neighborhood of the played action, whereas one round would suffice with a deterministic feedback graph. Note that Theorem 5.5 implies that, if $\operatorname{supp}\big([\mathcal{G}]_{\varepsilon(K,T)}\big)$ is not observable for $\varepsilon(K, T)$ as in Theorem 5.4, then there is no hope to achieve sublinear regret, as the lower bounds for both strongly and weakly observable supports are linear in T for all $\varepsilon \leq \varepsilon(K, T)$.

5.5 Refined Graph-Theoretic Parameters

Although the results from Section 5.3 are worst-case optimal up to logarithmic factors, we may find that the factors $\sqrt{\alpha(G_\varepsilon)/\varepsilon}$ and $(\delta(G_\varepsilon)/\varepsilon)^{1/3}$ for strongly and weakly observable $G_\varepsilon := \operatorname{supp}([\mathcal{G}]_\varepsilon)$, respectively, may be improved upon in certain cases. In particular, we show that, under additional assumptions on the feedback that we receive, we can obtain better regret bounds. To understand our results, we need some initial definitions. First, we introduce the definition for a weighted version of the independence number.

Definition 5.2 (Weighted independence number). *First, the* weighted independence number *for a graph $H = (V, E)$ and positive vertex weights $w\colon V \to \mathbb{R}_{>0}$ is defined as*

$$\alpha_{\mathsf{w}}(H, w) := \max_{S \in \mathcal{I}(H)} \sum_{i \in S} w(i) ,$$

where $\mathcal{I}(H)$ denotes the family of independent sets in H.

We consider two different weight assignments computed in terms of any stochastic feedback graph $\mathcal{G}$ with edge probabilities $[p(i,j)]_{i,j\in V}$ and $\operatorname{supp}(\mathcal{G}) = G$. For any $i \in V$, they are defined as the $w^-_{\mathcal{G}}(i) \coloneqq \big(\min_{j\in N^{\mathrm{in}}_G(i)} p(j,i)\big)^{-1}$, the inverse of the least probability of observing i from any other vertex, and $w^+_{\mathcal{G}}(i) \coloneqq \big(\min_{j\in N^{\mathrm{out}}_G(i)} p(i,j)\big)^{-1}$, the inverse of the least probability of observing any other vertex from i. Then, the two corresponding weighted independence numbers are $\alpha^-_{\mathsf{w}}(\mathcal{G}) \coloneqq \alpha_{\mathsf{w}}(G, w^-_{\mathcal{G}})$ and $\alpha^+_{\mathsf{w}}(\mathcal{G}) \coloneqq \alpha_{\mathsf{w}}(G, w^+_{\mathcal{G}})$. The parameter of interest for the results in this section is $\alpha_{\mathsf{w}}(\mathcal{G}) \coloneqq \alpha^-_{\mathsf{w}}(\mathcal{G}) + \alpha^+_{\mathsf{w}}(\mathcal{G})$.[§] For more details on the weighted independence number, see Appendix C.5.

Furthermore, we analogously introduce a weighted version of the weak domination number.

Definition 5.3 (Weighted weak domination number)**.** *The* weighted weak domination number δ_{w} *for a graph $H = (V,E)$ and positive vertex weights $w\colon V \to \mathbb{R}_{>0}$ is defined as*

$$\delta_{\mathsf{w}}(H,w) \coloneqq \min_{D\in\mathcal{D}(H)} \sum_{i\in D} w(i) \ ,$$

where $\mathcal{D}(H)$ denotes the family of weakly dominating sets in H.

In this section, we focus on the weighted weak domination number $\delta_{\mathsf{w}}(\mathcal{G}) \coloneqq \delta_{\mathsf{w}}(G, w^+_{\mathcal{G}})$. We also define what we call the *self-observability* parameter $\sigma(\mathcal{G})$ of the stochastic feedback graph $\mathcal{G}$ as

$$\sigma(\mathcal{G}) \coloneqq \sum_{i\in V : i\in N^{\mathrm{in}}_G(i)} \frac{1}{p(i,i)} \ .$$

To gain some intuition on the graph-theoretic parameters introduced above, consider the graph with only self-loops, also used in Example 5.1 below. If all $p(i,i) = \varepsilon$, the learner needs to pull a single arm $1/\varepsilon$ times for one observation in expectation, and K/ε times to see the losses of all arms once. However, when the edge probabilities are different we need to pull arms for $\sum_{i=1}^K 1/p(i,i)$ times. The weighted independence number, weighted weak domination number, and self-observability parameter generalize this intuition and tell us how many observations the learner needs in order to see all losses at least once in expectation. We now state the main result of this section.

Theorem 5.6 (Informal)**.** *There exists an algorithm with per-round running time of $O(K^4)$ and whose regret R_T, ignoring logarithmic factors, satisfies*

$$R_T \lesssim \min\Bigg\{ T,\ \min_{\varepsilon}\Big\{ \sqrt{\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)T} \ :\ \operatorname{supp}([\mathcal{G}]_\varepsilon) \ \textit{is strongly observable}\Big\}, \\ \min_{\varepsilon}\Big\{ \big(\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\big)^{1/3} T^{2/3} + \sqrt{\sigma([\mathcal{G}]_\varepsilon)T} \ :\ \operatorname{supp}([\mathcal{G}]_\varepsilon) \ \textit{is observable}\Big\} \Bigg\} \ .$$

The regret bound in Theorem 5.6 follows from Theorem C.5 in Appendix C.4. In that section of Appendix C.4, we essentially design a version of the Exponential Weights algorithm tailored to handle stochastic feedback graphs, analogously to the Exp3.G algorithm (Alon et al., 2015) for deterministic graphs. However, this description alone is only partial. The main missing detail regards

[§] One may equivalently define $\alpha_{\mathsf{w}}(\mathcal{G})$ as the average, or the maximum, of the two weighted independence numbers. This preserves the same shape of the regret guarantees within this section, by only loosing a multiplicative factor of $\sqrt{2}$ at most.

the absence of any a priori information of the stochastic feedback graph, not even on the structure of the support given by the best thresholding (contrarily to our first algorithm from Section 5.3 that first determines the graph structure to play over). We observe that avoiding separating the learning phases and, instead, simultaneously learning both the stochastic feedback graph $\mathcal{G}$ and the losses appears to be crucial if we desire an improved instance-dependence of our regret bound. To satisfy this requirement, as briefly anticipated before, our algorithm first optimistically assumes the support from the best thresholding to be strongly observable, and eventually switches to the weakly observable regime in a timely manner when it is determined to be preferable in terms of regret. Since the construction and the analysis of the above algorithm are both particularly involved, we defer them to Appendix C.4. We nevertheless provide here some intuition to understand the design of this algorithm and how its properties guarantee the final regret bound from Theorem 5.6.

5.5.1 Proof Sketch for the Improved Regret Analysis

We begin from the per-round running time, which is mainly determined by approximating δ_{w} for all K^2 possible thresholds. In each of the thresholded graphs, we can compute a $(\ln(K)+1)$-approximation for the weighted weak domination number in $\mathcal{O}(K^2)$ time by reduction to set cover (Chvatal, 1979, Vazirani, 2001). Doing so only introduces an extra factor of order $\ln^{1/3}(K)$ in the regret bound.

Regarding the regret analysis, an important property of the bound in Theorem 5.6 is that it is never worse than the bounds obtained before. The following example shows that the regret bound in Theorem 5.6 can also be better than previously obtained regret bounds.

Example 5.1 (Faulty bandits). *Consider a stochastic feedback graph $\mathcal{G}$ for the K-armed bandit setting: $p(i,i)=\varepsilon_i \in (0,1]$ for all $i \in V$ and $p(i,j)=0$ for all $i \neq j$. In this case, the regret of* EdgeCatcher *is $\widetilde{O}\big(\sqrt{KT/(\min_i \varepsilon_i)}\big)$. On the other hand, Theorem 5.6 provides the bound $\widetilde{O}\big(\sqrt{T\sum_i(1/\varepsilon_i)}\big)$, as $\alpha_{\mathsf{w}}(\mathcal{G})=2\sum_i 1/\varepsilon_i$. In the special case when $\varepsilon_i=\varepsilon\in(0,1]$ for some $i \in V$ while $\varepsilon_j = 1$ for all $j \neq i$, the regret of* EdgeCatcher *is $\widetilde{O}(\sqrt{KT/\varepsilon})$, while Theorem 5.6 guarantees a $\widetilde{O}(\sqrt{(K+1/\varepsilon)T})$ regret bound.*

To derive these tighter bounds, we exploit the additional assumption that the realized feedback graph G_t is observed at the end of each round. This allows us to *simultaneously* estimate the feedback graph and control the regret, rather than performing these two tasks sequentially as in Section 5.3. In particular, we use this extra information to construct a novel importance-weighted estimator for the loss, which for rounds $t \geq 2$ is defined to be

$$\widetilde{\ell}_t(i) \coloneqq \frac{\ell_t(i)}{\widehat{P}_t(i)}\mathbb{I}\big\{i \in N^{\text{out}}_{G_t}(I_t) \wedge i \in N^{\text{out}}_{\widehat{G}_t}(I_t)\big\} \qquad \forall i \in V\ , \tag{5.6}$$

where $\widehat{P}_t(i) \coloneqq \sum_{j \in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\widehat{p}_t(j,i)$ is the estimated probability of observing the loss of arm i at round t, $\pi_t(i) \in \Delta_V$ is the distribution we sample I_t from, and $\widehat{G}_t$ is the support of the estimated graph $\widehat{\mathcal{G}}_t$. Note that we ignore losses that we receive due to missing edges in $\widehat{G}_t$, even when they are realized. We demonstrate that we do pay an additive term in the regret for wrongly estimating an edge, which is why it is important to control which edges are in $\widehat{G}_t$. Ideally, we would use $P_t(i) \coloneqq \sum_{j \in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)p(j,i)$ rather than $\widehat{P}_t(i)$, as this is the true probability of observing the loss of arm i in round t. However, since we do not have access to $p(j,i)$, we instead use an upper-confidence estimate of $p(j,i)$ for rounds $t \geq 2$ given by

$$\widehat{p}_t(j,i) \coloneqq \widetilde{p}_t(j,i) + \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)\;,$$

where $\widetilde{p}_t(j,i) \coloneqq \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}\{(j,i) \in E_s\}$. We denote by $\widehat{\mathcal{G}}_t^{\mathrm{UCB}}$ the stochastic graph with edge probabilities $\widehat{p}_t(j,i)$. Note that the support of $\widehat{\mathcal{G}}_t^{\mathrm{UCB}}$ is a complete graph because $\widehat{p}_t(j,i) > 0$ for all $(j,i) \in V \times V$. These estimators for the edge probabilities are sufficiently good for our purposes whenever some "good" event $\mathcal{K}$ occurs, which we define as

$$\mathcal{K} \coloneqq \bigcap_{t\geq 2}\bigcap_{i,j\in V}\left\{|\widetilde{p}_t(j,i) - p(j,i)| \leq \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)\right\}\;.$$

An important property of $\widetilde{\ell}_t$ can be found in Lemma 5.1 below. It tells us that we may treat $\widetilde{\ell}_t$ as if event $\mathcal{K}$ is always realized, i.e., the estimator $\widehat{p}_t(j,i)$ is always an upper bound on $p(j,i)$. The proof of Lemma 5.1 is implied by Lemma C.5 in Appendix C.4.

Lemma 5.1 (Informal)**.** *Let e_k denote the basis vector with $e_k(i) \coloneqq \mathbb{I}\{i = k\}$ as i-th entry for each $i \in [K]$. The loss estimate $\widetilde{\ell}_t$ defined in* (5.6) *satisfies*

$$R_T = \widetilde{\mathcal{O}}\left(\mathbb{E}\left[\sum_{t=2}^{T}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)}{(t-1)\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right] + \max_{k\in V}\mathbb{E}\left[\sum_{t=2}^{T}\sum_{i=1}^{K}\big(\pi_t(i) - e_k(i)\big)\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]\right). \tag{5.7}$$

Lemma 5.1 shows that we only suffer $\widetilde{\mathcal{O}}\Big(\sqrt{\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\pi_t(i)}{P_t(i)}}\Big)$ additional regret compared to when we know $p(j,i)$. Lemma 5.1 also shows that $\widetilde{\ell}_t$ behaves nicely in the sense that, conditioned on $\mathcal{K}$, we have $\widetilde{\ell}_t(i) \leq \frac{\ell_t(i)}{P_t(i)}\mathbb{I}\big\{i \in N_{G_t}^{\mathrm{out}}(I_t) \wedge i \in N_{\widehat{G}_t}^{\mathrm{out}}(I_t)\big\}$. This is a crucial property to bound the regret of our algorithm. We show that, with a modified version of Exp3.G, the second sum on the right-hand side of Equation (5.7) is bounded from above by a term of order $\sqrt{\sum_{t=2}^{T}\sum_{i=1}^{K}\frac{\pi_t(i)}{P_t(i)}}$ too, meaning that the overall regret is also bounded similarly. The final step of the analysis is to prove that the above term is bounded in terms of the minimum of the weighted independence number and the weighted weak domination number plus self-observability. As we finally manage to do so, this concludes the proof for Theorem 5.6.

Chapter 6

Delayed Bandits: When Do Intermediate Observations Help?

We study a K-armed bandit problem with delayed bandit feedback and intermediate observations. In this model, an intermediate observation is any element from a finite state space $\mathcal{S}$ and is observed immediately after taking an action, whereas the loss is observed after an adversarially chosen delay. We show that the regime of the mapping of states to losses determines the complexity of the problem, irrespective of whether the mapping of actions to states is stochastic or adversarial. If the state-loss mapping is adversarial, then we prove that intermediate observations cannot help. Otherwise, if the same mapping is stochastic, we design an algorithm whose regret grows at rate $\sqrt{\big(K + \min\{|\mathcal{S}|, d\}\big)T}$ without logarithmic factors, implying that intermediate observations can reduce the negative effect of the total delay if their number $|\mathcal{S}|$ sufficiently small. We also provide refined high-probability regret bounds for non-uniform delays, together with experimental validation of our results.

6.1 Introduction

Delay is an ubiquitous phenomenon that many sequential decision makers have to deal with. For example, outcomes of medical treatments are often observed with delay, purchase events happen with delay after advertisement impressions, and acceptance/rejection decisions for scientific papers are observed with delay after manuscript submissions. The impact of delay on the performance of sequential decision makers, measured by regret, has been extensively studied under full information and bandit feedback, and in stochastic and adversarial environments. Yet, in many real-life situations, *intermediate observations* may be available to the learner. For example, a health check-up might give a preliminary indication on the effect of a treatment, an advertisement click might be a precursor for an upcoming purchase, and preliminary reviews might provide some information regarding an upcoming acceptance or rejection decision. In this chapter, we investigate when and how intermediate observations can be used to reduce the impact of delays in observing the final outcome of an action in a multi-armed bandit setting.

Online learning with delayed feedback and intermediate observations was studied by Mann, Gowal, György, Hu, Jiang, Lakshminarayanan, and Srinivasan (2019) in a full-information setting, and subsequently by Vernade, György, and Mann (2020) in a non-stationary stochastic bandit setting. In the paper of Vernade et al. (2020), at each round the learner chooses an action and immediately observes a signal (also called state) belonging to a finite set. The actual loss (i.e., feedback) incurred by the learner in that time step is only received with delay, which can be fixed or random. More formally, the observed state is drawn from a distribution that only depends on the chosen action, and the incurred loss is drawn from a distribution that only depends on the observed state (and not on the chosen action), forming a Markov chain.

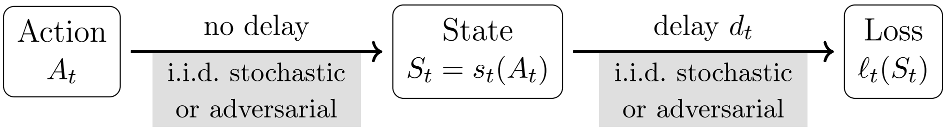


Figure 6.1: Scheme depicting the delayed feedback setting with intermediate observations.

The work of Vernade et al. (2020) studies a setting where mappings s_t from actions to states are non-stationary and losses ℓ_t over states are i.i.d. stochastic. In this chapter, we instead consider two possible regimes for the action-state mappings s_t (stochastic and adversarial) and two possible regimes for the mappings ℓ_t from states to losses (also stochastic and adversarial). Altogether, we study four different regimes, defined by the combination of the first and the second mapping type (see Figure 6.1).

We characterize (within logarithmic factors) the minimax regret rates for all of them, by giving upper and lower bounds. Similar to Vernade et al. (2020), we assume that the states are observed instantaneously, and that the losses are observed with some delay $d \in \mathbb{N}$. We show that the minimax regret rate is fully determined by the regime of the state-loss mapping, regardless of the regime of the action-state mapping. The results are informally summarized in Table 6.1, where K denotes the number of actions, S denotes the number of states, and T denotes the time horizon. It is assumed that the losses belong to the $[0, 1]$ interval. All of our upper bounds hold with high probability (with respect to the learner's internal randomization) irrespective of the regime of the action-state mapping.

State-loss mapping	Regret bounds	References
Adversarial	$\sqrt{dT} + \sqrt{KT}$	Cesa-Bianchi et al. (2019) Theorem 6.7
Stochastic	$\min\left\{\sqrt{ST} + d\sqrt{S}, \sqrt{dT}\right\} + \sqrt{KT}$	Theorems 6.2 and 6.3 Corollary 6.2

Table 6.1: Summary of our results with fixed delay d, ignoring logarithmic factors.

We recall that, up to logarithmic factors, the minimax regret rate in multi-armed bandits with delays without intermediate observations is of order $\sqrt{(K + d)T}$ (Cesa-Bianchi et al., 2019, Zimmert and Seldin, 2020). Therefore, given our findings we conclude that, if the mapping from states to losses is adversarial, then intermediate observations do not help (in the minimax sense) because the regret rates are the same irrespective of whether the intermediate observations are used or not, and irrespective of whether the mapping from actions to states is stochastic or adversarial. However, if the mapping from states to losses is stochastic, and the number S of states is smaller than the delay d, then intermediate observations are helpful, and we provide an algorithm, `AdaMetaBIO`, which is able to exploit them. Our result improves on the $\widetilde{\mathcal{O}}\big(\sqrt{KST}\big)$ regret bound obtained by Vernade et al. (2020) for the case of stochastic and stationary action-state mapping. Our algorithm also applies to a more general setting of non-uniform delays $(d_t)_{t\in[T]}$ where we achieve a high-probability regret bound of order $\sqrt{KT+\min\{ST,\mathcal{D}_T\}}$, ignoring logarithmic factors once more and terms not depending on T. This improves upon the total delay term $\mathcal{D}_T = d_1 + \cdots + d_T$ similarly to the respective term in the fixed delay setting.

Roadmap. We provide a formal definition of the problem in Section 6.2. In Section 6.3, we introduce two algorithms, `MetaBIO` and `AdaMetaBIO`, for the model of bandits with intermediate observations. Section 6.4 contains the analysis of both algorithms, where we prove high-probability regret bounds for the setting of adversarial action-state mappings and stochastic losses. We provide regret lower bounds in Section 6.5, and experimental validation of our results in Section 6.6, concluding with a short discussion in Section 6.7.

6.1.1 Related Work

Adaptive clinical trials have served an inspiration for the multi-armed bandit model (Thompson, 1933) and, interestingly, they have also pushed the field to study the effect of delayed feedback (Simon, 1977, Eick, 1988). In the bandit setting, Joulani et al. (2013) have studied a stochastic setting with random delays, whereas Neu, György, Szepesvári, and Antos (2010, 2014) have studied an adversarial setting with constant delays. Cesa-Bianchi et al. (2019) have shown an $\Omega(\max\{\sqrt{KT},\sqrt{dT\ln K}\})$ lower bound for adversarial bandits with uniformly delayed feedback, and an upper bound matching the lower bound within logarithmic factors by using an Exp3-style algorithm (Auer et al., 2002b), whereas Zimmert and Seldin (2020) have reduced the gap to the lower bound down to constants by using a Tsallis-INF approach (Zimmert and Seldin, 2021). Follow up works have studied adversarial multi-armed bandits with non-uniform delays (Thune et al., 2019, Bistritz, Zhou, Chen, Bambos, and Blanchet, 2019, 2022, György and Joulani, 2021, Van der Hoeven and Cesa-Bianchi, 2022) with Zimmert and Seldin (2020) providing a near-optimal algorithm, and Masoudian et al. (2022) and Masoudian, Zimmert, and Seldin (2024) deriving best-of-both-worlds extensions and a matching lower bound for special sequences of delays. Two key techniques for handling non-uniform delays are the skipping technique, introduced by Thune et al. (2019), and algorithm parametrization by the number of outstanding observations (an observed quantity at action time related to delays), as opposed to the delays (an unobserved quantity at action time), introduced by Zimmert and Seldin (2020). Finally, the presence of delays has been further considered in more complex extensions of multi-armed bandits (Van der Hoeven, Zierahn, Lancewicki, Rosenberg, and Cesa-Bianchi, 2023).

6.2 Problem Setting

We consider an online learning setting with a finite set $\mathcal{A} \coloneqq [K]$ of $K \geq 2$ actions and a finite set $\mathcal{S} \coloneqq [S]$ of $S \geq 2$ states. In each round $t \in [T]$, the learner picks an action $A_t \in \mathcal{A}$ and receives a state $S_t \coloneqq s_t(A_t) \in \mathcal{S}$ as an intermediate observation according to some unknown action-state mapping $s_t \in \mathcal{S}^{\mathcal{A}}$. The learner then incurs a loss $\ell_t(S_t) \in [0,1]$, which *exclusively* depends on the state associated to the selected action and is only observed at the end of round $t+d_t$, where the delay $d_t \geq 0$ is (fully) revealed to the learner only when the loss observation is received. The difficulty of this learning task depends on three elements, all initially unknown to the learner:

- the sequence of action-state mappings $s_1, \ldots, s_T \in \mathcal{S}^{\mathcal{A}}$;

- the sequence of loss vectors $\ell_1, \dots, \ell_T \in [0,1]^{\mathcal{S}}$;
- the sequence of delays $d_1, \dots, d_T \in \mathbb{N}$, where $d_t \leq T - t$ for all $t \in [T]$ without loss of generality.

Note that unlike standard bandits, as remarked above, here the losses are functions of the states instead of the actions. However, since actions are chosen without a-priori information on the action-state mappings, learners have no direct control on the losses they will incur and, because of the delays, they also have no immediate feedback on the loss associated with the observed states. Note also that, for all $t \geq 1$, the states $s_t(a)$ for $a \neq A_t$ and the losses $\ell_t(s)$ for $s \neq S_t$ are never revealed to the algorithm. For brevity, we refer to this setting as (delayed) Bandits with Intermediate Observations (BIO).

In the setting of stochastic losses, we assume the loss vectors $\ell_t \in [0,1]^{\mathcal{S}}$ are sampled i.i.d. from some fixed but unknown distribution Q, and let $\theta \in [0,1]^{\mathcal{S}}$ be the unknown vector of expected losses for the states. That is, $\ell_t(s) \sim Q(\cdot \mid s)$ has mean $\theta(s)$ for each $t \in [T]$ and $s \in \mathcal{S}$. Note that we allow dependencies between the stochastic losses of distinct states in the same round, but require losses to be independent across rounds. In the setting of stochastic action-state mappings, we assume that each observed state S_t is independently drawn from a fixed but unknown distribution $P(\cdot \mid A_t)$. If both losses and action-state mappings are stochastic, then $\ell_t(S_t)$ is independent of A_t given S_t. When losses or action-state mappings are adversarial, we assume an oblivious adversary as in previous chapters.

Our main quantity of interest is the regret measured via the learner's cumulative loss $\sum_{t=1}^T \ell_t(S_t)$, where $S_t = s_t(A_t)$ and $(A_t)_{t\in[T]}$ is the sequence of actions chosen by the learner. In the case of stochastic losses, we define the performance of the learner by $\sum_{t=1}^T \theta(S_t)$. In the case of stochastic action-state mappings, we average each instantaneous loss over the random choice of the state: $\sum_s \ell_t(s)P(s \mid A_t)$ for adversarial losses and $\sum_s \theta(s)P(s \mid A_t)$ for stochastic losses. Regret is always computed according to the best fixed action in hindsight with respect to some appropriate notion of cumulative loss. In particular, for stochastic state-action mappings, the cumulative losses of the best action are

$$\min_{a\in\mathcal{A}} \sum_{t=1}^T \sum_{s\in\mathcal{S}} \ell_t(s)P(s \mid a) \qquad \text{and} \qquad \min_{a\in\mathcal{A}} \sum_{t=1}^T \sum_{s\in\mathcal{S}} \theta(s)P(s \mid a) \ ,$$

respectively, whereas for adversarial state-action mappings they are, intuitively,

$$\min_{a\in\mathcal{A}} \sum_{t=1}^T \ell_t(s_t(a)) \qquad \text{and} \qquad \min_{a\in\mathcal{A}} \sum_{t=1}^T \theta(s_t(a)) \ .$$

6.3 A Reduction to Standard Delayed Feedback

In this section, we introduce `MetaBIO` (Algorithm 6.1), a meta-algorithm that transforms any algorithm $\mathcal{B}$ tailored for the delayed setting *without* intermediate observations into an algorithm for our setting. We then propose `AdaMetaBIO`, a modification of `MetaBIO` that delivers an improved regret bound for our setting. The idea of `MetaBIO` is to reduce the impact of delays using the information we get from intermediate observations. More precisely, if we have *enough* observations for the current state S_t at time t, we immediately feed to $\mathcal{B}$ an *estimate* of the mean loss of this state as if it were the actual loss at time t; otherwise, we wait for d_t time steps and refine our estimate using the additional loss observations.

The are two key steps in the design of our algorithm: *how* we construct the mean estimate and *when* we use it instead of waiting for the actual loss. They are the steps highlighted in green in Algorithm 6.1 (Lines 10 and 16). For all $t \in [T]$ and all $s \in \mathcal{S}$, we use $\widetilde{\theta}_t(s)$ to denote the estimate of $\theta(s)$ at round t and $n_t(s)$ to denote the number of observations for state s that we want to observe before using $\widetilde{\theta}_t(s)$. We add a subscript t to $\mathcal{L}(s)$ in Algorithm 6.1 to denote the set of loss observations $\mathcal{L}_t(s) \coloneqq \{(j, \ell_j(s)) : j + d_j \le t, S_j = s\}$ for state s that we have collected by the end of round t. Thus, $\widetilde{\theta}_t(s)$ is computed by using $N_t(s) \coloneqq |\mathcal{L}_t(s)|$ loss observations.

Algorithm 6.1: `MetaBIO`

```
 1: input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0,1)$
 2: initialize $\mathcal{L}(s) \leftarrow \emptyset$ for all $s \in \mathcal{S}$
 3: for $t = 1, \dots, T$ do
 4:     get $A_t$ from $\mathcal{B}$ and play it
 5:     observe $S_t = s_t(A_t)$
 6:     for $j : j + d_j = t$ do
 7:         receive $(j, \ell_j(S_j))$
 8:         update $\mathcal{L}(S_j) \leftarrow \mathcal{L}(S_j) \cup \{(j, \ell_j(S_j))\}$
 9:     initialize feedback set $\mathcal{M}_t \leftarrow \emptyset$
10:     compute $n_t(S_t)$
11:     if $|\mathcal{L}(S_t)| \ge n_t(S_t)$ then
12:         add $t$ to $\mathcal{M}_t$
13:     for $j : j + d_j = t \wedge |\mathcal{L}(S_j)| < n_j(S_j)$ do
14:         add $j$ to $\mathcal{M}_t$
15:     for $j \in \mathcal{M}_t$ do
16:         compute $\widetilde{\theta}_j(S_j)$ from $\mathcal{L}(S_j)$                ▷ using $\delta$
17:         feed $\big(j, A_j, \widetilde{\theta}_j(S_j)\big)$ to $\mathcal{B}$
```

Fixed delay setting. When all rounds have delay d, we simply choose $n_t(s) \coloneqq d$ for all $s \in \mathcal{S}, t \in [T]$. In other words, if we have at least d observations for some state, then we can compensate for the effect of delays and construct a well-concentrated mean estimate around the actual mean. Let $\widehat{\theta}_t(s) \coloneqq \sum_{j \in \mathcal{L}_t(s)} \ell_j(s)/N_t(s)$. Then our mean loss estimate is a lower confidence bound for $\theta(s)$ defined by

$$\widetilde{\theta}_t(s) \coloneqq \max\left\{0, \widehat{\theta}_t(s) - \frac{1}{2}\varepsilon_t(s)\right\} \tag{6.1}$$

for $\varepsilon_t(s) \coloneqq \sqrt{\frac{2}{N_t(s)} \ln \frac{4ST}{\delta}}$.

Arbitrary delay setting. In the arbitrary delay setting, where we do not have preliminary knowledge of delays, we cannot really use the delays to set $n_t(s)$. Instead, at the *end* of time t, we have access to the number of outstanding observations $\sigma_t \coloneqq \big|\{j \in [t] \,:\, j + d_j > t\}\big|$, which is the number of yet-to-arrive loss observations at the end of round t.* Then, for any $s \in \mathcal{S}$, we may set $n_t(s) \coloneqq \sigma_t$. With this choice, incurring zero delay at some round implies that we received at least half of all the loss observations we could have received in the no-delay setting (see Appendix D.2.4). In Section 6.4 we see that this ensures our mean estimate is well concentrated around its mean.

*This differs from prior work that considers outstanding observations at the *beginning* of the round.

Since Algorithm 6.1 waits for the actual loss at time t only if $N_t(S_t) < \sigma_t$, then $\widetilde{d}_t \coloneqq d_t \mathbb{I}\{N_t(S_t) < \sigma_t\}$ is the actual delay incurred by the algorithm, and $\mathcal{L}_{t+\widetilde{d}_t}(s)$ is the set of loss observations used to compute the estimate of the mean loss at time t. Because some losses may arrive at the same time, the high-probability analysis of `MetaBIO` requires these observations to be ordered. More precisely, we construct our mean estimate at time $t + \widetilde{d}_t$ for the feedback of round t using the set

$$\mathcal{L}'_t(s) \coloneqq \left\{(j, \ell_j(s)) \in \mathcal{L}_{t+\widetilde{d}_t}(s) \,\middle|\, j + \widetilde{d}_j < t + \widetilde{d}_t \vee j < t\right\} . \tag{6.2}$$

Letting $N'_t(s) \coloneqq |\mathcal{L}'_t(s)|$, we define the empirical mean

$$\widehat{\theta}_t(s) \coloneqq \sum_{j \in \mathcal{L}'_t(s)} \frac{\ell_j(s)}{N'_t(s)} . \tag{6.3}$$

Then, we set $\varepsilon_t(s) \coloneqq \sqrt{\frac{2}{N'_t(s)} \ln \frac{4ST}{\delta}}$ and define the mean loss estimator $\widetilde{\theta}_t(s)$ as a lower confidence bound similarly to Equation (6.1). We remark that, while $\widetilde{\theta}_t(s)$ is employed for the estimation of the mean loss $\theta_t(s)$ of the state s, the estimator is only ever adopted starting from time $t + \widetilde{d}_t$ with some (possibly nonzero) delay $\widetilde{d}_t$. We may thus use all the collected losses in $\mathcal{L}'_t(s) \subseteq \mathcal{L}_{t+\widetilde{d}_t}(s)$ for its definition. Therefore, once receiving the losses at the end of round t, Algorithm 6.1 constructs the estimator $\widetilde{\theta}_j(S_j)$ for the incurred loss at any (previous) round $j \in \mathcal{M}_t$ from the feedback set $\mathcal{M}_t$ using as much information as possible gathered thus far, i.e., losses in $\mathcal{L}'_j(S_j) \subseteq \mathcal{L}_t(S_j)$.

The `AdaMetaBIO` algorithm. As we already anticipated, the goal of intermediate observations is to reduce the impact of delays. However, if the number of states is too large compared to the average delay, then the information we get from intermediate observations could be misleading. We introduce `AdaMetaBIO` (Algorithm 6.2) to address this issue. Given a horizon T,[†] this algorithm runs $\mathcal{B}$ (which is tailored for the setting *without* intermediate observations) until the total incurred delay exceeds ST, and then switches to `MetaBIO`. We precise that `AdaMetaBIO` computes $\mathfrak{D}_t \coloneqq \sum_{j \le t} \sigma_j$ as the sum of outstanding observation counts up to round t, which is then used in the switching condition.

Algorithm 6.2: `AdaMetaBIO`

```
input: Algorithm $\mathcal{B}$ for standard delayed bandits, confidence parameter $\delta \in (0,1)$
initialize $\mathfrak{D}_0 \leftarrow 0$
for $t = 1, \dots, T$ do
    get $A_t$ from $\mathcal{B}$
    for $j : j + d_j = t$ do
        receive $(j, \ell_j(S_j))$
        feed $(j, A_j, \ell_j(S_j))$ to $\mathcal{B}$
    set $\sigma_t \leftarrow \sum_{j=1}^{t-1} \mathbb{I}\{j + d_j > t\}$
    update $\mathfrak{D}_t \leftarrow \mathfrak{D}_{t-1} + \sigma_t$
    if $\mathfrak{D}_t (3 \ln K + \ln(6/\delta)) > 49ST \ln \frac{8ST}{\delta}$ then
        break
if $t < T$ then
    run MetaBIO($\mathcal{B}, \delta/2$) for the remaining rounds
```

[†] Note that we may remove the a-priori knowledge of T by using a doubling trick at the cost of a polylog factor in the regret. See Remark 6.1 for further details.

6.4 Regret Analysis

We analyze `MetaBIO` and `AdaMetaBIO` in the setting of adversarial action-state mappings and stochastic losses where the regret is defined by

$$R_T \coloneqq \sum_{t=1}^{T} \theta(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a)) \ .$$

Our analysis guarantees a bound on R_T that holds with high probability (and not just in expectation), hence the reason why R_T is not defined by taking the expectation over the internal randomization of the learner or the stochasticity of the environment (as done in all previous chapters). A related notion of regret is

$$\mathcal{R}_T \coloneqq \sum_{t=1}^{T} \ell_t(S_t) - \min_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(s_t(a)) \ ,$$

which considers the realized losses instead of their means. The two quantities are close with high probability: each inequality in

$$-\sqrt{2T \ln(2K/\delta)} \leq R_T - \mathcal{R}_T \leq \sqrt{2T \ln(2/\delta)} \tag{6.4}$$

individually holds with probability at least $1 - \delta$ for any given $\delta \in (0, 1)$; see Lemma D.1 in Appendix D.

Let $\mathcal{D}_T \coloneqq \sum_{t=1}^{T} d_t$ be the total delay. We start by showing an upper bound on the total actual (or effective) delay $\widetilde{\mathcal{D}}_T = \sum_{t=1}^{T} d_t \mathbb{I}\{N_t(S_t) < \sigma_t\} \leq \mathcal{D}_T$ incurred by `MetaBIO`. Then, we provide a high-probability regret analysis of both `MetaBIO` and `AdaMetaBIO`.

More precisely, we can show that `MetaBIO` incurs the delays of no more than $\min\{2S\sigma_{\max}, T\}$ rounds, where $\sigma_{\max} \coloneqq \max_{t \in [T]} \sigma_t$. In the worst case, these rounds correspond with those from the set

$$\Phi \in \operatorname*{arg\,max}_{\mathcal{J} \subseteq [T]} \Big\{ \mathcal{D}_{\mathcal{J}} : |\mathcal{J}| = \min\{2S\sigma_{\max}, T\} \Big\} \ . \tag{6.5}$$

where we denote $\mathcal{D}_{\mathcal{J}} \coloneqq \sum_{t \in \mathcal{J}} d_t$ for any $\mathcal{J} \subseteq [T]$. Note that the set Φ is fully determined by the delay sequence $d_1, \dots, d_T$. Moreover, the total delay incurred by `MetaBIO` cannot be worse than the sum of delays corresponding to the rounds in Φ, as stated in the lemma below.

Lemma 6.1 (Total effective delay)**.** *If* `MetaBIO` *is run with any algorithm $\mathcal{B}$ on delays $(d_t)_{t \in [T]}$, then its total effective delay is $\widetilde{\mathcal{D}}_T \leq \mathcal{D}_\Phi$.*

Lemma 6.1 (proof in Appendix D.2.1) implies that, if all delays are bounded by $d_{\max}$, then $\widetilde{\mathcal{D}}_T \leq 2S\sigma_{\max} d_{\max}$, which does not depend on T. In the fixed-delay setting with delay d, for example, we get a total effective delay of at most $2Sd^2$, rather than the total delay dT we would incur without access to intermediate observations (when T is large enough).

We now turn `MetaBIO` into a concrete algorithm by instantiating $\mathcal{B}$. Specifically, we use `DAda-Exp3` (György and Joulani, 2021), a variant of `Exp3` which does not use intermediate observations and is robust to delays. `DAda-Exp3` guarantees the following regret bound.

Theorem 6.1 (György and Joulani (2021, Corollary 4.2))**.** *For any $\delta \in (0, 1)$, the regret of*

DAda-Exp3 *with respect to the realized losses in the adversarial bandits with arbitrary delays satisfies*

$$\mathcal{R}_T \leq 2\sqrt{3(2KT + \mathcal{D}_T)\ln K} + \left(\sqrt{\frac{2KT + \mathcal{D}_T}{3\ln K}} + \frac{\sigma_{\max}}{2} + 1\right)\ln\frac{2}{\delta}$$

with probability at least $1 - \delta$*.*

While Theorem 6.1 shows a high-probability bound on $\mathcal{R}_T$, Equation (6.4) shows that a high-probability bound for one notion of regret ensures a high-probability bound for the other. Although the original bound by György and Joulani (2021) was stated with $d_{\max}$ instead of $\sigma_{\max}$, we can replace the former with the latter by observing that, in the analysis of György and Joulani (2021, Theorem 4.1), they only use $d_{\max}$ to upper bound the number of outstanding observations. Note that $\sigma_{\max}$ is never larger than $d_{\max}$, indicating it is a well-behaved term that is not vulnerable to a few large delays. See Masoudian et al. (2022, Lemma 3) for a refined quantification of the relation between $\sigma_{\max}$ and $d_{\max}$.

If we consider a fixed confidence level $\delta \in (0, 1)$, then we can make the learning rate η_t and the implicit-exploration term γ_t in DAda-Exp3 depend on the specific value of δ so as to achieve an improved regret bound (see Appendix D.2.2). This allows us to show that in the BIO setting with adversarial action-state mappings and stochastic losses, the regret $\mathcal{R}_T$ of DAda-Exp3 is bounded from above by

$$2\sqrt{2KTC_{K,6\delta}} + 2\sqrt{D_T C_{K,6\delta}} + \frac{\sigma_{\max} + 2}{2}\ln\frac{2}{\delta} \tag{6.6}$$

with probability at least $1 - \delta$, where

$$C_{K,\delta} \coloneqq 3\ln K + \ln\frac{12}{\delta} \tag{6.7}$$

is a negligible logarithmic factor in K and $1/\delta$ only.

Next, we state the regret bound for MetaBIO. We remark that we initialize DAda-Exp3 with confidence parameter $\delta/2$ so as to guarantee the high-probability bound as in Equation (6.6) with probability at least $1 - \delta/2$ as required.

Theorem 6.2. *Let* $\delta \in (0, 1)$*. If we run* MetaBIO *using* DAda-Exp3*, then the regret of* MetaBIO *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$R_T \leq 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST\ln\frac{4ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max} + 2}{2}\ln\frac{4}{\delta} \tag{6.8}$$

with probability at least $1 - \delta$*.*

We begin the analysis of Theorem 6.2 by decomposing the regret into two parts: (i) the regret $\mathcal{R}_T$ of DAda-Exp3 with losses $\widetilde{\theta}_t(S_t)$, and (ii) the gap $R_T - \mathcal{R}_T$, corresponding to the cumulative error of the estimates fed to DAda-Exp3. For the first part, we follow an approach similar to György and Joulani (2021) and apply Neu (2015, Lemma 1) to obtain a concentration bound for the loss estimates defined using importance weighting along with implicit exploration. When using the actual losses, the application of Neu (2015, Lemma 1) is straightforward. However, when the mean loss estimate $\widetilde{\theta}_t(S_t)$ is used rather than the actual loss, there is a potential dependency between the chosen action A_t and $\widetilde{\theta}_t(S_t)$. In Appendix D.2.3 we carefully design a filtration to show that we may indeed use the high-probability regret bound of `DAda-Exp3` in order to upper bound the first part (regret $\mathcal{R}_T$ defined in terms of the estimates $\widetilde{\theta}_t$).

The second part requires to bound the cumulative error of our estimator in Equation (6.3) for the observed states $(S_t)_{t\in[T]}$. To this end, we use the Azuma-Hoeffding inequality to control the error of these estimates. Doing so causes a $\widetilde{\mathcal{O}}(\sqrt{ST})$ term to appear in the regret bound. The detailed proof of this part is in Appendix D.2.4, together with the proof of Theorem 6.2.

The presence of the additive $\widetilde{\mathcal{O}}(\sqrt{ST})$ term in the regret bound implies that, when $S \gg \max\{\mathcal{D}_T/T, K\}$, using intermediate feedback leads to no advantage over ignoring it. So we ideally want to recover the original bound in Equation (6.6) when this happens. `AdaMetaBIO` is an adaptive extension of `MetaBIO` that solves this issue and gives the following regret guarantee. The proof of this result is deferred to Appendix D.2.5. We remark that, to achieve this bound, before the eventual switch at some round t^* we use algorithm `DAda-Exp3` with confidence parameter set to $\delta/3$ so as to guarantee a high-probability bound on R_{t^*} with probability at least $1-\delta/2$ over the first t^* rounds (during which `DAda-Exp3` runs by itself).

Theorem 6.3. *Let $\delta \in (0,1)$. If we run* `AdaMetaBIO` *with* `DAda-Exp3`*, then the regret of* `AdaMetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$R_T \le 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta} \quad (6.9)$$

with probability at least $1-\delta$.

If we consider any upper bound $d_{\max}$ on the delays $(d_t)_{t\in[T]}$, we can further observe that the regret R_T of `AdaMetaBIO` (with `DAda-Exp3`) satisfies

$$R_T = \widetilde{\mathcal{O}}\Big(\sqrt{KT} + \min\Big\{\sqrt{S}\big(\sqrt{T}+d_{\max}\big), \sqrt{d_{\max}T}\Big\}\Big)$$

with high probability. This also follows from the fact that, as previously mentioned, we can bound the total delay of `MetaBIO` by $\mathcal{D}_\Phi \le 2Sd_{\max}^2$.

Given the previous regret bounds, we observe that we may further improve the dependency on the delays by adopting the idea of skipping rounds with large delays when computing the learning rates. This "skipping" idea was introduced by Thune et al. (2019) and has been leveraged by György and Joulani (2021) to show that `DAda-Exp3` can achieve a refined high-probability regret bound—see György and Joulani (2021, Theorem 5.1). As a consequence, we can indeed provide an improved bound in our setting by following similar steps as in the proof of Theorem 6.2. The only main change is the adoption of the version of `DAda-Exp3` that uses the skipping procedure.

Corollary 6.1. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *with* `DAda-Exp3` *with skipping (György and Joulani, 2021, Theorem 5.1), then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$R_T = \mathcal{O}\bigg(\sqrt{KTC_{K,\delta}} + \sqrt{ST\ln\frac{ST}{\delta}} + \ln\frac{1}{\delta} + \sqrt{C_{K,\delta}\ln K}\min_{R\subseteq\Phi}\Big\{|R| + \sqrt{\mathcal{D}_{\Phi\setminus R}\ln K}\Big\}\bigg)$$

with probability at least $1-\delta$.

This result could also be extended in a similar way to `AdaMetaBIO`, so as to achieve the best result from the presence of intermediate feedback.

So far, we have provided some high-probability guarantees for the regret of both `MetaBIO` and `AdaMetaBIO`, by which we can derive some expectation bounds as well (e.g., by setting $\delta \approx 1/T$). However, using the empirical mean estimators $\widehat{\theta}_t$ as the mean loss estimators at time t and working directly with the expected regret allows us to improve the achievable bound by a polylogarithmic factor. Hence, for the expected regret we use `Tsallis-INF` (Zimmert and Seldin, 2020), a learning algorithm for the standard delayed bandit problem that uses a hybrid regularizer to deal with delays and gives a minimax-optimal expected regret bound in the standard delayed setting. The proof of this expected regret upper bound is in Appendix D.2.6.

Proposition 6.1. *If we execute* `AdaMetaBIO` *with* `Tsallis-INF` *(Zimmert and Seldin, 2020), and use the switching condition* $\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST\ln(2ST)}$ *at each round* $t \in [T]$*, where* $\mathfrak{D}_t = \sum_{j=1}^{t} \sigma_j$*, then the regret of* `AdaMetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\mathbb{E}\left[R_T\right] \leq 4\sqrt{2KT} + 2\sqrt{2\mathcal{D}_\Phi \ln K} + 4\min\Big\{3\sqrt{ST\ln(2ST)}, \sqrt{2\mathcal{D}_T \ln K}\Big\}\ .$$

Remark 6.1. *In* `MetaBIO`*, we can replace* T *by* t^2 *in the definition of the confidence intervals for Equation* (6.3) *and remove the need for prior knowledge of the time horizon* T*. In* `AdaMetaBIO`*, we could use a doubling trick to avoid the prior knowledge of* T *in the switching condition. On the other hand, it is not required to know the number of states* S *for expectation bounds on the regret of* `MetaBIO`*. However, removing the prior knowledge of* S *in the high-probability regret bounds is challenging. Indeed, to the best of our knowledge, there is no result in the BIO setting that avoids prior knowledge on the number of states. Lifting this requirement in the high-probability analysis is thus an interesting question for future work.*

6.5 Lower Bounds

The lower bounds in this section are for the expected regret $\mathbb{E}\left[R_T\right]$. Since our algorithms provide high-probability guarantees, the upper bounds also apply to the expected regret. Throughout this section we will make use of constant delay, i.e., $d_t = d$ for all $t \in [T]$. We will first prove a general $\sqrt{KT}$ lower bound for all algorithms in BIO, after which we specialize to particular cases.

We start by proving a $\Omega\big(\sqrt{KT}\big)$ lower bound for any algorithm in our setting and for any combination of stochastic or adversarial action-state mappings and loss vectors. The construction is a reduction to the standard bandits lower bound construction.

Theorem 6.4. *Irrespective to whether the action-state mappings and loss vectors are stochastic or adversarial, there exists a sequence of losses such that any (possibly randomized) algorithm in BIO suffers regret* $\mathbb{E}\left[R_T\right] = \Omega\big(\sqrt{KT}\big)$*.*

Proof. Our construction only uses two states h_1 and h_2. The loss vectors, which are deterministic and do not change over time, are defined as follows: $\ell_t(h_1) \coloneqq 1$ and $\ell_t(h_2) \coloneqq 0$ for all $t \geq 0$. The stochastic action-state mapping, which is also constant over time, is given by

$$s_t(a) := \begin{cases} h_1 & \text{with probability } p_a \\ h_2 & \text{with probability } 1 - p_a \end{cases}$$

for all $a \in \mathcal{A}$ and $t \geq 0$, where the probabilities p_a are to be determined. Thus, the loss of an arm a is $\ell_t(s_t(a)) := \ell_t(h_1) = 1$ with probability p_a and $\ell_t(s_t(a)) := \ell_t(h_2) = 0$ with probability $1 - p_a$. Since the loss is determined by the state, the learner receives bandit feedback without delay. We can then choose p_a for $a \in \mathcal{A}$ to mimic the standard $\Omega\big(\sqrt{KT}\big)$ distribution-free bandit lower bound—e.g., see Slivkins (2019, Chapter 2). By Yao's minimax principle, the same lower bound also applies to the case with adversarial action-state mappings. Since the loss vectors are deterministic, this covers all possible cases in BIO. □

Adversarial action-state mapping and stochastic losses. We first prove a lower bound of order $\sqrt{ST}$ for any number $K \geq 2$ of actions. However, we do need a minor generalization of our setting to allow correlation between unseen losses. Specifically, we allow all pairs of losses $\ell_j(s), \ell_{j'}(s')$ of distinct states $s \neq s'$ to be correlated if $j > j'$ and $j - j' \leq d$, while we guarantee the i.i.d. nature of losses for any fixed state. Since $\mathbb{E}\left[\ell_t(S_t)\right] = \mathbb{E}\left[\theta(S_t)\right]$, this does not affect the analysis for the upper bound on the regret of our algorithms since $\mathbb{E}\left[R_T\right] \leq \mathbb{E}\left[\mathcal{R}_T\right]$ (see Lemma D.3). However, for a high-probability upper bound, we need to relate R_T and $\mathcal{R}_T$, which now leads to an additive $\widetilde{\mathcal{O}}(\sqrt{ST})$ term rather than an additive $\widetilde{\mathcal{O}}(\sqrt{T})$ term as in Equation (6.4).

In the proof of the $\sqrt{ST}$ lower bound, we leverage the fact that losses are independent only across time steps for a fixed state, while they may depend on the losses of the other states. Note that our lower bound holds even when the learner knows the action-state assignments beforehand. We provide a sketch of the proof of Theorem 6.5 below; see Appendix D.3 for the full proof.

Theorem 6.5. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that $d_t = d$ for all $t \in [T]$. If $T \geq \min\{S, d\}$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}\left[R_T\right] = \Omega\big(\sqrt{\min\{S, d\}T}\big)$.*

Proof sketch. First, suppose that $S \leq 2d$. For the construction of the lower bound we only consider two actions and equally split the states over these two actions. Then, we divide the T time steps in blocks of length $S/2 \leq d$. In each block, each state has the same loss. Since the block length is smaller then the delay, we have effectively created a two-armed bandit problem with $T' = T/(S/2)$ rounds and loss range $[0, S/2]$, for which we can prove a $\Omega\big(S\sqrt{T'}\big) = \Omega\big(\sqrt{ST}\big)$ lower bound by showing an equivalent lower bound for the full information setting. If $S > 2d$, we use the same construction with only $2d$ states, and obtain a $\Omega\big(\sqrt{dT}\big)$ lower bound. □

Finally, we can show the following lower bound, whose proof can be found in Appendix D.3.

Theorem 6.6. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq d + 1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}\left[R_T\right] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) .$$

This term is also present in the dynamic regret bound of NSD-UCRL2, but it is necessarily incurred from their analysis even in the stationary case (Vernade et al., 2020, Theorem 1).

This last lower bound implies that the regret of our algorithm is near-optimal. Since the lower bound of Theorem 6.4 applies to the case where the action-state mapping is adversarial and the losses are stochastic, we find the following result as a corollary of Theorem 6.4, Theorem 6.5, and Theorem 6.6.

Corollary 6.2. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \geq 1 + \min\{S, d\}$, then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}\left[R_T\right] = \Omega\Big(\max\big\{\sqrt{KT}, \sqrt{\min\{S,d\}T}, (d+1)\sqrt{S}\big\}\Big) .$$

Stochastic action-state mappings and adversarial losses. In this case, we recover the standard lower bound for adversarial bandits with bounded delay.

Theorem 6.7. *Suppose that the action-state mapping is stochastic, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists a stochastic action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret $\mathbb{E}\left[R_T\right] = \Omega\big(\max\left\{\sqrt{KT}, \sqrt{dT}\right\}\big)$.*

Proof. Since by Theorem 6.4 we already know that any algorithm must suffer $\Omega\big(\sqrt{KT}\big)$ regret, we only need to show a $\Omega(\sqrt{dT})$ lower bound. We use two states, h_1 and h_2. Our action-state mapping is deterministic and, for all $t \geq 0$, assigns $s_t(a) := h_1$ to all but one action $a^\star$, to which the mapping assigns $s_t(a^\star) := h_2$. We now have constructed a two-armed bandit problem with delayed feedback and T rounds, for which a $\Omega(\sqrt{dT})$ lower bound is known (Cesa-Bianchi et al., 2019). □

Adversarial action-state mappings, adversarial losses. Since we can recover the construction of the lower bound in Theorem 6.7, we immediately have the following result.

Corollary 6.3. *Suppose that the action-state mapping is adversarial, the losses are adversarial, and that $d_t = d$ for all $t \in [T]$. Then there exists an action-state mapping and a sequence of losses such that any (possibly randomized) algorithm suffers regret $\mathbb{E}\left[R_T\right] = \Omega\big(\max\big\{\sqrt{KT}, \sqrt{dT}\big\}\big)$.*

6.6 Experiments

We empirically compare our algorithm MetaBIO with the following baselines: DAda-Exp3 (György and Joulani, 2021) for adversarial delayed bandits without intermediate observations (which we used to instantiate the algorithm $\mathcal{B}$), the standard UCB1 algorithm (Auer et al., 2002a) for stochastic bandits without delays and intermediate observations, and NSD-UCRL2 (Vernade et al., 2020) for non-stationary stochastic action-state mappings and stochastic losses. We run all experiments with a time horizon of $T = 10^4$. All our plots show the cumulative regret of the algorithms considered as a function of time. The performance of each algorithm is averaged over 20 independent runs in every experiment, and the shaded areas consider a range centered around the mean with half-width corresponding to the empirical standard deviation of these 20 repetitions. In the first two experiments, we consider both fixed delays $d \in \{50, 100, 200\}$ and random delays $d_t \sim \text{Laplace}(50, 25)$ sampled i.i.d. from the Laplace distribution with $\mathbb{E}\left[d_t\right] = 50$.

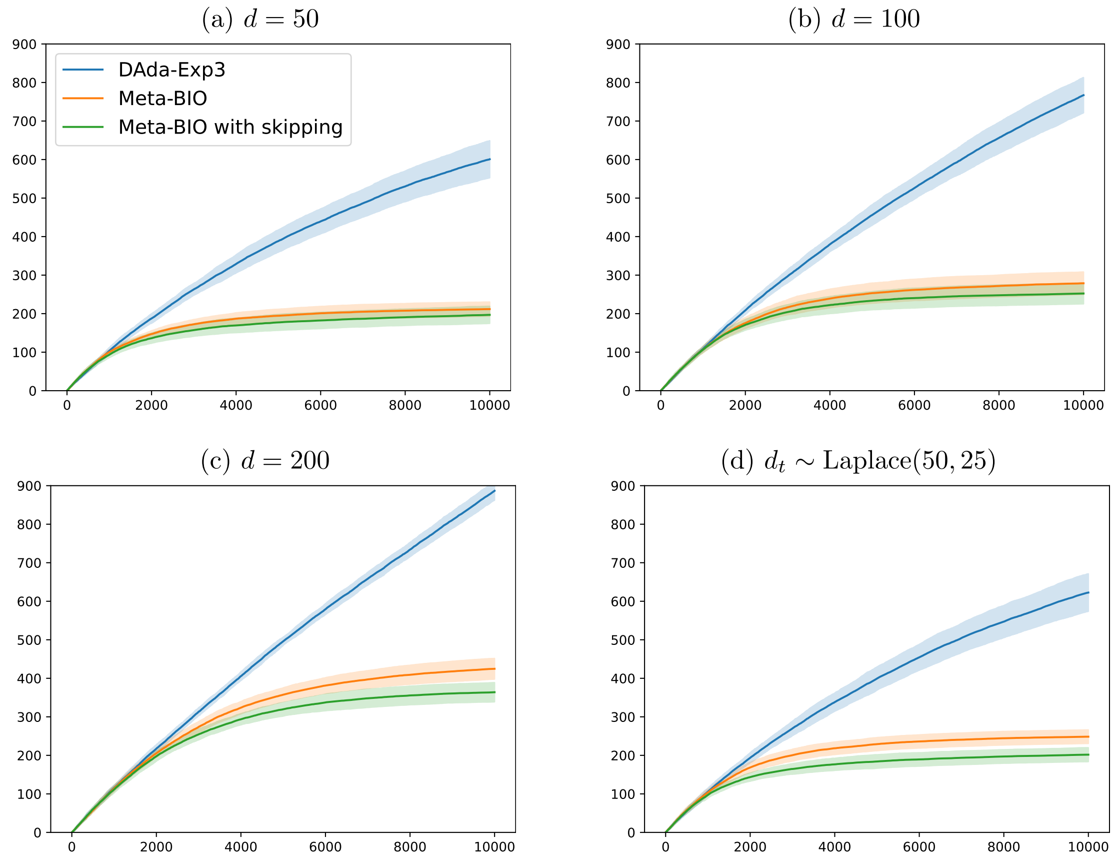


Figure 6.2: Cumulative regret over time for the stochastic action-state mapping when delays are fixed or random.

Experiment 1: stochastic action-state mappings. Here we use a stationary version of the experiments in Vernade et al. (2020)—see Table D.1 in Appendix D.4 for details. We set $K = 4$ and $S = 3$, while we repeat this experiment for the previously mentioned values of delays. Figure 6.2 shows that, across all delay regimes, MetaBIO largely improves on the performance of DAda-Exp3 by exploiting intermediate observations.

Experiment 2: adversarial action-state mappings. In this construction, we simulate the adversarial mapping using a construction adapted from Zimmert and Seldin (2021): we alternate between two stochastic mappings while keeping the loss means fixed. We set $K = 4$, $S = 3$, and we consider multiple instances for the different values of delays as in the previous experiment. The interval between two consecutive changes in the distribution of action-state mappings grows exponentially. See Table D.2 in Appendix D.4 for details. Figure 6.3 shows that MetaBIO and MetaBIO with “skipping” outperform both UCB1 and NSD-UCRL2.

Experiment 3: utility of intermediate observations. Here we set $K = 8$, $d = 100$, and investigate how the performance of MetaBIO changes when the number S of states varies in $\{4, 6, 8, 10, 12\}$. The mean loss is always 0.2 for the optimal state and 1 for the others. The optimal action always maps to the optimal state. The suboptimal actions map to the optimal state with probability 0.6 and map to a random suboptimal state with probability 0.4. This implies that the expected loss of each arm remains constant when the number of states changes. Figure 6.4 shows that the regret

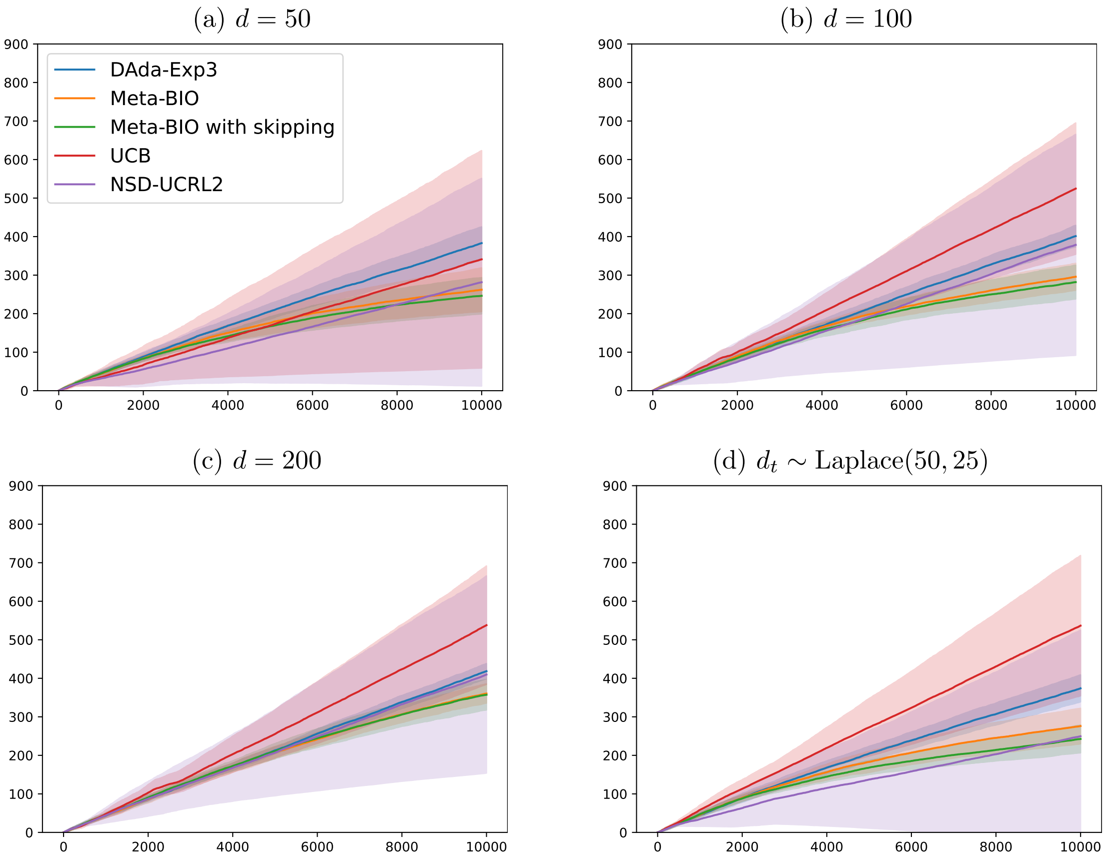


Figure 6.3: Cumulative regret over time for the adversarial action-state mapping when delays are fixed or random. All algorithms have small variance except for UCB1 and NSD-UCRL2.

gap between MetaBIO and DAda-Exp3 shrinks as the number of states increases. This observation confirms our theoretical findings about the dependency of the regret on the number of states, which leads to a larger improvement the fewer they are.

Experiment 4: performance of AdaMetaBIO when $S < d$. We use the same setting as in Experiment 1 with delay $d = 20$.[‡] Figure 6.6 shows the performance of AdaMetaBIO compared with both DAda-Exp3 and MetaBIO. Before the switching point, AdaMetaBIO runs DAda-Exp3 (up to independent internal randomization). Afterwards, AdaMetaBIO switches to MetaBIO (which in turn runs DAda-Exp3 as a subroutine) and quickly aligns with its performance. Note that, at the switching time, AdaMetaBIO uses (via MetaBIO) the same instance of DAda-Exp3 that was already running, rather than starting a new instance. It can be shown that our analysis of AdaMetaBIO applies to this variant as well without changes in the order of the bound.

Experiment 5: performance of AdaMetaBIO when $S > d$. We use a setting that is almost identical to that of Experiment 3, except we set $d = 4$ and $S = 14$. The performance of the three algorithms is shown in Figure 6.5. We can observe that AdaMetaBIO does not switch to MetaBIO and its performance is thus the same as that of DAda-Exp3, whereas MetaBIO incurs a larger regret.

[‡] Compared to the switching condition used for the analysis of AdaMetaBIO, we replace $49ST \ln \frac{8ST}{\delta}$ with ST. This change allows the switching condition to be triggered more easily to provide a better visualization of the behaviour of AdaMetaBIO, while it only introduces a polylog factor in its regret bound.

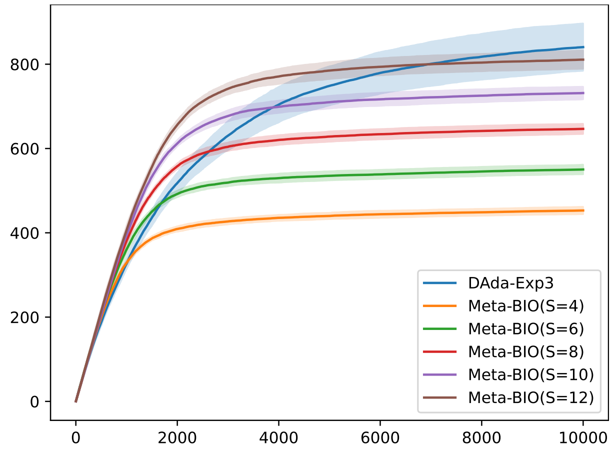


Figure 6.4: Cumulative regret over time of both `DAda-Exp3` and `MetaBIO` with different numbers of states $S \in \{4, 6, 8, 10, 12\}$.

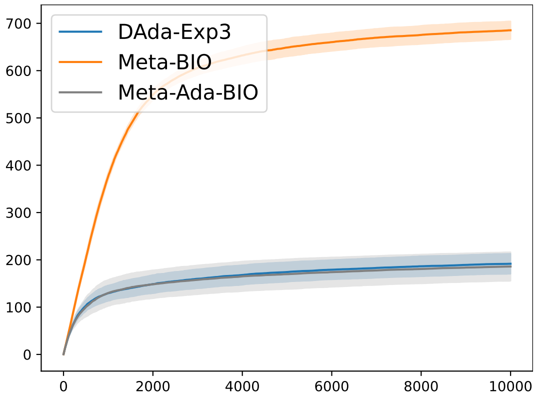


Figure 6.5: Cumulative regret over time of `DAda-Exp3`, `MetaBIO` and `AdaMetaBIO` when $S > d$.

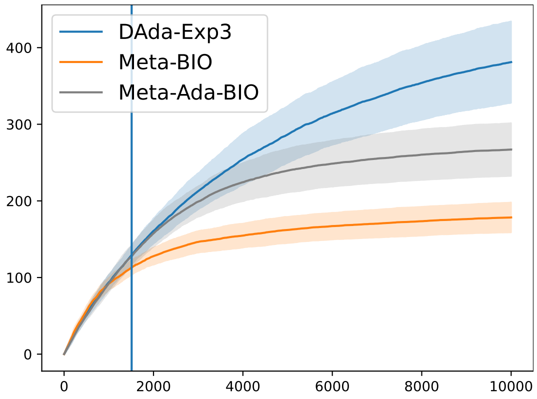


Figure 6.6: Cumulative regret over time of `DAda-Exp3`, `MetaBIO` and `AdaMetaBIO`. The vertical blue line marks the switching point of `AdaMetaBIO`.

6.7 Conclusions

The work of Vernade et al. (2020) also considers a non-stationary action-state mapping and derive regret bounds for the switching regret. Preliminary results suggest that, as long as there is an algorithm that can provide bounds on the switching regret with delayed feedback, our ideas also transfer to this setting. To the best of our knowledge, there is currently no algorithm that can provide bounds on the switching regret with delayed feedback and we leave this as a promising direction for future work.

Part II

Uniform Convergence and a Theory of Interpretability

Chapter 7

Statistical Learning Theory

This introductory chapter provides the basics of the established and well-studied statistical learning framework, accompanied by fundamental and classical results from the learning theory literature. Part of the content in this introductory chapter is inspired from the textbook by Shalev-Shwartz and Ben-David (2014), which is recommended to read for further details and remarks. Throughout this part of the manuscript, we implicitly assume measurability whenever required for ease of exposition.

7.1 The Statistical Learning Framework

We begin this chapter by introducing the statistical learning framework, which is a general mathematical model for the analysis of learning algorithms. This framework determines how any example (x, y), which is a pair composed by a data point x and a label y, received by a learning algorithm is generated and how the performance of any label predictor is evaluated. Let $\mathcal{X}$ be the domain space and let $\mathcal{Y}$ be the label space. We assume that each example $Z = (X, Y) \sim \mathcal{D}$ is drawn independently from a fixed but unknown distribution $\mathcal{D}$ over the example space $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$. These examples are then collected within a *training set* $S \coloneqq (Z_1, \dots, Z_m) \in \mathcal{Z}^m$ of size $m \in \mathbb{N}$, where $Z_1, \dots, Z_m \sim \mathcal{D}$ are i.i.d., and provided to the learning algorithm.

The aim of a learning algorithm is to provide an appropriate *predictor*, or *hypothesis*, $h \colon \mathcal{X} \to \mathcal{Y}$ given a training set. To measure the quality of the labels predicted by h, we adopt a nonnegative *loss function* $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Then, in the statistical learning framework, an instance is generally described by the pair $(\mathcal{D}, \ell)$ and the performance of a given predictor h is measured by the *statistical risk*

$$\ell_{\mathcal{D}}(h) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell(Y, h(X))\big] \ .$$

On the other hand, the *empirical risk*, or training error, $\ell_S(h)$ of h is the average loss over the examples $Z_i \coloneqq (X_i, Y_i)$ from the training set S, that is,

$$\ell_S(h) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \ell(Y_i, h(X_i)) \ .$$

Observe that, while a learner might be unable to compute the statistical risk $\ell_{\mathcal{D}}(h)$ of a given hypothesis h because $\mathcal{D}$ is unknown, it is possible to calculate its empirical risk $\ell_S(h)$ over the training set S provided access to the loss function. Also notice that, assuming h does not depend on

S, the empirical risk of h is an unbiased estimator for its statistical risk since $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\ell_S(h)\right] = \ell_{\mathcal{D}}(h)$ by linearity of expectation, where $\mathcal{D}^m$ is the product probability measure over $\mathcal{Z}^m$.

A learning algorithm can be seen as a function

$$A\colon \bigcup_{m\in\mathbb{N}} \mathcal{Z}^m \to \mathcal{Y}^{\mathcal{X}}$$

that outputs a hypothesis $A(S) \in \mathcal{Y}^{\mathcal{X}}$ upon receiving a certain training set $S \in \mathcal{Z}^m$ of size m. While in general the predictor $A(S)$ can be any function from a class $\mathcal{H}_m \subseteq \mathcal{Y}^{\mathcal{X}}$ that depends on the training set size m, we henceforth consider learning algorithms that output hypotheses from a fixed *hypothesis class* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. The restriction of the output of A to $\mathcal{H}$ introduces what is known as an *inductive bias*.

Ideally, the most desirable behavior for a learning algorithm consists of providing the best possible mapping from data points to labels in terms of statistical risk. This concept is well known and takes the name of Bayes optimal predictor.

Definition 7.1 (Bayes optimal predictor). *Given any instance $(\mathcal{D}, \ell)$, the* Bayes optimal predictor *$f^*\colon \mathcal{X} \to \mathcal{Y}$ is the best possible predictor $f^* \in \arg\min_{f\in\mathcal{Y}^{\mathcal{X}}} \ell_{\mathcal{D}}(f)$ which is defined as*

$$f^*(x) = \arg\min_{y\in\mathcal{Y}} \mathbb{E}_{\mathcal{D}}\left[\ell(Y,y) \mid X = x\right] \qquad \forall x \in \mathcal{X} \ .$$

Its risk $\ell_{\mathcal{D}}(f^*)$ is also known as Bayes risk *or* Bayes error.

Since we restrict the learning algorithm to select a hypothesis from the class $\mathcal{H}$, it may be impossible for the learner to compute the Bayes optimal predictor f^* if the latter does not belong to $\mathcal{H}$. Hence, we may only hope for such an algorithm to find the best predictor in the class

$$h^* \in \arg\min_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) \ ,$$

which might not perform as well as f^*. Moreover, the problem about computing either h^* or f^* lies in the necessity of knowing $\mathcal{D}$, which is an unsatisfiable requirement in many scenarios. We should consequently expect the risk of a hypothesis output by the learner to be larger than the Bayes risk, and the design of the learning algorithm should aim at closing this gap as much as possible.

This fact is more explicitly depicted by the well-known *bias-variance decomposition* of the risk of a given hypothesis $h \in \mathcal{H}$:

$$\ell_{\mathcal{D}}(h) = \underbrace{\ell_{\mathcal{D}}(h) - \ell_{\mathcal{D}}(h^*)}_{\text{variance error}} + \underbrace{\ell_{\mathcal{D}}(h^*) - \ell_{\mathcal{D}}(f^*)}_{\text{bias error}} + \underbrace{\ell_{\mathcal{D}}(f^*)}_{\text{Bayes error}} \ .$$

This decomposition shows that, except for the unavoidable Bayes error, the learner mainly incurs two types of errors: the *bias error*, which is the error of approximating f^* introduced by the inductive bias from the choice of $\mathcal{H}$, and the *variance error*, which consists of the error in estimating h^* within $\mathcal{H}$ by using the information contained in the training set S.

7.2 PAC Learning and Uniform Convergence

As mentioned in the previous section, the objective of the learner A is to output a hypothesis from $\mathcal{H}$ that minimizes the risk. The main issue we pointed out is that the learning algorithm cannot directly optimize the risk given the lack of knowledge about $\mathcal{D}$. Hence, it has to use only the information provided by the training set S. The most natural approach would be for the algorithm to optimize the empirical risk over S instead. Such an algorithm is commonly known as an *Empirical Risk Minimization* (ERM) algorithm and its output is

$$A(S) = h_S^{\mathrm{ERM}} \in \arg\min_{h \in \mathcal{H}} \ell_S(h) \ .$$

The hope of adopting the ERM rule is that optimizing the empirical risk leads to minimizing the true risk as well. This is known to be approximately the case, e.g., for binary classification tasks and it is also known to suffice for more general learning problems under some conditions. Moreover, the quality of h_S^{ERM} in terms of statistical risk intuitively depends on the size of S.

Let us begin with a common simplifying assumption called realizability, which is a property of $\mathcal{H}$ given $\mathcal{D}$ and ℓ.

Assumption 7.1 (Realizability)**.** *There exists $h \in \mathcal{H}$ such that $\ell_{\mathcal{D}}(h) = 0$.*

Under the realizability assumption, $\ell_{\mathcal{D}}(h^*) = 0$ must also hold. Furthermore, both h^* and h_S^{ERM} have empirical risk

$$0 \leq \ell_S\left(h_S^{\mathrm{ERM}}\right) \leq \ell_S(h^*) = 0$$

almost surely for any i.i.d. sample S, where the first inequality follows by nonnegativity of the loss function, the second one is due to the optimality of h_S^{ERM} with respect to the empirical risk over S, and the equality holds with probability 1 since $\ell_{\mathcal{D}}(h^*) = 0$. On the other hand, we cannot reach the same conclusion for $\ell_{\mathcal{D}}\left(h_S^{\mathrm{ERM}}\right)$. Indeed, while the empirical risk is an unbiased estimator of the true risk for any fixed hypothesis in $\mathcal{H}$, as mentioned in the previous section, $\ell_S\left(h_S^{\mathrm{ERM}}\right)$ is generally not because h_S^{ERM} depends on the entire training set S. Consequently, the above reasoning is insufficient for concluding that h_S^{ERM} is able to learn and requires a more sophisticated argument.

The notion of learnability for a given learning task has been a main object of study in statistical learning theory. The most notable formalization of this concept is that of Probably Approximately Correct learnability introduced by Valiant (1984).

Definition 7.2 (PAC learnability)**.** *A class $\mathcal{H}$ is* Probably Approximately Correct (PAC) learnable *if there exists a function $\widetilde{m}_{\mathcal{H}}^{\mathrm{PAC}} \colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that the following holds. For any $\varepsilon, \delta \in (0,1)$ and any distribution $\mathcal{D}$ over $\mathcal{Z}$ for which the realizability assumption holds, then algorithm A, given an i.i.d. sample $S \sim \mathcal{D}^m$ of size $m \geq \widetilde{m}_{\mathcal{H}}^{\mathrm{PAC}}(\varepsilon, \delta)$, it returns $h_S = A(S)$ such that*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left(\ell_{\mathcal{D}}(h_S) > \varepsilon\right) \leq \delta \ .$$

As the name states, a learning algorithm is said to PAC learn whenever its output h_S, given an i.i.d. training set S, incurs with probability $1 - \delta$ (“probably”) a risk bounded from above by ε (“approximately correct”).

By relaxing the realizability assumption, we can extend the definition of PAC learnability for more general settings commonly known as *agnostic*. Observe that, in the absence of realizability, the learning algorithm cannot aspire to obtain risk close to zero. Hence, we can only compare its performance to the risk $\ell_{\mathcal{D}}(h^*)$ of the best predictor in the class $\mathcal{H}$. This is the reason why the learnability notion in the agnostic setting revolves around the variance error $\ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*)$ rather than the risk $\ell_{\mathcal{D}}(h_S)$ only.

Definition 7.3 (Agnostic PAC learnability)**.** *A class $\mathcal{H}$ is* agnostic PAC learnable *if there exists a function $m_{\mathcal{H}}^{\mathrm{PAC}} \colon (0,1)^2 \to \mathbb{N}$ and a learning algorithm A such that the following holds. For any $\varepsilon, \delta \in (0,1)$ and any distribution $\mathcal{D}$ over $\mathcal{Z}$, then algorithm A, given an i.i.d. sample $S \sim \mathcal{D}^m$ of size $m \geq m_{\mathcal{H}}^{\mathrm{PAC}}(\varepsilon, \delta)$, it returns $h_S = A(S)$ such that*

$$\mathbb{P}_{S \sim \mathcal{D}^m}\big(\ell_{\mathcal{D}}(h_S) - \ell_{\mathcal{D}}(h^*) > \varepsilon\big) \leq \delta \,.$$

The property of agnostic PAC learnability is a formalization of the computational aspect of a statistical learning task, describing the feasibility of the algorithmic learnability via a certain class of predictors up to any given accuracy and probability of success. Additionally, showing that an ERM algorithm is an successful agnostic PAC learner would suffice to prove the agnostic PAC learnability of $\mathcal{H}$. The former condition can be shown to hold under some potentially stronger property of $\mathcal{H}$. In particular, we can first notice that the variance error of an ERM algorithm is

$$\begin{aligned}
\ell_{\mathcal{D}}\left(h_S^{\mathrm{ERM}}\right) - \ell_{\mathcal{D}}(h^*) &= \ell_{\mathcal{D}}\left(h_S^{\mathrm{ERM}}\right) - \ell_S\left(h_S^{\mathrm{ERM}}\right) + \ell_S\left(h_S^{\mathrm{ERM}}\right) - \ell_{\mathcal{D}}(h^*) \\
&\leq \ell_{\mathcal{D}}\left(h_S^{\mathrm{ERM}}\right) - \ell_S\left(h_S^{\mathrm{ERM}}\right) + \ell_S(h^*) - \ell_{\mathcal{D}}(h^*) && \text{by optimality of } h_S^{\mathrm{ERM}} \\
&\leq \left|\ell_{\mathcal{D}}\left(h_S^{\mathrm{ERM}}\right) - \ell_S\left(h_S^{\mathrm{ERM}}\right)\right| + \left|\ell_S(h^*) - \ell_{\mathcal{D}}(h^*)\right| \\
&\leq 2 \sup_{h \in \mathcal{H}} \left|\ell_S(h) - \ell_{\mathcal{D}}(h)\right| \,.
\end{aligned}$$

Then, deriving a upper bound on the absolute difference between empirical risk and true risk, uniformly over all hypotheses in $\mathcal{H}$, implies the desired property for ERM.

This type of uniform bound is more generally captured by the convergence in probability of the empirical mean of a function in a given class over an i.i.d. process to its expectation, uniformly over all functions in a given class. Said property takes the name of uniform convergence in probability or, simply, *uniform convergence*.

Definition 7.4 (Uniform convergence)**.** *Let $\mathcal{Z}$ be any domain. A class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ of real-valued functions has the* uniform convergence *property if there exists a function $m_{\mathcal{F}} \colon (0,1)^2 \to \mathbb{N}$ such that, for any $\varepsilon, \delta \in (0,1)$ and any distribution $\mathcal{D}$ over $\mathcal{Z}$, if $Z, Z_1, \dots, Z_m \sim \mathcal{D}$ are i.i.d. random variables for any $m \geq m_{\mathcal{F}}(\varepsilon, \delta)$, then*

$$\mathbb{P}_{Z_1, \dots, Z_m \sim \mathcal{D}}\left(\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(Z_i) - \mathbb{E}_{Z \sim \mathcal{D}}\left[f(Z)\right]\right| > \varepsilon\right) \leq \delta \,.$$

We now show that uniform convergence is a stronger notion than agnostic PAC learnability in general. Recalling that the domain is the example space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, we define the function class

$$\mathcal{L}_{\mathcal{H},\ell} \coloneqq \big\{f_h : f_h(x,y) \coloneqq \ell(y, h(x)), \forall (x,y) \in \mathcal{Z}, \forall h \in \mathcal{H}\big\}$$

where each function f_h corresponds to the evaluation of the loss ℓ between an input label $y \in \mathcal{Y}$ and the prediction of the hypothesis h over the input point $x \in \mathcal{X}$. It is then immediate to show that

$$\frac{1}{m}\sum_{i=1}^{m} f_h(Z_i) = \ell_S(h) \qquad \text{and} \qquad \mathbb{E}_{Z\sim\mathcal{D}}\left[f_h(Z)\right] = \ell_{\mathcal{D}}(h)$$

for any sample $S = (Z_1, \ldots, Z_m) \in \mathcal{Z}^m$. Then, the uniform convergence property of $\mathcal{L}_{\mathcal{H},\ell}$ guarantees a bound in probability on the supremum

$$\sup_{f_h\in\mathcal{L}_{\mathcal{H},\ell}}\left|\frac{1}{m}\sum_{i=1}^{m} f_h(Z_i) - \mathbb{E}_{Z\sim\mathcal{D}}\left[f_h(Z)\right]\right| = \sup_{h\in\mathcal{H}}\left|\ell_S(h) - \ell_{\mathcal{D}}(h)\right| .$$

It is therefore clear that uniform convergence is an extremely powerful tool, with strong implications in statistical learning theory and beyond. In Chapter 9, for instance, we leverage uniform convergence—together with other mathematical tools—to derive surprising results in the context of a learning-theoretic model for the interpretability of binary concepts.

7.3 Combinatorial Dimensions for Learning

In this part we introduce purely combinatorial quantities that relate to the hypothesis class $\mathcal{H}$, and we will briefly analyze the relationship between them. Surprisingly enough, while these combinatorial parameters only measure the "richness" of function classes, they are able to characterize properties of these classes such as uniform convergence and PAC learnability.

Let us first restrict ourselves to the simpler binary classification setting. We particularly assume that the label space is $\mathcal{Y} = \{0,1\}$ and hence that any predictor $h \in \mathcal{H}$ is a binary function $h\colon \mathcal{X} \to \{0,1\}$. We define the loss function ℓ to be the 0-1 loss function, i.e., $\ell(y,y') \coloneqq \mathbb{I}\{y \neq y'\}$ for any $y, y' \in \{0,1\}$. The first quantity we introduce is the growth function (Vapnik and Chervonenkis, 1971) and it counts the maximum number of distinct ways that m data points from $\mathcal{X}$ can be labeled by hypotheses in $\mathcal{H}$, for any $m \in \mathbb{N}$; this is also commonly called the m-th shatter coefficient.

Definition 7.5 (Growth function)**.** *The* growth function *for a binary function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ is a function $\Pi_{\mathcal{F}}\colon \mathbb{N} \to \mathbb{N}$ defined as*

$$\Pi_{\mathcal{F}}(m) \coloneqq \sup_{x_1,\ldots,x_m\in\mathcal{X}}\left|\left\{(f(x_1),\ldots,f(x_m)) : f \in \mathcal{F}\right\}\right| \qquad \forall n \in \mathbb{N} .$$

Then, by a simple combinatorial reasoning we know that the growth function for $\mathcal{H}$ satisfies $\Pi_{\mathcal{H}}(m) \le 2^m$, where 2^m is the maximum number of dichotomies in the binary classification setting. By definition, if $\Pi_{\mathcal{H}}(m) = 2^m$ then there exists $S \in \mathcal{X}^m$ that is *shattered* by $\mathcal{H}$, meaning that the points in S can be classified in all possible ways by predictors in $\mathcal{H}$.

The other notion that we introduce here is the *Vapnik-Chervonenkis dimension* (Vapnik and Chervonenkis, 1971), commonly abbreviated in VC dimension.

Definition 7.6 (VC dimension)**.** *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a class of binary functions. The* VC dimension *$\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$ is the size of the largest set of points in $\mathcal{X}$ that can be shattered by $\mathcal{F}$, that is,*

$$\mathrm{VC}(\mathcal{F}) \coloneqq \max\left\{m \in \mathbb{N} : \Pi_{\mathcal{F}}(m) = 2^m\right\} .$$

We define $\mathrm{VC}(\mathcal{F}) \coloneqq \infty$ *if* $\Pi_{\mathcal{F}}(m) = 2^m$ *for all* $m \in \mathbb{N}$.

Even though the VC dimension is an exclusively combinatorial property of binary function classes, and learnability is never mentioned or explicitly considered within its definition, it is capable of characterizing both PAC learnability and uniform convergence for binary classification problems. The complete result states a stronger characterization and it is known as the fundamental theorem of statistical learning. Here we report its qualitative version.

Theorem 7.1 (The fundamental theorem of statistical learning)**.** *Consider any domain* $\mathcal{X}$*. Let* $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ *be a binary hypothesis class and let* ℓ *be the* 0-1 *loss function. Then, the following statements are equivalent:*

1. *$\mathcal{H}$ has finite VC dimension;*
2. *$\mathcal{L}_{\mathcal{H},\ell}$ has the uniform convergence property;*
3. *$\mathcal{H}$ is agnostic PAC learnable;*
4. *any ERM algorithm is a successful agnostic PAC learner;*
5. *$\mathcal{H}$ is PAC learnable (under Assumption 7.1);*
6. *any ERM algorithm is a successful PAC learner (under Assumption 7.1);*

Additionally, the fundamental theorem even quantifies the sample complexity for which the above properties of $\mathcal{H}$ are guaranteed in terms of its VC dimension. Namely, assuming $\mathrm{VC}(\mathcal{H}) < \infty$, there exists universal constants $c, C > 0$ such that the sample complexities for uniform convergence and agnostic PAC learnability are both

$$c \cdot \frac{\mathrm{VC}(\mathcal{H}) + \ln(1/\delta)}{\varepsilon^2} \leq m_{\mathcal{H}}(\varepsilon,\delta), m_{\mathcal{H}}^{\mathrm{PAC}}(\varepsilon,\delta) \leq C \cdot \frac{\mathrm{VC}(\mathcal{H}) + \ln(1/\delta)}{\varepsilon^2} \ ,$$

whereas the sample complexity for PAC learnability in the realizable setting is

$$c \cdot \frac{\mathrm{VC}(\mathcal{H}) + \ln(1/\delta)}{\varepsilon} \leq \widetilde{m}_{\mathcal{H}}^{\mathrm{PAC}}(\varepsilon,\delta) \leq C \cdot \frac{\mathrm{VC}(\mathcal{H}) \ln(1/\varepsilon) + \ln(1/\delta)}{\varepsilon} \ .$$

It clearly follows that the VC dimension is a crucial dimension that characterizes fundamental properties of binary function classes. Despite the fact that binary classification tasks have numerous use-cases, many fundamental learning problems concern the adoption of predictors whose outputs consist of real values, taking the name of real-valued regression problems. These problems also have multiple applications in fields such as medicine and economics. Therefore, one may wonder about the feasibility of generalizing these notions to classes of real-valued functions. This question has already been thoroughly addressed in the past, leading to the introduction of other combinatorial dimensions. A first notable example is given by Pollard's *pseudodimension* (Pollard, 1990).

Definition 7.7 (Pseudodimension)**.** *Let* $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ *be a class of real-valued functions. A set* $S \subseteq \mathcal{X}$ *of points are* pseudo-shattered *by* $\mathcal{F}$ *if there exists a function* $r \colon S \to \mathbb{R}$ *such that, for every* $B \subseteq S$ *there exists* $f_B \in \mathcal{F}$ *for which*

$$\begin{aligned} f_B(x) &\geq r(x) \qquad \forall x \in B \ , \\ f_B(x) &< r(x) \qquad \forall x \in S \setminus B \ . \end{aligned}$$

The pseudodimension $\mathrm{Pdim}(\mathcal{F})$ *of* $\mathcal{F}$ *is the largest number of points in* $\mathcal{X}$ *that can be pseudo-shattered by* $\mathcal{F}$*. We define* $\mathrm{Pdim}(\mathcal{F}) \coloneqq \infty$ *if every finite set of points can be pseudo-shattered.*

The finiteness of the pseudodimension is known to guarantee the uniform convergence property. The issue is that, as per related definitions of dimensions (Vapnik, 1989), their finiteness does not provide an exact characterization of uniform convergence. The main cause of this problem lies in the excessive generality of their definition. Thus arises the need for more refined, scale-sensitive dimensions of real-valued function families. One such case is constituted by the *fat-shattering dimension* (Kearns and Schapire, 1994), whose definition indeed resembles that of the pseudodimension while relying on a scale parameter.

Definition 7.8 (γ-fat-shattering dimension)**.** *Let* $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ *be a class of real-valued functions and let* $\gamma > 0$ *be a scale (or width) parameter. A set* $S \subseteq \mathcal{X}$ *of points are* γ-shattered *by* $\mathcal{F}$ *if there exists a function* $r \colon S \to \mathbb{R}$ *such that, for every* $B \subseteq S$ *there exists* $f_B \in \mathcal{F}$ *for which*

$$\begin{aligned} f_B(x) &\geq r(x) + \gamma \qquad \forall x \in B \ , \\ f_B(x) &\leq r(x) - \gamma \qquad \forall x \in S \setminus B \ . \end{aligned}$$

The γ-fat-shattering dimension $\mathrm{fat}_{\mathcal{F}}(\gamma)$ *of* $\mathcal{F}$ *is the largest number of points in* $\mathcal{X}$ *that can be* γ*-shattered by* $\mathcal{F}$*. We define* $\mathrm{fat}_{\mathcal{F}}(\gamma) \coloneqq \infty$ *if every finite set of points can be* γ*-shattered.*

Restricting to the world of $[0, 1]$-valued functions,[*] Alon, Ben-David, Cesa-Bianchi, and Haussler (1997) demonstrate the equivalence between finite fat-shattering dimension and the uniform convergence property, showing the crucial role of scale-sensitivity in the complexity measure of the hypothesis class; this equivalence is known for the realizable setting too (Shalev-Shwartz, Shamir, Srebro, and Sridharan, 2010). The fat-shattering dimension has been shown to characterize even agnostic PAC learnability with respect to the absolute loss $\ell(y, y') = |y - y'|$ (Bartlett, Long, and Williamson, 1996, Bartlett and Long, 1998) and the square loss $\ell(y, y') = (y - y')^2$ (Alon et al., 1997). Nonetheless, the landscape of learnability for real-valued function classes in the realizable setting is fundamentally different compared to the binary case (Shalev-Shwartz et al., 2010); e.g., see Attias, Hanneke, Kalavasis, Karbasi, and Velegkas (2023) for a detailed overview and recent results.

While not equivalent under realizability (Attias et al., 2023, Example 1), uniform convergence remains a sufficient condition for the PAC learnability via any ERM algorithm. It thus remains an important problem to study the sample complexity $m_{\mathcal{F}}(\varepsilon, \delta)$ for classes with finite fat-shattering dimension. The best bounds known so far exhibit a spurious polylogarithmic factor in the upper bound, contrarily to analogous bounds in the VC dimension or in the pseudodimension. In the upcoming Chapter 8, we address this problem by deriving improved bounds on the sample complexity of uniform convergence.

[*] Similar results are known for real-valued function classes with any finite range.

Chapter 8

An Improved Uniform Convergence Bound with Fat-Shattering Dimension

The fat-shattering dimension characterizes the uniform convergence property of real-valued function classes. The state-of-the-art upper bounds on the sample complexity (Bartlett and Long, 1995) feature a multiplicative squared logarithmic factor in the accuracy, leaving an open gap with the existing lower bound. In this chapter, by relying on a refined packing number bound by Rudelson and Vershynin (2006), we provide an improved uniform convergence bound that closes this gap.

8.1 Introduction

We first recall the definition of uniform convergence. Given a class of real-valued functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ with domain $\mathcal{X}$, it is said that $\mathcal{F}$ enjoys the uniform convergence property if, for any $\mathcal{X}$-valued i.i.d. process $X, X_1, X_2, \dots$, the sequence of empirical means $\frac{1}{m}\sum_{i=1}^{m} f(X_i)$ converges in probability to its expectation $\mathbb{E}\left[f(X)\right]$, uniformly over any $f \in \mathcal{F}$. Formally, $\mathcal{F}$ enjoys the uniform convergence property if, for every $\varepsilon, \delta > 0$, there exists $\widehat{m}_{\mathcal{F}} \coloneqq \widehat{m}_{\mathcal{F}}(\varepsilon, \delta) \in \mathbb{N}$ such that, for every $m \geq \widehat{m}_{\mathcal{F}}$ and every $\mathcal{X}$-valued i.i.d. process $X, X_1, X_2, \dots$, it holds that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(X_i) - \mathbb{E}\left[f(X)\right]\right| > \varepsilon\right) \leq \delta \ .$$

Among all the functions $\widehat{m}_{\mathcal{F}}$ for which the previous property holds, the smallest of them (pointwise in ε, δ), namely $m^*_{\mathcal{F}}$, is called the sample complexity for the uniform convergence of $\mathcal{F}$.

Uniform convergence is a fundamental tool in learning theory as argued in Chapter 7. Indeed, we can learn any class of functions $\mathcal{F}$ that enjoys uniform convergence via empirical risk minimization. Besides learnability, uniform convergence has notable practical applications. In particular, whenever $\mathcal{F}$ enjoys uniform convergence, we can estimate the risk of any model in $\mathcal{F}$ by computing its empirical risk over the same dataset used to select the model, an aspect that can be especially useful when the model is selected via (heuristic) approximations of algorithms featuring theoretical guarantees. In all these applications, it is crucial to have sharp estimates of the sample complexity $m^*_{\mathcal{F}}$.

A large body of work has focused on identifying conditions implying uniform convergence (Vapnik and Chervonenkis, 1971, Pollard, 1986, Ben-David, Cesa-Bianchi, and Long, 1992, Alon et al., 1997, Bartlett and Long, 1998). In particular, Alon et al. (1997) showed that the fat-shattering dimension—introduced by Kearns and Schapire (1994)—characterizes the uniform convergence property for real-valued functions. However, state-of-the-art estimates on the sample complexity $m^*_{\mathcal{F}}$ for the uniform convergence (Bartlett and Long, 1995) have an accuracy gap of order $\ln^2(1/\varepsilon)$ when compared to the corresponding lower bound for the special case of binary functions (Vapnik and Chervonenkis, 1971). These bounds are extensively used in the current literature (see, e.g., Attias, Kontorovich, and Mansour (2022), Attias and Hanneke (2023), Belkin (2018), Hu, Peale, and Reingold (2022)) and, regrettably, have not been improved ever since.

In this chapter, we close this gap by removing the exogenous $\ln^2(1/\varepsilon)$ factor. Our improvement builds upon a carefully designed chaining argument leveraging sharp estimates (up to constants) for the metric entropy based on the fat-shattering dimension (Rudelson and Vershynin, 2006).

8.2 Preliminaries

Throughout the current chapter we use the following specific notation. For any $m \in \mathbb{N}$ and any $p \in [1, \infty]$, we let $d_p \colon \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$ be the metric defined, for any $g, h \in \mathbb{R}^m$, by $d_p(g, h) \coloneqq \big(\frac{1}{m}\sum_{i=1}^m |g(i) - h(i)|^p\big)^{1/p}$ if $p < \infty$, and by $d_p(g, h) \coloneqq \max_{i\in[m]} |g(i) - h(i)|$ if $p = \infty$. If $(\mathcal{X}, d)$ is a metric space, $\varepsilon > 0$ and $x \in \mathcal{X}$, the closed ball of radius ε centered at x is denoted by $B_\varepsilon(x) \coloneqq \{y \in \mathcal{X} : d(x, y) \le \varepsilon\}$. In this case, for any $\varepsilon > 0$ and any $\widetilde{\mathcal{X}} \subset \mathcal{X}$, we recall that $\widetilde{\mathcal{X}}$ is said to be an ε-net if $\mathcal{X} \subset \bigcup_{x\in\widetilde{\mathcal{X}}} B_\varepsilon(x)$, while $\widetilde{\mathcal{X}}$ is said to be an ε-separated set if $d(x_1, x_2) > \varepsilon$ for any two distinct points $x_1, x_2 \in \widetilde{\mathcal{X}}$. The ε-packing number $\mathcal{P}\big(\mathcal{X}, d, \varepsilon\big)$ of the metric space $(\mathcal{X}, d)$ is the maximum number of elements of any ε-separated set, whenever this maximum exists; otherwise, we set it to ∞. For any $n \in \mathbb{N}^+$, if $A_1, \dots, A_n$ are non-empty subsets of some vector space V, we denote their Minkowski sum using the notation $A_1 + \dots + A_n \coloneqq \{v_1 + \dots + v_n : \forall i \in [n], v_i \in A_i\}$. We recall that a Rademacher random variable (with respect to some underlying probability measure $\mathbb{P}$) is any random variable Z such that $\mathbb{P}(Z = 1) = 1/2 = \mathbb{P}(Z = -1)$.

8.3 The Uniform Convergence Bound

In this section we present our result, which improves on state-of-the-art bounds based on the fat-shattering dimension, and whose proof is deferred to Section 8.5. We start by recalling that we defined the fat-shattering dimension (Definition 7.8) $\mathrm{fat}_{\mathcal{F}}(\gamma)$ as the maximum number of elements that are γ-shattered by $\mathcal{F}$, when this maximum exists; otherwise, we set $\mathrm{fat}_{\mathcal{F}}(\gamma) = \infty$. The fat-shattering dimension is a scale-sensitive generalization to real-valued functions of the classical VC dimension for binary functions. It is well known (Alon et al., 1997) that the finiteness of the fat-shattering dimension for a class of functions $\mathcal{F}$ characterizes the uniform convergence of $\mathcal{F}$.

We are now ready to state our main theorem for classes of functions with bounded range.

Theorem 8.1. *There exists a universal constant $C > 0$ such that the following holds. For any $a < b$, any $\mathcal{F} \subset [a, b]^{\mathcal{X}}$,[*] and any probability measure $\mathbb{P}$, if $X, X_1, X_2, \dots$ is a $\mathbb{P}$-i.i.d. $\mathcal{X}$-valued sequence*

[*] To avoid measurability pathologies (see Ben-David (2015)), and for the sake of simplicity, we carry out the proof under the further assumption that the class $\mathcal{F}$ is countable. This assumption can be greatly relaxed (Alon et al., 1997) relying on measurability conditions such as the "image admissible Suslin" property (Dudley, 1984, Section 10.3.1, page 101). For further discussion on some well-behavedness conditions, see, e.g., Blumer, Ehrenfeucht, Haussler, and Warmuth (1989).

of random variables, then, for every $\varepsilon > 0$ *satisfying* $\mathrm{fat}_{\mathcal{F}}(\varepsilon/32) < \infty$*, every* $\delta \in (0,1)$*, and every* $m \in \mathbb{N}$ *satisfying*

$$m \geq C \cdot \frac{(b-a)^2}{\varepsilon^2}\left(\mathrm{fat}_{\mathcal{F}}(\varepsilon/32) + \ln\frac{1}{\delta}\right) , \tag{8.1}$$

we have that, with probability at least $1-\delta$*,*

$$\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(X_i) - \mathbb{E}\left[f(X)\right]\right| \leq \varepsilon \; .$$

Before presenting a proof of Theorem 8.1, some remarks are in order. The best previously known bound on the sample complexity of the uniform convergence of $[a,b]$-valued functions was of the order of

$$\frac{(b-a)^2}{\varepsilon^2}\left(\mathrm{fat}_{\mathcal{F}}(\varepsilon/5)\ln^2\left(\frac{b-a}{\varepsilon}\right) + \ln\frac{1}{\delta}\right) ; \tag{8.2}$$

see Theorem 9, Eq. (5) in Bartlett and Long (1995).[†] The bound of Equation (8.1) improves on Equation (8.2) by removing the extra $\ln^2\bigl((b-a)/\varepsilon\bigr)$ factor whenever $\mathrm{fat}_{\mathcal{F}}(\gamma)$ is at most of the order $\bigl((b-a)/\gamma\bigr)^{\alpha}$ for some constant $\alpha \geq 0$, which is typical. Indeed, this is the case for linear separators in Hilbert spaces (Mendelson and Schechtman, 2004, Attias and Kontorovich, 2024), polyhedra in Euclidean spaces with finitely many facets and margin (Gottlieb, Kaufman, Kontorovich, and Nivasch, 2022), general Lipschitz functions in bounded metric spaces with finite doubling dimension (Gottlieb, Kontorovich, and Krauthgamer, 2014), uniform Donsker classes of bounded functions (Rudelson and Vershynin, 2006, Theorem 1.1), and many types of finite aggregations of such classes (Attias and Kontorovich, 2024). However, when $\mathrm{fat}_{\mathcal{F}}(\gamma)$ grows as $\exp(1/\gamma)$, then the bound of Equation (8.1) is worse than that of Equation (8.2). Nevertheless, to our knowledge, there are no notable examples in learning theory that achieve this growth rate.

On the one hand, our bound is optimal as for the dependence on ε and δ. Indeed, if $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ and $\varepsilon < 16$, then $\mathrm{fat}_{\mathcal{F}}(\varepsilon/32) = \mathrm{VC}(\mathcal{F})$. In this case, it is well known that to ensure uniform convergence, at least order of $\varepsilon^{-2}(\mathrm{VC}(\mathcal{F}) + \ln(1/\delta))$ samples are required. On the other hand, it shows a worse dependence on the constant factor in the scale of the fat-shattering dimension, which we did not attempt to optimize. It is thus an interesting question whether our analysis could be further refined to improve said constant; we leave this task to future work.

Our proof does not rely on discretizing the range $[a,b]$ as was done in previous work (Bartlett and Long, 1995, Alon et al., 1997). We avoid this use of discretization by relying on a technical lemma (Lemma 8.3) and the breakthrough result of Rudelson and Vershynin (2006), which bounds directly the packing number of certain metric spaces of functions in terms of a quantity depending on the fat-shattering dimension. We further note that an alternative route one might consider goes through Dudley's entropy integral and an application of the metric entropy bound due to Mendelson and Schechtman (2004). It is unclear whether this approach may lead to a better constant in the scale of the fat-shattering dimension, as the bound in Mendelson and Schechtman (2004) leaves it unspecified. We leave the interesting question of finding better constants as an open problem.

Finally, we emphasize that, while our result is the first sample complexity bound with finite fat-shattering dimension that removes the spurious polylogarithmic factor in the accuracy to the

[†] The original bound was stated for $[0,1]$-valued functions, but with a straightforward adaptation of the proof, it can be extended to $[a,b]$-valued functions while preserving the scale of the fat-shattering dimension.

best of our knowledge, a bound on the empirical Rademacher complexity of function classes with finite fat-shattering dimension via Dudley's integral is known (e.g., see Rakhlin and Sridharan (2014, Corollary 12.8)) to take the form

$$\mathcal{R}_m(\mathcal{F};x) \coloneqq \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right] \le \inf_{\alpha\ge 0}\left\{4\alpha + \frac{12}{\sqrt{m}}\int_\alpha^1 \sqrt{c_1 \mathrm{fat}_{\mathcal{F}}(c_2\beta)\ln\frac{2}{\beta}}\,\mathrm{d}\beta\right\} ,$$

where $x = (x_1,\dots,x_m) \in \mathcal{X}^m$ is a collection of m arbitrary points in $\mathcal{X}$, $\sigma = (\sigma_1,\dots,\sigma_m)$ is a vector of m i.i.d. Rademacher random variables, and $c_1, c_2 > 0$ are absolute constants. This quantity is tightly related to the quantities that we introduce in our proofs, but the above inequality alone can only provide a guarantee on $\sup_{x\in\mathcal{X}^m}\mathcal{R}_m(\mathcal{F};x)$, whereas here we are interested in the uniform convergence property of $\mathcal{F}$. Anyhow, the notion of empirical Rademacher complexity can be leveraged to derive such bounds, and this further supports the intuition that analyzing the sample complexity through the Rademacher complexity bound via Dudley's integral might help in improving the constants in our bound, depending on how c_1 and c_2 can be quantified.

8.4 Auxiliary Results

The proof follows the pattern of chaining techniques (Talagrand, 1994) and, for the sake of clarity, it is provided here with the aid of a sequence of technical lemmas. In all the lemmas in the current section, consider $\mathcal{F}, \mathbb{P}, a, b, \varepsilon, \delta$, and the sequence $X, X_1, X_2, \dots,$ to be defined as in the statement of Theorem 8.1. Also, fix $m \in \mathbb{N}^+$.

The first tool we introduce is a symmetrization lemma, which can be proved along the lines of the corresponding symmetrization lemma for $[0,1]$-valued functions; the latter is provided, e.g., by Bartlett and Long (1995, Lemma 10).

Lemma 8.1 (Symmetrization)**.** *If* $m \ge 4\ln(2)\cdot\big((b-a)/\varepsilon\big)^2$*, then*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(X_i) - \mathbb{E}[f(X)]\right| > \varepsilon\right) \le 2\cdot\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \big(f(X_i) - f(X_{m+i})\big)\right| > \frac{\varepsilon}{2}\right) .$$

The importance of symmetrization lies in bounding the probability of the desired event from above in terms of the probability of a similar event based on two i.i.d. samples of same size m. To give some vague intuition, this is performed by essentially replacing the true mean with the empirical mean over the second sample; in a way, we use the latter quantity as an estimate of the former.

The second tool we require is a permutation lemma, which can be proved following the lines of the corresponding permutation lemma found, e.g., in Anthony and Bartlett (1999, Lemma 4.5). Its usefulness consists of removing any dependence on the unknown probability distribution of the i.i.d. process $X_1, X_2, \dots,$ by focusing instead on rescaled i.i.d. Rademacher random variables.

Lemma 8.2 (Permutation lemma)**.** *Let* $Z_1,\dots,Z_m$ *be a family of* $\mathbb{P}$*-independent Rademacher random variables. Then,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \big(f(X_i) - f(X_{m+i})\big)\right| > \frac{\varepsilon}{2}\right) \le \sup_{x_1,\dots,x_{2m}\in\mathcal{X}} \mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(f(x_i) - f(x_{m+i})\big)\right| > \frac{\varepsilon}{2}\right) .$$

In the light of the previous two lemmas, it will be sufficient to estimate the last probability involving the supremum of linear combinations of Rademacher random variables. Luckily enough, the latter has been thoroughly studied in the literature, and there is a plethora of concentration bounds involving this kind of random variables. Most notable is Hoeffding's inequality (Hoeffding, 1963), which we indeed employ later on.

From here onwards, we fix a sequence $Z_1, \dots, Z_m$ of $\mathbb{P}$-i.i.d. Rademacher random variables. For each vector $\mathbf{x} \coloneqq (x_1, \dots, x_{2m}) \in \mathcal{X}^{2m}$, define the family

$$\mathcal{F}(\mathbf{x}) \coloneqq \Big\{ \big(f(x_i)\big)_{i\in[2m]} : f \in \mathcal{F} \Big\}$$

of vectors in $\mathbb{R}^{2m}$ that represent the restrictions of the functions in $\mathcal{F}$ to the sample $\mathbf{x}$. To be more explicit, for any $f \in \mathcal{F}$ there is some $\phi \in \mathcal{F}(\mathbf{x})$ such that $\phi(i) = f(x_i)$ for all $i \in [2m]$.

For each $\mathbf{x} \in \mathcal{X}^{2m}$, we fix an $\varepsilon/8$-separated $\varepsilon/8$-net $\mathcal{F}_\varepsilon(\mathbf{x})$ of the metric space $\big(\mathcal{F}(\mathbf{x}), d_2\big)$. These sets can be built following an iterative procedure where, at each step, we add another element whose distance from any already selected element is greater than $\varepsilon/8$. This procedure terminates after at most $\mathcal{P}\big(\mathcal{F}(\mathbf{x}), d_2, \varepsilon/8\big)$ steps, and we note explicitly that $\mathcal{P}\big(\mathcal{F}(\mathbf{x}), d_2, \varepsilon/8\big) < \infty$ as a consequence of a subsequent technical result (see Lemma 8.6). When this procedure stops, every element of $\mathcal{F}(\mathbf{x})$ is within $\varepsilon/8$ distance from some element in $\mathcal{F}_\varepsilon(\mathbf{x})$.

The next ingredient is a lemma whose purpose is to reduce the problem of bounding the supremum over the whole family $\mathcal{F}$ to another problem where the supremum is taken with respect to the $\varepsilon/8$-separated set $\mathcal{F}_\varepsilon(\mathbf{x})$, over which we plan to implement a chaining procedure. We explicitly note that, to prove the following result, the sole property of $\mathcal{F}_\varepsilon(\mathbf{x})$ we use is that it is an $\varepsilon/8$-net of the metric space $\big(\mathcal{F}(\mathbf{x}), d_2\big)$.

Lemma 8.3. *Given that $\mathcal{F}_\varepsilon(\mathbf{x})$ is an $\varepsilon/8$-net of $(\mathcal{F}(\mathbf{x}), d_2)$ for any $\mathbf{x} \in \mathcal{X}^{2m}$, it holds that*

$$\begin{aligned}\sup_{(x_1,\dots,x_{2m})\in\mathcal{X}^{2m}} &\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(f(x_i) - f(x_{m+i})\big)\right| > \frac{\varepsilon}{2}\right) \\ &\le \sup_{\mathbf{x}\in\mathcal{X}^{2m}} \mathbb{P}\left(\sup_{\phi\in\mathcal{F}_\varepsilon(\mathbf{x})}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(\phi(i) - \phi(m+i)\big)\right| > \frac{\varepsilon}{4}\right) .\end{aligned}$$

Proof. For each sample $\mathbf{x} \coloneqq (x_1, \dots, x_{2m}) \in \mathcal{X}^{2m}$ and each vector $\xi \coloneqq (\xi_1, \dots, \xi_m) \in \{-1, 1\}^m$, select $\phi_{\mathbf{x},\xi} \in \mathcal{F}(\mathbf{x})$ such that

$$\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\phi_{\mathbf{x},\xi}(i) - \phi_{\mathbf{x},\xi}(m+i)\big)\right| > \varepsilon/2$$

whenever it is feasible to do so, otherwise select $\phi_{\mathbf{x},\xi} \in \mathcal{F}(\mathbf{x})$ arbitrarily. Notice that, if it holds that

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(f(x_i) - f(x_{m+i})\big)\right| > \varepsilon/2$$

then also

$$\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\phi_{\mathbf{x},\xi}(i) - \phi_{\mathbf{x},\xi}(m+i)\big)\right| > \varepsilon/2$$

holds, and vice versa.

For each $\mathbf{x} \in \mathcal{X}^{2m}$ and each $\xi \in \{-1,1\}^m$, let $\varphi^{\varepsilon}_{\mathbf{x},\xi} \in \mathcal{F}_\varepsilon(\mathbf{x})$ be such that

$$d_2(\phi_{\mathbf{x},\xi}, \varphi^{\varepsilon}_{\mathbf{x},\xi}) \le \frac{\varepsilon}{8} \,. \tag{8.3}$$

Then, for each $\mathbf{x} := (x_1, \dots, x_{2m}) \in \mathcal{X}^{2m}$ we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(f(x_i) - f(x_{m+i})\big)\right| > \frac{\varepsilon}{2}\right) \\
&= \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(f(x_i) - f(x_{m+i})\big)\right| > \frac{\varepsilon}{2}\right\} \\
&= \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\phi_{\mathbf{x},\xi}(i) - \phi_{\mathbf{x},\xi}(m+i)\big)\right| > \frac{\varepsilon}{2}\right\} \\
&\le \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\varphi^{\varepsilon}_{\mathbf{x},\xi}(i) - \varphi^{\varepsilon}_{\mathbf{x},\xi}(m+i)\big)\right| > \frac{\varepsilon}{4}\right\} + \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{d_1(\phi_{\mathbf{x},\xi}, \varphi^{\varepsilon}_{\mathbf{x},\xi}) > \frac{\varepsilon}{8}\right\} \\
&\le \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\varphi^{\varepsilon}_{\mathbf{x},\xi}(i) - \varphi^{\varepsilon}_{\mathbf{x},\xi}(m+i)\big)\right| > \frac{\varepsilon}{4}\right\} + \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{d_2(\phi_{\mathbf{x},\xi}, \varphi^{\varepsilon}_{\mathbf{x},\xi}) > \frac{\varepsilon}{8}\right\} \\
&= \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\varphi^{\varepsilon}_{\mathbf{x},\xi}(i) - \varphi^{\varepsilon}_{\mathbf{x},\xi}(m+i)\big)\right| > \frac{\varepsilon}{4}\right\} \\
&\le \frac{1}{2^m}\sum_{\xi\in\{-1,1\}^m} \mathbb{I}\left\{\sup_{\phi\in\mathcal{F}_\varepsilon(\mathbf{x})}\left|\frac{1}{m}\sum_{i=1}^m \xi_i\big(\phi(i) - \phi(m+i)\big)\right| > \frac{\varepsilon}{4}\right\} \\
&= \mathbb{P}\left(\sup_{\phi\in\mathcal{F}_\varepsilon(\mathbf{x})}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(\phi(i) - \phi(m+i)\big)\right| > \frac{\varepsilon}{4}\right) \,,
\end{aligned}$$

where the second equality follows by definition of $\phi_{\mathbf{x},\xi}$, the second inequality follows from the fact that $d_1(\mathbf{x},\mathbf{y}) \le d_2(\mathbf{x},\mathbf{y})$ for all vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^{2m}$, and the third equality holds by definition of $\varphi^{\varepsilon}_{\mathbf{x},\xi}$ and the fact that $\mathcal{F}_\varepsilon(\mathbf{x})$ is an $\varepsilon/8$-net. □

Leveraging Hoeffding's inequality (Hoeffding, 1963), we can prove the following lemma, which can be viewed as a multi-scale concentration inequality.

Lemma 8.4. *Let $L \in \mathbb{N}$. Consider $\varepsilon_0, \dots, \varepsilon_L > 0$ such that $\sum_{j=0}^L \varepsilon_j \le \varepsilon/4$. For each $\mathbf{x} \in \mathcal{X}^{2m}$, let $\widetilde{\mathcal{H}}_0(\mathbf{x}), \dots, \widetilde{\mathcal{H}}_L(\mathbf{x}) \subset \mathcal{F}_\varepsilon(\mathbf{x})$ such that $\mathcal{F}_\varepsilon(\mathbf{x}) \subset \widetilde{\mathcal{H}}_0(\mathbf{x}) + \cdots + \widetilde{\mathcal{H}}_L(\mathbf{x})$. Then,*

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}^{2m}} \mathbb{P}\left(\sup_{\phi\in\mathcal{F}_\varepsilon(\mathbf{x})}\left|\frac{1}{m}\sum_{i=1}^m Z_i\big(\phi(i) - \phi(m+i)\big)\right| > \frac{\varepsilon}{4}\right) \\
&\qquad \le 2\sum_{j=0}^L \sup_{\mathbf{x}\in\mathcal{X}^{2m}} \sum_{h\in\widetilde{\mathcal{H}}_j(\mathbf{x})} \exp\left(-\frac{\varepsilon_j^2 m^2}{2\sum_{i=1}^m \big(h(i) - h(m+i)\big)^2}\right) .
\end{aligned}$$

Proof. Fix $\mathbf{x} := (x_1, \dots, x_{2m}) \in \mathcal{X}^{2m}$. For each $\phi \in \mathcal{F}_\varepsilon(\mathbf{x})$, since $\mathcal{F}_\varepsilon(\mathbf{x}) \subset \widetilde{\mathcal{H}}_0(\mathbf{x}) + \cdots + \widetilde{\mathcal{H}}_L(\mathbf{x})$, we can (and do) select $h_0^\phi \in \widetilde{\mathcal{H}}_0(\mathbf{x}), \dots, h_L^\phi \in \widetilde{\mathcal{H}}_L(\mathbf{x})$ such that $\phi = h_0^\phi + \cdots + h_L^\phi$. Furthermore, notice that for each $(\xi_1, \dots, \xi_m) \in \{-1, 1\}^m$ it holds

$$\left\{ \sup_{\phi \in \mathcal{F}_\varepsilon(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(\phi(i) - \phi(m+i)\big) \right| > \frac{\varepsilon}{4} \right\} \subset \bigcup_{j=0}^{L} \left\{ \sup_{\phi \in \mathcal{F}_\varepsilon(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(h_j^\phi(i) - h_j^\phi(m+i)\big) \right| > \varepsilon_j \right\} .$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\phi \in \mathcal{F}_\varepsilon(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} Z_i \big(\phi(i) - \phi(m+i)\big) \right| > \frac{\varepsilon}{4} \right) \\
&= \frac{1}{2^m} \sum_{\xi \in \{-1,1\}^m} \mathbb{I}\left\{ \sup_{\phi \in \mathcal{F}_\varepsilon(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(\phi(i) - \phi(m+i)\big) \right| > \frac{\varepsilon}{4} \right\} \\
&\le \sum_{j=0}^{L} \frac{1}{2^m} \sum_{\xi \in \{-1,1\}^m} \mathbb{I}\left\{ \sup_{\phi \in \mathcal{F}_\varepsilon(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(h_j^\phi(i) - h_j^\phi(m+i)\big) \right| > \varepsilon_j \right\} \\
&\le \sum_{j=0}^{L} \frac{1}{2^m} \sum_{\xi \in \{-1,1\}^m} \mathbb{I}\left\{ \sup_{h \in \widetilde{\mathcal{H}}_j(\mathbf{x})} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(h(i) - h(m+i)\big) \right| > \varepsilon_j \right\} \\
&\le \sum_{j=0}^{L} \sum_{h \in \widetilde{\mathcal{H}}_j(\mathbf{x})} \frac{1}{2^m} \sum_{\xi \in \{-1,1\}^m} \mathbb{I}\left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i \big(h(i) - h(m+i)\big) \right| > \varepsilon_j \right\} \\
&= \sum_{j=0}^{L} \sum_{h \in \widetilde{\mathcal{H}}_j(\mathbf{x})} \mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} Z_i \big(h(i) - h(m+i)\big) \right| > \varepsilon_j \right) =: (\star) .
\end{aligned}$$

Now, for each $j \in \{0, 1, \dots, L\}$, and each $h \in \widetilde{\mathcal{H}}_j(\mathbf{x})$, the sequence $\big(W_i^h(\mathbf{x})\big)_{i \in [m]}$ is a sequence of bounded zero-mean independent random variables

$$W_i^h(\mathbf{x}) \coloneqq Z_i \big(h(i) - h(m+i)\big) \qquad \forall i \in [m] .$$

More precisely, we notice that for each $i \in [m]$ it holds that

$$-|h(i) - h(m+i)| \le W_i^h(\mathbf{x}) \le |h(i) - h(m+i)| .$$

By Hoeffding's inequality (Hoeffding, 1963), we obtain

$$(\star) \le 2 \sum_{j=0}^{L} \sum_{h \in \widetilde{\mathcal{H}}_j(\mathbf{x})} \exp\left(- \frac{\varepsilon_j^2 m^2}{2 \sum_{i=1}^{m} \big(h(i) - h(m+i)\big)^2} \right) .$$

Taking the supremum over $\mathbf{x} \in \mathcal{X}^{2m}$ on the first and the last term of this chain of inequalities, and switching said supremum with the sum over $j \in \{0, \dots, L\}$ on the last expression, we conclude the proof. □

Now, for each $\mathbf{x} \in \mathcal{X}^{2m}$, we need to build a suitable sequence $\widetilde{\mathcal{H}}_0(\mathbf{x}), \dots, \widetilde{\mathcal{H}}_L(\mathbf{x})$ to which we want to apply Lemma 8.4. Our choice for such a sequence follows a standard chaining argument (Talagrand, 1994). From now on, we fix $l \coloneqq \lfloor \log_2\big((b-a)/\varepsilon\big) \rfloor + 4$ and, for each $\mathbf{x} \in \mathcal{X}^{2m}$, we define by induction over $j \in \{0, 1, \dots, l\}$ the sets $\mathcal{G}_0(\mathbf{x}), \dots, \mathcal{G}_l(\mathbf{x}) \subset \mathcal{F}_\varepsilon(\mathbf{x})$ in the following way:

- $\mathcal{G}_0(\mathbf{x}) \coloneqq \{g_0\}$, for an arbitrary choice of $g_0 \in \mathcal{F}_\varepsilon(\mathbf{x})$.
- For any $j \in [l]$, we initially define $\mathcal{G}_j(\mathbf{x}) \coloneqq \mathcal{G}_{j-1}(\mathbf{x})$. Then, iteratively, we add elements $\phi \in \mathcal{F}_\varepsilon(\mathbf{x})$ to $\mathcal{G}_j(\mathbf{x})$ for which $d_2(\phi, g) > (b-a)\cdot 2^{-j}$ for every other element g already in $\mathcal{G}_j(\mathbf{x})$. The procedure is carried out until we can no longer add other elements.[‡]

Notice that, by construction, for each $\mathbf{x} \in \mathcal{X}^{2m}$ and each $j \in \{0,1,\dots,l\}$, the set $\mathcal{G}_j(\mathbf{x})$ is a $(b-a)\cdot 2^{-j}$-net and $(b-a)\cdot 2^{-j}$-separated set of $\big(\mathcal{F}_\varepsilon(\mathbf{x}), d_2\big)$, which implies that, for any $\phi \in \mathcal{F}_\varepsilon(\mathbf{x})$, there exists—and hence we can and do select—an element $\pi_j(\phi) \in \mathcal{G}_j(\mathbf{x})$ such that $d_2\big(\phi, \pi_j(\phi)\big) \le (b-a)\cdot 2^{-j}$. For each $\mathbf{x} \in \mathcal{X}^{2m}$, we finally define

$$\begin{aligned} &\mathcal{H}_0(\mathbf{x}) \coloneqq \mathcal{G}_0(\mathbf{x})\ , \\ &\mathcal{H}_j(\mathbf{x}) \coloneqq \big\{g - \pi_{j-1}(g) : g \in \mathcal{G}_j(\mathbf{x})\big\},\ \forall j \in [l]\ . \end{aligned} \tag{8.4}$$

The relevant properties of this sequence of sets are summarized by the following lemma.

Lemma 8.5. *For any $\mathbf{x} \in \mathcal{X}^{2m}$, consider $\mathcal{H}_0(\mathbf{x}), \dots, \mathcal{H}_l(\mathbf{x})$ defined as in* (8.4). *It holds that*

1. $\mathcal{F}_\varepsilon(\mathbf{x}) \subset \mathcal{H}_0(\mathbf{x}) + \cdots + \mathcal{H}_l(\mathbf{x})$ *;*
2. $\forall j \in \{0,\dots,l\}$, $\forall h \in \mathcal{H}_j(\mathbf{x})$, $\sum_{i=1}^m \big(h(i) - h(m+i)\big)^2 \le 16m(b-a)^2 4^{-j}$ *;*
3. $\forall j \in \{0,\dots,l\}$, $|\mathcal{H}_j(\mathbf{x})| \le \mathcal{P}\big(\mathcal{F}_\varepsilon(\mathbf{x}), d_2, (b-a)\cdot 2^{-j}\big)$ *.*

Proof. Fix an arbitrary $\mathbf{x} \in \mathcal{X}^{2m}$. First, we prove that $\mathcal{F}_\varepsilon(\mathbf{x}) = \mathcal{G}_l(\mathbf{x})$. We know that $\mathcal{G}_l(\mathbf{x}) \subset \mathcal{F}_\varepsilon(\mathbf{x})$ by construction. Consider now any $\phi \in \mathcal{F}_\varepsilon(\mathbf{x})$. By our choice of l, we have that

$$d_2(\phi, \pi_l(\phi)) \le (b-a)\cdot 2^{-l} \le \frac{\varepsilon}{8}\ .$$

If $\phi \ne \pi_l(\phi)$ were true then we would have that $d_2(\phi, \pi_l(\phi)) > \varepsilon/8$ because $\mathcal{F}_\varepsilon(\mathbf{x})$ is an $\varepsilon/8$-separated set, which is a contradiction. Then, it must hold that $\phi = \pi_l(\phi)$ and thus $\mathcal{F}_\varepsilon(\mathbf{x}) \subset \mathcal{G}_l(\mathbf{x})$.

Second, for each $j \in \{0,\dots,l\}$ and each $g \in \mathcal{G}_j(\mathbf{x})$, we prove that there exist $h_0 \in \mathcal{H}_0(\mathbf{x}), \dots, h_j \in \mathcal{H}_j(\mathbf{x})$ such that $g = \sum_{k=0}^j h_k$. We prove this claim by induction over $j \in \{0,1,\dots,l\}$. The base case $j=0$ is trivial. Assuming the claim holds for $j < l$ and that $g \in \mathcal{G}_{j+1}(\mathbf{x})$, we have that $\pi_j(g) = \sum_{k=0}^j h_k$ for some $h_0 \in \mathcal{H}_0(\mathbf{x}), \dots, h_j \in \mathcal{H}_j(\mathbf{x})$ by the inductive hypothesis, and thus $g = (g - \pi_j(g)) + \pi_j(g) = \sum_{k=0}^{j+1} h_k$ for $h_{j+1} \coloneqq g - \pi_j(g) \in \mathcal{H}_{j+1}(\mathbf{x})$, hence proving the claim.

The above property for the specific case of $j = l$ implies that $\mathcal{F}_\varepsilon(\mathbf{x}) = \mathcal{G}_l(\mathbf{x}) \subset \mathcal{H}_0(\mathbf{x}) + \cdots + \mathcal{H}_l(\mathbf{x})$. This shows that the first point in the statement holds.

Consider now any $j \in [l]$ and any $h \in \mathcal{H}_j(\mathbf{x})$. By definition of h, there exists some $g \in \mathcal{G}_j(\mathbf{x})$ such that $h = g - \pi_{j-1}(g)$. Then,

$$\begin{aligned} \sum_{i=1}^m \big(h(i) - h(m+i)\big)^2 &\le 2\sum_{i=1}^{2m} h(i)^2 = 4m\cdot d_2(g, \pi_{j-1}(g))^2 \\ &\le 4m(b-a)^2 \cdot 2^{-2(j-1)} = 16m(b-a)^2 \cdot 4^{-j}\ . \end{aligned}$$

[‡] Note that this process has to terminate, since $|\mathcal{F}_\varepsilon(\mathbf{x})| \le \mathcal{P}\big(\mathcal{F}(\mathbf{x}), d_2, \varepsilon/8\big) < \infty$, as a consequence of Lemma 8.6.

On the other hand, for the case $j = 0$ we have that

$$\sum_{i=1}^{m} \big(g_0(i) - g_0(m+i)\big)^2 \leq m(b-a)^2 ,$$

thus proving the second point of the statement.

Finally, noting that the map $\pi_j \colon \mathcal{G}_j(\mathbf{x}) \to \mathcal{H}_j(\mathbf{x})$ is a surjective map for each $j \in \{0, 1, \dots, l\}$, we have that $|\mathcal{H}_j(\mathbf{x})| \leq |\mathcal{G}_j(\mathbf{x})|$, which, together with the fact that $\mathcal{G}_j(\mathbf{x})$ is a $(b-a) \cdot 2^{-j}$-separated set of $\big(\mathcal{F}_\varepsilon(\mathbf{x}), d_2\big)$, yields the third point of the statement. □

The last ingredient to prove the main theorem is the following lemma, which is an immediate corollary of Rudelson and Vershynin (2006, Corollary 5.4), observing that the metric d_2 is the natural metric on $L^2(\mu)$ when the underlying measure μ is the uniform probability measure on the set $[m]$.

Lemma 8.6. *There exists a universal constant* $\widetilde{C} > 0$ *for which the following holds. For any* $\mathbf{x} \in \mathcal{X}^{2m}$ *and any* $0 < \zeta < (b-a)/2$ *such that* $\mathrm{fat}_{\mathcal{F}}(\zeta/2) < \infty$,

$$\mathcal{P}(\mathcal{F}(\mathbf{x}), d_2, \zeta) \leq \left(\frac{2(b-a)}{\zeta}\right)^{\widetilde{C}\mathrm{fat}_{\mathcal{F}}(\zeta/2)} .$$

Proof. Let $\widetilde{C}, \widetilde{c}$ be universal constants as in Rudelson and Vershynin (2006, Corollary 5.4). Define $\mathcal{F}'(\mathbf{x}) \coloneqq \{(\phi - a)/(b-a) : \phi \in \mathcal{F}(\mathbf{x})\}$ and $\zeta' \coloneqq \zeta/(b-a) \in (0, 1/2)$. A direct computation shows that $\mathrm{fat}_{\mathcal{F}'(\mathbf{x})}\big(\widetilde{c}\zeta'\big) = \mathrm{fat}_{\mathcal{F}(\mathbf{x})}\big(\widetilde{c}\zeta\big)$ and $\mathcal{P}(\mathcal{F}'(\mathbf{x}), d_2, \zeta') = \mathcal{P}(\mathcal{F}(\mathbf{x}), d_2, \zeta)$. Now, further observing that $\mathcal{F}'(\mathbf{x})$ is 1-bounded in d_∞ since $\mathcal{F}(\mathbf{x}) + \{-a\}$ is $(b-a)$-bounded in d_∞, we may apply Rudelson and Vershynin (2006, Corollary 5.4) to $\mathcal{F}'(\mathbf{x})$ with ζ' to infer that

$$\mathcal{P}(\mathcal{F}(\mathbf{x}), d_2, \zeta) = \mathcal{P}(\mathcal{F}'(\mathbf{x}), d_2, \zeta') \leq \left(\frac{1}{\widetilde{c}\zeta'}\right)^{\widetilde{C}\mathrm{fat}_{\mathcal{F}'(\mathbf{x})}(\widetilde{c}\zeta')} = \left(\frac{b-a}{\widetilde{c}\zeta}\right)^{\widetilde{C}\mathrm{fat}_{\mathcal{F}(\mathbf{x})}(\widetilde{c}\zeta)} .$$

Finally, we arrive at the conclusion by observing that $\mathrm{fat}_{\mathcal{F}(\mathbf{x})}(\widetilde{c}\zeta) \leq \mathrm{fat}_{\mathcal{F}}(\widetilde{c}\zeta)$ and that, in our specific case, Rudelson and Vershynin (2006, Corollary 5.4) holds with the choice $\widetilde{c} \coloneqq 1/2$. □

8.5 Proof of the Main Result

We are now ready to present the proof of our main result, that is, the uniform convergence bound in Theorem 8.1. In this proof, each of the technical results from the previous section has a relevant role.

Proof of Theorem 8.1. We may assume that $\varepsilon < b - a$ without loss of generality, since otherwise

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(X_i) - \mathbb{E}\left[f(X)\right] \right| \leq \varepsilon \right) = 1 .$$

Pick $\widetilde{C}$ as the universal constant whose existence is stated in Lemma 8.6. Let $\kappa \coloneqq \mathrm{fat}_{\mathcal{F}}(\varepsilon/32)$ and $R \coloneqq b - a$. Furthermore, define $c_j \coloneqq \frac{1}{44}\sqrt{4^{2-j}(j+1)}$ and $\rho_j \coloneqq \min\big\{2^{-(j+1)}, 1/8\big\}$ for each $j \in \{0, 1, \dots, l\}$. Then, by carefully combining the technical lemmas introduced thus far, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(X_i)-\mathbb{E}\left[f(X)\right]\right|>\varepsilon\right)\\
&\overset{\text{(a)}}{\le} 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}(f(X_i)-f(X_{m+i}))\right|>\frac{\varepsilon}{2}\right)\\
&\overset{\text{(b)}}{\le} 2\sup_{(x_1,\dots,x_{2m})\in\mathcal{X}^{2m}}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m}Z_i(f(x_i)-f(x_{m+i}))\right|>\frac{\varepsilon}{2}\right)\\
&\overset{\text{(c)}}{\le} 2\sup_{\mathbf{x}\in\mathcal{X}^{2m}}\mathbb{P}\left(\sup_{\phi\in\mathcal{F}_\varepsilon(\mathbf{x})}\left|\frac{1}{m}\sum_{i=1}^{m}Z_i(\phi(i)-\phi(m+i))\right|>\frac{\varepsilon}{4}\right)\\
&\overset{\text{(d)}}{\le} 4\sum_{j=0}^{l}\sup_{\mathbf{x}\in\mathcal{X}^{2m}}\sum_{h\in\mathcal{H}_j(\mathbf{x})}\exp\left(-\frac{\varepsilon^2c_j^2m^2}{2\sum_{i=1}^{m}\big(h(i)-h(m+i)\big)^2}\right)\\
&\overset{\text{(e)}}{\le} 4\sum_{j=0}^{l}\sup_{\mathbf{x}\in\mathcal{X}^{2m}}|\mathcal{H}_j(\mathbf{x})|\exp\left(-\frac{4^j\varepsilon^2c_j^2m}{32R^2}\right)\\
&\overset{\text{(f)}}{\le} 4\sum_{j=0}^{l}\sup_{\mathbf{x}\in\mathcal{X}^{2m}}\mathcal{P}\big(\mathcal{F}(\mathbf{x}),d_2,2^{-j}R\big)\exp\left(-\frac{4^j\varepsilon^2c_j^2m}{32R^2}\right)\\
&\overset{\text{(g)}}{\le} 4\sum_{j=0}^{l}\rho_j^{-\widetilde{C}\mathrm{fat}_{\mathcal{F}}(\rho_jR)}\cdot\exp\left(-\frac{4^j\varepsilon^2c_j^2m}{32R^2}\right)\\
&\overset{\text{(h)}}{\le} 4\cdot 8^{\widetilde{C}\kappa}\sum_{j=0}^{l}\exp\left(j\cdot\widetilde{C}\kappa\ln(2)-\frac{4^j\varepsilon^2c_j^2m}{32R^2}\right)\\
&\overset{\text{(i)}}{=} 4\cdot 8^{\widetilde{C}\kappa}\cdot\exp\left(-\frac{\varepsilon^2m}{2\cdot 44^2R^2}\right)\sum_{j=0}^{l}\exp\left(j\left(\widetilde{C}\kappa\ln(2)-\frac{\varepsilon^2m}{2\cdot 44^2R^2}\right)\right)\\
&\overset{\text{(j)}}{\le} 8^{\widetilde{C}\kappa+1}\cdot\exp\left(-\frac{\varepsilon^2m}{2\cdot 44^2R^2}\right), \qquad (8.5)
\end{aligned}$$

where the marked inequalities respectively follow as explained (in order) by the following points:

(a) By Lemma 8.1, assuming $m\ge 4\ln(2)\cdot(R/\varepsilon)^2$.

(b) By Lemma 8.2, in the light of the fact that $Z_1,\dots,Z_m$ is a family of $\mathbb{P}$-independent Rademacher random variables.

(c) By Lemma 8.3.

(d) By Lemma 8.4 with the choice $L:=l$, for each $j\in\{0,1,\dots,l\}$, $\widetilde{\mathcal{H}}_j(\mathbf{x}):=\mathcal{H}_j(\mathbf{x})$ and $\varepsilon_j:=c_j\varepsilon$. Note that the assumptions of Lemma 8.4 are satisfied as a consequence of the first point of Lemma 8.5, and the fact that $\sum_{j=0}^{l}c_j\le 1/4$.

(e) By the second point in Lemma 8.5.

(f) By the third point in Lemma 8.5 and the fact that $\mathcal{P}\big(\mathcal{F}_\varepsilon(\mathbf{x}),d_2,2^{-j}R\big)\le\mathcal{P}\big(\mathcal{F}(\mathbf{x}),d_2,2^{-j}R\big)$.

(g) By Lemma 8.6. Specifically, if $j\ge 2$, we set $\zeta:=R\cdot 2^{-j}$ (and upper bound). Instead, if $j\in\{0,1\}$, we first bound $\mathcal{P}\left(\mathcal{F}(\mathbf{x}),d_2,2^{-j}R\right)$ from above with $\mathcal{P}\left(\mathcal{F}(\mathbf{x}),d_2,R/4\right)$, then apply the lemma setting $\zeta:=R/4$ (and upper bound again).

(h) By the fact that the function $\gamma \mapsto \mathrm{fat}_{\mathcal{F}}(\gamma)$ is monotonically non-increasing, and $\rho_l = 2^{-(l+1)} \geq \varepsilon/(32R)$ because $l = \lfloor \log_2\big(R/\varepsilon\big) \rfloor + 4$.

(i) By our choice of $c_0, \dots, c_l$.

(j) Assuming $m \geq 2 \cdot 44^2 \ln(2) \cdot (R/\varepsilon)^2 \cdot \big(\widetilde{C}\kappa + 1\big)$.

Finally, observe that the right-hand side of Equation (8.5) is at most δ for

$$m \geq \frac{2 \cdot 44^2 \cdot R^2}{\varepsilon^2} \left(\Big(\widetilde{C}\kappa + 1\Big) \ln(8) + \ln \frac{1}{\delta} \right) . \tag{8.6}$$

Therefore, for a sufficiently large universal constant $C > 0$ in the statement of the theorem, we see that any value of m that satisfies Equation (8.1) suffices to guarantee Equation (8.6) and the assumptions in items (a) and (j), concluding the proof. □

Chapter 9

A Theory of Interpretable Approximations

The growing demand for machine learning models that are *interpretable* by humans has become more prominent in recent years. In this chapter, we study such questions by introducing *interpretable approximations*, a notion that captures the idea of approximating a target concept c by a small aggregation of concepts from some base class $\mathcal{H}$. In particular, we consider the approximation of a binary concept c by decision trees based on a simple class $\mathcal{H}$ (e.g., of bounded VC dimension), and use the tree depth as a measure of complexity. Our primary contribution is a remarkable trichotomy for any pair of $\mathcal{H}$ and c: either c cannot be approximated by $\mathcal{H}$, or c can be approximated by $\mathcal{H}$ but without any universal complexity rate, or else c can be approximated by $\mathcal{H}$ with a complexity bounded from above by a constant, for any data distribution and any desired accuracy, which depends only on $\mathcal{H}$ and c. This taxonomy stands in stark contrast to the landscape of supervised classification, which offers a complex array of distribution-free and universally learnable scenarios. We show that, in the case of interpretable approximations, even a slightly nontrivial a-priori guarantee on the complexity of approximations implies approximations with constant (distribution-free and accuracy-free) complexity. We extend our trichotomy to classes $\mathcal{H}$ of unbounded VC dimension and give characterizations of interpretability based on the algebra generated by $\mathcal{H}$.

9.1 Introduction

Many machine learning techniques, such as deep neural networks, produce large and complex models whose inner workings are difficult to grasp. In sectors such as healthcare and law enforcement, where the stakes of automated decisions are high, this is a serious problem: complex models make it hard to explain the rationale behind an outcome, or why two similar inputs produce different outcomes. In those cases, *interpretable* models may become the preferred choice. Although there is an ongoing debate around the notion of interpretability (Erasmus, Brunet, and Fisher, 2021), decision trees are typically considered as the quintessential example of interpretable models (Molnar, 2022): ones that favor a transparent decision-making process, and that allow users to understand how individual features influence predictions. A line of research in this area studies the extent to which small decision trees can approximate some specific learning models, such as neural networks (Craven and Shavlik, 1995) and k-means classifiers (Dasgupta, Frost, Moshkovitz, and Rashtchian, 2020).

Inspired by these results, we develop a general theory of interpretability viewed as approximability via simple decision trees. Our guiding principle can be summarized as follows.

Interpretable approximations = Small aggregations of simple hypotheses

In analogy with PAC learning, we focus on binary classification tasks and view a classifier (e.g., a neural network) as a concept $c \subseteq \mathcal{X}$, where $\mathcal{X}$ is the data domain. Now let $\mathcal{H} \subseteq 2^{\mathcal{X}}$ be a family of simple hypotheses, for instance decision stumps or halfspaces. Our goal is to understand how well c can be approximated by aggregating a small set of elements in $\mathcal{H}$. To formalize this goal in the language of decision trees we introduce two notions. First, we say that c is *approximable* by $\mathcal{H}$ if, under any given data distribution, there exists a finite decision tree using splitting functions from $\mathcal{H}$ that approximates c arbitrarily well. Moreover, if the approximation can be always achieved using a shallow tree, we say that c is *interpretable* by $\mathcal{H}$. It is easy to see that, depending on c and $\mathcal{H}$, one may have interpretability, approximability but not interpretability, or even non-approximability. In Section 9.4, we give explicit examples of pairs $(c, \mathcal{H})$ for each one of the three above cases.

Note that in this initial investigation of the general structure of interpretable approximations we focus on the fundamental question of what conditions ensure the existence of accurate approximations and interpretations. Important topics, such as the informational or computational complexity of obtaining accurate interpretations, are not addressed in this chapter. Note also that we do not make any specific assumption on the data distribution P. Our approach is thus in line with standard notions and theories in machine learning—e.g., universal Bayes consistency (Devroye, Györfi, and Lugosi, 2013), PAC learnability (Shalev-Shwartz and Ben-David, 2014), and universal learnability (Bousquet, Hanneke, Moran, van Handel, and Yehudayoff, 2021)—as it encompasses both distribution-free and distribution-dependent guarantees.

While our primary focus is not algorithmic, our results reveal profound connections within the algorithmic framework of boosting. Indeed, there is a clear relationship between boosting, which involves the aggregation of weak hypotheses to learn a target concept, and interpretable approximation, which concerns the aggregation of simple hypotheses to approximate a target concept. However, our findings uncovers and exploits deeper links at a technical level. In particular, our general construction that gives decision trees whose depth depends logarithmically on the accuracy is based on boosting decision trees, and its analysis uses potential functions from this line of work. Our improved bound for the case of VC classes, which provides approximating decision trees with constant (accuracy- and distribution-free) depth, is somewhat more subtle; it is also based on a boosting perspective, this time using majority-vote based algorithms and Von Neumann's minimax theorem. However, to eliminate the dependency on the accuracy, we utilize tools from VC theory, particularly uniform convergence (which we have already defined in Chapter 7 and further investigated for real-valued function classes in Chapter 8).

9.1.1 Contributions

Here we provide a summary of the contributions contained in the current chapter.

Degrees of interpretability (Section 9.4). We introduce our learning-theoretic notions of approximability and interpretability. Informally speaking, we use the depth of the shallowest approximating tree to measure the extent to which a certain concept c is interpretable by a given class $\mathcal{H}$ (e.g., hyperplanes or single features). Approximability is our weakest notion, as we do not constrain the rate at which the shallowest approximating tree grows as a function of the desired accuracy. Our strongest notion is instead interpretability with a tree depth that is constant with respect to both accuracy and data distribution. In between these two extremes, a wide variety of behaviors is possible, as the tree depth may grow at different rates that may be uniform, or depend on the data distribution (similarly to the distinction between PAC learning and universal learning).

Collapse of the degrees (Section 9.5). We prove that the range of possible behaviors collapses dramatically, and only three cases are actually possible: c is uniformly interpretable by $\mathcal{H}$, c is approximable but not interpretable by $\mathcal{H}$, c is not approximable by $\mathcal{H}$. If the class $\mathcal{H}$ of splits has bounded VC dimension, which conforms to our request that $\mathcal{H}$ be simple, we show that whenever c is interpretable (possibly with a distribution-dependent rate) then it is uniformly interpretable by $\mathcal{H}$ *at constant depth*. This means that, for every data distribution P and every accuracy $\varepsilon > 0$, there exists an $\mathcal{H}$-based decision tree that approximates c with accuracy ε and whose depth is bounded *by a constant* depending only on c and $\mathcal{H}$ (but not on P or ε). Thus, whenever c is interpretable at some arbitrary rate, it is in fact interpretable at a constant rate. We show a similar collapse for classes $\mathcal{H}$ of unbounded VC dimension: in this case, we show that interpretability collapses to uniform interpretability at logarithmic depth $\mathcal{O}\big(\log \frac{1}{\varepsilon}\big)$.

Algebraic characterizations (Section 9.6). We prove that the trichotomy described above can be characterized in terms of algebras and closures over $\mathcal{H}$. For example, we show that if $\mathcal{H}$ has bounded VC dimension, then c is interpretable at constant depth if and only if c is in the algebra generated by the *closure* of $\mathcal{H}$, i.e., the family of all the concepts that can be approximated arbitrarily well by single hypotheses of $\mathcal{H}$. We also present a simpler characterization when the domain $\mathcal{X}$ is countable.

Extension to other complexity measures (Section 9.7). Finally, we exploit the equivalence between $\mathcal{H}$-based decision trees and Boolean formulae over $\mathcal{H}$ to show that the trichotomy above holds for a large class of complexity measures, including not only tree-depth but also, for example, circuit size. In particular, we show that for any complexity measure in our class, interpretability collapses to uniform interpretability at constant complexity rate for VC classes and at polynomial complexity rate for non-VC classes.

9.2 Related Work

According to Molnar (2022), there are different approaches to interpretability in learning. One important distinction is between local explanation, where we explain the prediction of the model on a single data point, and global interpretation, where we explain the model itself. The content of this chapter will focus on the latter. A common approach to global interpretation is to use simpler "interpretable" models (e.g., decision trees) to approximate more complex ones (Craven and Shavlik, 1995). This is known as *post-hoc interpretability* (Molnar, 2022). For example, Zhang, Yang, Ma, and Wu (2019) used decision trees to interpret convolutional neural networks. Formally, interpretability can be modeled as a property of a classifier. For example, Dziugaite, Ben-David, and Roy (2020) define a variant of empirical risk minimization (ERM), where each classifier in a given class $\mathcal{H}$ is either interpretable or not, and the task is to learn an interpretable one even though the target concept is not necessarily interpretable. We generalize this setup by assigning a complexity measure to each classifier, e.g., the depth for decision trees. This allows to trade-off the desired accuracy ε and the maximum depth of a decision tree one is willing to call interpretable. Learning-theoretic perspectives on interpretability are rare and typically not covered in standard books and surveys. One important line of work initiated by Dasgupta et al. (2020) deals with the problem of approximating a given k-means or k-median clustering with decision trees. From this perspective, our setup can be seen as a generalization from clusterings to arbitrary concepts. However, that line of work focuses on efficient algorithms to compute decision trees with k leaves and approximation guarantees in terms of the k-means or k-medians cost function, not in terms of classification error under a distribution as we do here. Bastani, Kim, and Bastani (2017) discuss a related problem setup where a given classifier is approximated using a decision tree. Under strong assumptions, the authors state convergence results for the proposed decision tree. However, they do not state bounds on the required depth which is assumed to be given as a hyperparameter. Some algorithmic analyses exist for specific cases of hypothesis spaces and standard explainers. For example, Garreau and Luxburg (2020) analyse LIME (Ribeiro, Singh, and Guestrin, 2016), one of the most used explanation techniques. Li, Nagarajan, Plumb, and Talwalkar (2021) discuss generalization bounds for local explainers. Blanc, Lange, and Tan (2021) introduce a local variant of our setup with the goal of explaining the classification $f(x)$ of a single instance x using a conjunction with small size (i.e., a small decision list). Their results cannot be used for our goal of global interpretation as one would have to take the union of all the local conjunctions for all (potentially infinite) instances x. Closer to our setup, Moshkovitz, Yang, and Chaudhuri (2021) state bounds on the depth of a decision tree required to fit a linear classifier with margin. Similarly to us, they also strongly rely on boosting arguments. Vidal and Schiffer (2020) give upper bounds on the number of nodes of a single decision tree to approximate an ensemble of trees. While mainly focusing on local explainability, Blanc et al. (2021) also state bounds on the depth of a decision tree required to fit an arbitrary classifier $f\colon \{0,1\}^d \to \{0,1\}$ under the uniform distribution on $\{0,1\}^d$. They do so by relying on classical bounds on the depth in terms of certificate complexity (Smyth, 2002, Tardos, 1989). As we focus instead on general hypothesis classes and distributions, their results are not directly comparable to ours.

9.3 Preliminaries and Definitions

Let $\mathcal{X}$ be any domain. We denote by P a distribution[*] on $\mathcal{X}$ and by $\mathcal{P}(\mathcal{X})$ the set of all distributions on $\mathcal{X}$, by $\mathcal{H} \subseteq 2^{\mathcal{X}}$ a hypothesis class on $\mathcal{X}$, and by $\mathrm{VC}(\mathcal{X},\mathcal{H})$ its VC dimension specifying the domain $\mathcal{X}$. We denote by $\mathrm{Alg}(\mathcal{H})$ the algebra generated by $\mathcal{H}$, i.e., the smallest set system $\mathcal{A} \subseteq 2^{\mathcal{X}}$ closed under complements and finite unions such that $\mathcal{H} \subseteq \mathcal{A}$ and $\emptyset, \mathcal{X} \in \mathcal{A}$. The σ-algebra $\sigma(\mathcal{H})$ is the smallest algebra containing $\mathcal{H}$ that is closed under countable unions. We denote by $c \in 2^{\mathcal{X}}$ an arbitrary concept (not necessarily in $\mathcal{H}$). As usual we also view c as a binary classification function $c\colon \mathcal{X} \to \{0,1\}$. Our goal is to understand how well c can be approximated using aggregations of hypotheses in $\mathcal{H}$.

[*] By default we assume a fixed but otherwise arbitrary σ-algebra on $\mathcal{X}$ and that all functions/sets discussed in our theorems are measurable. We also borrow standard assumptions on the underlying σ-algebra which allow us to use the VC Theorem (Vapnik and Chervonenkis, 1971). See, e.g., Blumer et al. (1989).

A decision tree over $\mathcal{X}$ is a full finite binary tree T with nodes $\mathcal{V}(T)$, where every leaf $z \in \mathcal{L}(T)$ holds a label $\lambda_z \in \{0,1\}$ and every internal node $v \in \mathcal{V}(T) \setminus \mathcal{L}(T)$ holds a *splitting criterion* (or decision stump) $f_v \colon \mathcal{X} \to \{0,1\}$. The depth (or height) of T is denoted as $\operatorname{depth}(T)$. We say T is *$\mathcal{H}$-based* if $f_v \in \mathcal{H}$ for all $v \in \mathcal{V}(T)$, and we denote by $\mathcal{T}_{\mathcal{H}}$ the set of all $\mathcal{H}$-based decision trees. We also use T to denote the binary classifier $T \colon \mathcal{X} \to \{0,1\}$ induced by T in the standard way. Note that $\mathcal{T}_{\mathcal{H}} \equiv \operatorname{Alg}(\mathcal{H})$, as any $\mathcal{H}$-based tree T can be rewritten as a Boolean formula and vice versa. For every $d \in \mathbb{N}^+$ we let $\operatorname{Alg}_d(\mathcal{H}) \coloneqq \{T \in \operatorname{Alg}(\mathcal{H}) : \operatorname{depth}(T) \leq d\}$. Given $P \in \mathcal{P}(\mathcal{X})$ and a concept $c \in 2^{\mathcal{X}}$, the *loss* of T with respect to c under P is $L_P(T,c) \coloneqq \mathbb{P}_{x \sim P}(T(x) \neq c(x)) = P\big(T^{-1}(1) \triangle c\big)$, where $A \triangle B \coloneqq (A \setminus B) \cup (B \setminus A)$ is the symmetric difference between A and B.

The central object of interest within the current chapter is represented by accurate approximations for a given concept c which, according to the main focus of this chapter, consist of decision trees that utilize hypothesis from the given class $\mathcal{H}$ as splitting criteria. To be precise, an *ε-accurate $\mathcal{H}$-approximation* of c under P is an $\mathcal{H}$-based decision tree T with $L_P(T,c) \leq \varepsilon$. The set of all such trees is denoted as $\mathcal{T}^c_{\mathcal{H}}(\varepsilon \mid P)$,[†] and their minimal depth is

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \coloneqq \inf_{T \in \mathcal{T}^c_{\mathcal{H}}(\varepsilon \mid P)} \operatorname{depth}(T) \; . \tag{9.1}$$

9.4 Approximability and Interpretability

This section introduces the key definitions used in our results. We start with the definition of approximability.

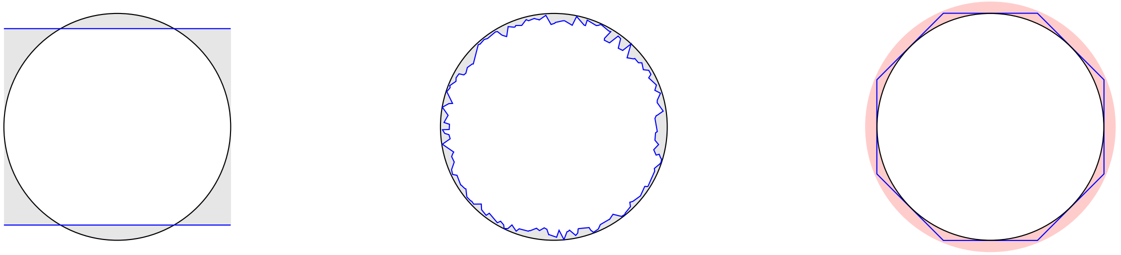
Definition 9.1 (Approximability). *A concept c is approximable by $\mathcal{H}$ if $\mathcal{T}^c_{\mathcal{H}}(\varepsilon \mid P) \neq \emptyset$ for every distribution $P \in \mathcal{P}(\mathcal{X})$ and every $\varepsilon > 0$.*

Approximability is our weakest notion, as it only requires that for any desired accuracy value a tree approximating c exists under any distribution, without any constraint on its depth. In fact, there may not even exist a function f such that $\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P)$ is bounded by $f(\varepsilon)$ for all distributions P.

For example, for $\mathcal{X} = \mathbb{R}^d$ let c be the unit d-dimensional Euclidean ball centered at the origin and $\mathcal{H}$ be the family of affine halfspaces whose boundary is orthogonal to, say, the d-th dimension. Then, any finite aggregation T of such halfspaces is unable to discern points that are aligned along the i-th dimension for any $i \neq d$ and, thus, necessarily incurs a constant $L_P(T,c)$ for some distribution P. On the other hand, if we extend $\mathcal{H}$ to be the family of all halfspaces in $\mathcal{X} = \mathbb{R}^d$, then it is possible to show that we can approximate the unit ball c up to any accuracy under any distribution. Indeed, it is known that a variant of the 1-nearest neighbour (1-NN) algorithm is universally strongly Bayes consistent in essentially separable metric spaces (Hanneke, Kontorovich, Sabato, and Weiss, 2021), and any 1-NN classifier corresponds to a finite Voronoi partition which can be represented as an $\mathcal{H}$-based decision tree. However, we expect the number of Voronoi cells, and thus the depth of the $\mathcal{H}$-based decision tree representing it, to grow larger as the distribution P concentrates around the decision boundary (that is, the surface of the unit ball c). Consider, for instance, a family of distributions P_α with $\alpha \in (0,1)$, where each P_α has support corresponding to the spherical shell $B^d(1+\alpha) \setminus B^d(1-\alpha)$ with inner radius $1-\alpha$ and outer radius $1+\alpha$ (here we denote by $B^d(r) \coloneqq \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ the origin-centered Euclidean ball of radius $r > 0$ in $\mathbb{R}^d$). Then,

[†] This is also known as the ε-Rashomon set (Fisher, Rudin, and Dominici, 2019).

we expect the number of Voronoi cells defining the decision boundary of the 1-NN classifiers that guarantee loss at most $0 < \varepsilon \leq 1$ to grow as $\alpha \to 0^+$. Figures 9.1a and 9.1b illustrate these examples in $\mathbb{R}^2$.



(a) Non-approximable (b) Approximable but not interpretable (c) Uniformly interpretable

Figure 9.1: Approximating a disk with halfspaces: the approximation error is the grey-shaded area, while the pink area is the margin region. In (a), we show inapproximability with x-axis-aligned halfspaces. In (b), we show the disk is approximable (but not interpretable) with arbitrary halfspaces, via a Voronoi tessellation with one-sided error. In (c), we show the disk with margin is uniformly interpretable with halfspaces.

Next, we define our notion of interpretability. Recall that we view an interpretation as an approximation via a tree of small depth. We formalize "small" by requiring the existence of a function that bounds the tree depth in terms of its accuracy.

Definition 9.2 (Interpretability)**.** *A concept c is* interpretable *by $\mathcal{H}$ if there is a function $f\colon (0,1] \to \mathbb{N}$ such that, for every distribution $P \in \mathcal{P}(\mathcal{X})$, there exists $\varepsilon_P > 0$ for which*

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq f(\varepsilon) \qquad \textit{for all} \quad 0 < \varepsilon \leq \varepsilon_P\,.$$

If this is the case, then we say that c is interpretable by $\mathcal{H}$ at depth rate f.
A concept c is uniformly interpretable *by $\mathcal{H}$ if there is a function $f'\colon (0,1] \to \mathbb{N}$ such that*

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq f'(\varepsilon) \qquad \textit{for all} \quad P \in \mathcal{P}(\mathcal{X}) \quad \textit{and} \quad 0 < \varepsilon \leq 1\,.$$

If this is the case, we say that c is uniformly interpretable by $\mathcal{H}$ at depth rate f'.

Note that interpretability requires the bound on the depth to hold only for values of ε that are smaller than a certain threshold ε_P which may depend on the distribution P. Uniform interpretability, instead, requires the depth bound to hold for all ε irrespective of the distribution.

Recalling the above example with the Euclidean space $\mathcal{X} = \mathbb{R}^d$ as domain and the family of halfspaces as the hypothesis class $\mathcal{H}$, if the concept c corresponds to the unit Euclidean ball with margin $\mu > 0$ then c is uniformly interpretable. More formally, such a concept c can be modeled as a partial function $c\colon \mathcal{X} \to \{0,1\}$ with natural domain $\widetilde{\mathcal{X}} \subset \mathcal{X}$, where points in the margin belong to $\mathcal{X} \setminus \widetilde{\mathcal{X}} = B^d(1+\mu) \setminus B^d(1)$ and $c^{-1}(1) = B^d(1)$. Then, without loss of generality, the same definitions and results apply as if the concept c was a total function by restricting the domain to $\widetilde{\mathcal{X}}$ and, for every distribution $P \in \mathcal{P}(\mathcal{X})$, considering the distribution $\widetilde{P}(\cdot) := P(\cdot \mid \widetilde{\mathcal{X}})$ instead. This follows from the fact that we incur no mistakes for any labeling of points that do not belong to the domain of the "partial" concept c, and the loss of any $\mathcal{H}$-approximation T of c is thus given by $L_{\widetilde{P}}(T,c)$. By reusing geometric results on the approximation of convex bodies, there exists a polytope Q such that $B^d(1) \subseteq Q \subseteq B^d(1+\mu)$, whose (finite) number of vertices is bounded from above by a function of d and μ (Naszódi, 2019). The polytope Q thus separates the positively labeled points $B^d(1)$ from the negatively labeled ones—achieving loss 0 under any distribution with support $\widetilde{\mathcal{X}}$—and it is equivalently representable as an $\mathcal{H}$-based decision tree with depth bounded by a function of d and μ only (i.e., the intersection of halfspaces associated to the facets of Q). See Figure 9.1c for an illustration of this example in $\mathbb{R}^2$.

At first glance, our notions of interpretability may appear a little narrow. Suppose that, for every distribution P, a concept c is interpretable by $\mathcal{H}$ at polynomial depth rate, but the degree can grow unbounded with P. In other words, for every $d \in \mathbb{N}^+$ there exists P_d such that c is interpretable by $\mathcal{H}$ at polynomial depth rate with degree d, but not at polynomial depth rate with any smaller degree $d' < d$. Then c is not interpretable by $\mathcal{H}$ according to our definition, but we could still say that c is interpretable at *polynomial* depth rate. More formally, we could consider the family $\mathcal{F}$ of all polynomials, and require that for every P there is some $f \in \mathcal{F}$ that bounds $\mathrm{depth}^c_{\mathcal{H}}(\cdot \mid P)$. By varying $\mathcal{F}$, we obtain a vast range of interpretability rates: logarithmic, sublinear, linear, polynomial, exponential, and so on. Surprisingly, our results show that this hierarchy collapses: an approximable concept c is either not interpretable at all, or is uniformly interpretable at logarithmic rate.

9.5 A Trichotomy for Interpretability

This section states our main result: as soon as a concept is interpretable at *some* rate, then it is uniformly interpretable at a constant rate for VC classes, and at a logarithmic rate in general.

Theorem 9.1 (Interpretability trichotomy)**.** *Let $\mathcal{X}$ be any domain. For every concept c and every VC hypothesis class $\mathcal{H}$ over $\mathcal{X}$ exactly one of the following cases holds:*

(1) c is not approximable by $\mathcal{H}$;

(2) c is approximable but not interpretable by $\mathcal{H}$;

(3) c is uniformly interpretable by $\mathcal{H}$ at constant depth rate.

If $\mathrm{VC}(\mathcal{X}, \mathcal{H}) = \infty$ *then all claims above hold true, but with (3) replaced by:*

(3′) c is uniformly interpretable by $\mathcal{H}$ at depth rate at most logarithmic.

Moreover, all cases are nonempty.

We emphasize again that Theorem 9.1 is in stark contrast with the behavior of excess risk in terms of training set size observed in statistical learning, where, in the non-uniform (or universal) setting, both exponential and linear rates are possible. It should also be noted that we do not know if case (3′) collapses into case (3)—that is, if a constant depth rate holds also for non-VC classes—or if a non-constant rate is in general unavoidable—e.g., because of the significantly stronger representative power of non-VC classes. This is one of the interesting questions left open.

We further observe that, while point (3) of Theorem 9.1 shows that c is uniformly interpretable by $\mathcal{H}$ at a constant depth rate, this does not necessarily imply the existence of a single $\mathcal{H}$-based decision tree providing such a guarantee for all values of $\varepsilon > 0$. For example, consider a domain $\mathcal{X} = \mathbb{N}$, a concept $c = \{0\}$, and a hypothesis class $\mathcal{H} = \{\{1, \dots, n\} : n \in \mathbb{N}^+\}$. Now let P be the distribution such that $P(0) = 0.5$ and $P(x) = 2^{-(x+1)}$ for all $x \in \mathbb{N}^+$. For any $\varepsilon > 0$ the depth-1 decision tree with splitting criterion $h = \{1, \ldots, \lceil \log_2(1/\varepsilon) \rceil\}$ is an ε-accurate approximation of c under P, but no $\mathcal{H}$-based tree is an ε-approximation of c for all ε simultaneously.

Our proof of Theorem 9.1 combines a variety of techniques from different contexts. The first step involves identifying a criterion which can be thought of as a form of "weak interpretability" (items (a) and (b) in the proof). The rest of the proof demonstrates that if a concept c fails to satisfy this criterion, then it is not interpretable by $\mathcal{H}$, and if it does, then it is uniformly interpretable by $\mathcal{H}$. The former impossibility result entails establishing a lower bound on the interpretation rate for an arbitrarily small accuracy with respect to a *fixed* and carefully tailored distribution. This type of lower bounds are more intricate than distribution-free lower bounds (such as those outlined in the No-Free-Lunch Theorem in the PAC setting) and were studied, e.g., by Antos and Lugosi (1998), Bousquet, Hanneke, Moran, Shafer, and Tolstikhin (2023). In the complementary case, when c satisfies the weak interpretability criterion with respect to $\mathcal{H}$, we prove that c is in fact uniformly interpretable by $\mathcal{H}$ with logarithmic depth, and if $\mathcal{H}$ has a finite VC dimension, then c is uniformly interpretable with constant depth. The logarithmic construction and its analysis builds on ideas and techniques originating from boosting algorithms for decision trees (Kearns and Mansour, 1999, Takimoto and Maruoka, 2003). The derivation of constant depth approximation when $\mathcal{H}$ is a VC class relies on a uniform convergence argument (Vapnik and Chervonenkis, 1971) combined with the Minimax Theorem (Von Neumann, 1928). This derivation is also linked to boosting theory and resembles the boosting-based sample compression scheme by Moran and Yehudayoff (2016).

Proof of Theorem 9.1. We start by proving the cases (1)-(3). Suppose (1) fails, so $\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) < \infty$ for all $P \in \mathcal{P}(\mathcal{X})$ and all $\varepsilon > 0$. This implies that, for any fixed $\gamma \in (0, \frac{1}{2})$, exactly one of the following two cases holds:

(a) for every $d \in \mathbb{N}$ there exists a distribution P_d such that $\operatorname{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_d\big) > d$;

(b) there exists $d \in \mathbb{N}$ such that $\operatorname{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq d$ for all distributions P.

Suppose (a) holds; we show this implies case (2) of the trichotomy. To this end, we prove that there is no function $r\colon (0,1] \to \mathbb{N}$ such that c is interpretable by $\mathcal{H}$ at depth rate r. Choose indeed any such r. For every $n \in \mathbb{N}^+$, let $d_n := r\big(2^{-n}(\frac{1}{2} - \gamma)\big)$, and consider the following distribution over $\mathcal{X}$:

$$P^* := \sum_{i \in \mathbb{N}^+} 2^{-i} \cdot P_{d_i} \ . \tag{9.2}$$

Since P_{d_n} appears in P^* with coefficient 2^{-n}, this implies that, for $\varepsilon_n := 2^{-n}(\frac{1}{2} - \gamma)$, any ε_n-accurate $\mathcal{H}$-interpretation of c under P^* is $(\frac{1}{2} - \gamma)$-accurate under P_{d_n} and so has depth larger than $d_n := r(\varepsilon_n)$. Indeed, for any $n \in \mathbb{N}^+$, any tree $T \in \mathcal{T}^c_{\mathcal{H}}(\varepsilon_n \mid P^*)$ satisfies that

$$\varepsilon_n \geq L_{P^*}(T, c) = \sum_{i \in \mathbb{N}^+} 2^{-i} L_{P_{d_i}}(T, c) \geq 2^{-n} L_{P_{d_n}}(T, c) \ ,$$

which means that $L_{P_{d_n}}(T, c) \leq 2^n \varepsilon_n = \frac{1}{2} - \gamma$ and thus $\mathcal{T}^c_{\mathcal{H}}(\varepsilon_n \mid P^*) \subseteq \mathcal{T}^c_{\mathcal{H}}(\frac{1}{2} - \gamma \mid P_{d_n})$. Consequently, we have that

$$\operatorname{depth}^c_{\mathcal{H}}(\varepsilon_n \mid P^*) \geq \operatorname{depth}^c_{\mathcal{H}}\left(\frac{1}{2} - \gamma \mid P_{d_n}\right) > d_n = r(\varepsilon_n) \tag{9.3}$$

holds for all $n \in \mathbb{N}^+$. We conclude that c is not interpretable by $\mathcal{H}$ at depth rate r, as desired.

Now suppose (b) holds; we show this implies case (3) of the trichotomy. Let $\mathcal{T}$ be the set of all binary classifiers that are represented by $\mathcal{H}$-based decision trees of depth at most d, where $d \in \mathbb{N}$ satisfies $\operatorname{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \le d$ for all $P \in \mathcal{P}(\mathcal{X})$. It is known that $\mathrm{VC}(\mathcal{X}, \mathcal{H}) < \infty$ implies $\mathrm{VC}(\mathcal{X}, \mathcal{T}) < \infty$ (Dudley, 1978).

We will first prove the claim by taking as domain an arbitrary but finite subset $U \subseteq \mathcal{X}$. Later on we will choose U appropriately as a function of the distribution $P \in \mathcal{P}(\mathcal{X})$, and this will prove the theorem's claim. Fix then any such U, and let $\mathcal{P}(U)$ be the family of all distributions over U. By definition of d,

$$\sup_{P \in \mathcal{P}(U)} \inf_{T \in \mathcal{T}} L_P(T, c) \le \frac{1}{2} - \gamma \,. \tag{9.4}$$

By Von Neumann's minimax theorem, recalling that the value of the game does not change if the column player uses a pure strategy, we have that

$$\sup_{P \in \mathcal{P}(U)} \inf_{T \in \mathcal{T}} L_P(T, c) = \inf_{D \in \mathcal{P}(\mathcal{T})} \sup_{x \in U} \mathop{\mathbb{E}}_{T \sim D} \left[L_{\delta_x}(T, c)\right] \,, \tag{9.5}$$

where $\mathcal{P}(\mathcal{T})$ is the set of all distributions over $\mathcal{T}$, δ_x is the Dirac delta at $x \in U$, and $\mathbb{E}_{T \sim D}\left[L_{\delta_x}(T, c)\right]$ is thus the expected loss on x of a tree T drawn from D. Hence, there exists $D^* \in \mathcal{P}(\mathcal{T})$ for which

$$\mathop{\mathbb{E}}_{T \sim D^*} \left[L_{\delta_x}(T, c)\right] \le \frac{1}{2} - \gamma \qquad \forall x \in U \,, \tag{9.6}$$

and therefore, since $c(x), T(x) \in \{0, 1\}$ for all x and T,

$$\left| c(x) - \mathop{\mathbb{P}}_{T \sim D^*}(T(x) = 1) \right| = \mathop{\mathbb{P}}_{T \sim D^*}\big(T(x) \neq c(x)\big) \le \frac{1}{2} - \gamma \qquad \forall x \in U \,. \tag{9.7}$$

Let $(\mathcal{T}, U)$ be the dual set system of $(U, \mathcal{T})$. Note that the dual VC dimension $\mathrm{VC}(\mathcal{T}, U)$ satisfies

$$\mathrm{VC}(\mathcal{T}, U) \le \mathrm{VC}(\mathcal{T}, \mathcal{X}) < 2^{\mathrm{VC}(\mathcal{X}, \mathcal{T}) + 1} < \infty \,, \tag{9.8}$$

where the second inequality shows a known relation (Assouad, 1983) between the primal VC dimension $\mathrm{VC}(\mathcal{X}, \mathcal{T})$ of $(\mathcal{X}, \mathcal{T})$ and its dual VC dimension $\mathrm{VC}(\mathcal{T}, \mathcal{X})$. By the classic uniform convergence result of Vapnik and Chervonenkis (1971) and the probabilistic method, there exists a multiset $R \subseteq \mathcal{T}$ with $|R| \le r := r(\mathrm{VC}(\mathcal{X}, \mathcal{T}), \gamma, d)$ such that, for every $x \in U$,

$$\left| \frac{|\{T \in R : T(x) = 1\}|}{|R|} - \mathop{\mathbb{P}}_{T \sim D^*}(T(x) = 1) \right| < \frac{\gamma}{2} \,. \tag{9.9}$$

Together with Equations (9.7) and (9.9), this yields

$$\left| \frac{|\{T \in R : T(x) = 1\}|}{|R|} - c(x) \right| < \frac{1}{2} - \frac{\gamma}{2} \tag{9.10}$$

by the triangle inequality.[‡] We now build a $\mathcal{H}$-based decision tree T^*_U that computes the majority

[‡] Note that, if no D^* achieves the infimum of the r.h.s. of Equation (9.5), the same result holds with, say, $(1 - \gamma)/2$ as the r.h.s. of Equation (9.6) because it suffices to show that the l.h.s. of Equation (9.10) is strictly less than 1/2 for our purposes.

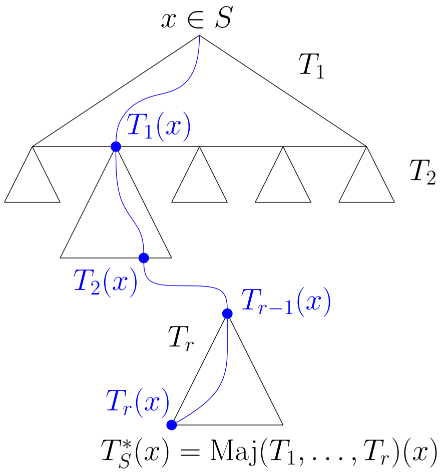


Figure 9.2: Construction of the majority-vote tree T^*_S for some finite $S \subseteq \mathcal{X}$ given the collection $T_1, \dots, T_r$, where $\mathrm{Maj}(T_1, \dots, T_r)(x)$ denotes the majority vote prediction computed for $x \in S$.

vote over all $T \in R$. This tree can be constructed as shown in Figure 9.2 and described as follows. Let $T_1, \dots, T_{|R|}$ be the trees in R. Replace each leaf of T_1 with a copy of T_2; in the resulting tree replace every leaf with a copy of T_3, and so on until obtaining T^*_U. For each leaf $z \in \mathcal{L}(T^*_U)$ of T^*_U, define its label λ_z as the majority vote given by leaves of (the copies of) $T_1, \dots, T_{|R|}$ that are encountered on the path from the root of T^*_U to z. Note that T^*_U has depth bounded by rd and, by Equation (9.10), computes $c(x)$ for all $x \in U$. Thus, $L_U(T^*_U, c) = 0$ where L_U is the expected loss over the uniform distribution over U.

We now choose the set U appropriately. Let $\mathcal{T}^*$ be the family of all $\mathcal{H}$-based decision trees whose depth is at most rd. Because, once again, $\mathrm{VC}(\mathcal{X}, \mathcal{T}^*) < \infty$, by uniform convergence there is a finite multiset $U \subseteq \mathcal{X}$ such that, for all $T \in \mathcal{T}^*$, $|L_P(T, c) - L_U(T, c)| \le \varepsilon$. Since $T^*_U \in \mathcal{T}^*$ and $L_U(T^*_U, c) = 0$, it follows that $L_P(T^*_U, c) \le \varepsilon$. This completes the proof of case (3). Case (3′) follows from Theorem 9.2 below, assuming (b) holds.

It remains to prove that all cases are nonempty. For (1) let $\mathcal{X} := \{a, b\}$, $\mathcal{H} := \{\mathcal{X}\}$, $c := \{a\}$, and note that under the uniform distribution no $\mathcal{H}$-interpretation of c is ε-accurate for $\varepsilon < \frac{1}{2}$. For (3) consider any $\mathcal{X}, \mathcal{H}$ with $\mathcal{H} \neq \emptyset$ and choose any $c \in \mathcal{H}$; this holds for (3′) too if $\mathcal{H}$ is not a VC class. For (2) we show $c, \mathcal{H}$ that satisfy case (a) above. Let $\mathcal{X} := \mathbb{N}$, $c := \mathbb{N}^+$, and $\mathcal{H} := \{\{i\} : i \in \mathbb{N}^+\}$. For every $n \in \mathbb{N}^+$ consider the distribution P_n with support $\{0, \dots, n\}$ such that $P_n(0) := \frac{1}{2}$ and that $P_n(i) := \frac{1}{2n}$ for every $i \in \{1, \dots, n\}$. To conclude note that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_n\big)$ is unbounded as a function of n for any constant $\gamma \in (0, \frac{1}{2})$. □

The proof of case (3′) of Theorem 9.1 uses the following result. Its proof can be found in Appendix E.1, and is an adaptation of the results by Kearns and Mansour (1999) and Takimoto and Maruoka (2003) on boosting decision trees. The main difference is that, via an adequate modification of the `TopDown` algorithm (Kearns and Mansour, 1999), we bound the depth rather than the size of the boosted decision tree.

Theorem 9.2. *Let $\mathcal{X}$ be any domain. For any concept c and any hypothesis class $\mathcal{H}$ over $\mathcal{X}$, if there exist $\gamma \in (0, \frac{1}{2})$ and $d \in \mathbb{N}$ such that $\mathrm{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \le d$ for all $P \in \mathcal{P}(\mathcal{X})$, then $\mathrm{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \le \frac{d}{2\gamma^2} \log \frac{1}{2\varepsilon}$ for all $P \in \mathcal{P}(\mathcal{X})$ and all $\varepsilon > 0$.*

9.6 Algebraic Characterizations

In this section we show that the notions of approximability and interpretability admit set-theoretical and measure-theoretical characterizations based on properties of $\mathcal{H}$ and the algebras it generates.

To begin with, we need a notion of closure of $\mathcal{H}$. Loosely speaking, we want to include all concepts that, under every distribution, can be approximated arbitrarily well by single elements of $\mathcal{H}$. In other words, these are the concepts that are approximable by $\mathcal{H}$ using decision trees of depth 1.

Definition 9.3. *The* closure *of* $\mathcal{H} \subseteq 2^{\mathcal{X}}$ *is*

$$\operatorname{clos}(\mathcal{H}) := \Big\{ h \subseteq \mathcal{X} \,\Big|\, \forall P \in \mathcal{P}(\mathcal{X}),\ \exists h_1, h_2, \ldots \in \mathcal{H}\, s.t. \lim_{n\to\infty} P(h \triangle h_n) = 0 \Big\}. \tag{9.11}$$

Observe that $\operatorname{clos}(\mathcal{H}) \supseteq \mathcal{H}$ by definition. To illustrate the closure let us discuss the hypothesis class $\mathcal{H}$ of rational halfspaces in $\mathbb{R}^2$, i.e., sets of the form $\big\{\{(x,y) : ax + by + d \geq 0\} : a, b, d \in \mathbb{Q}\big\}$. Every concept $c\colon \mathbb{R}^2 \to \{0,1\}$ is approximable by $\mathcal{H}$, as before, relying on the 1-NN algorithm. Halfspaces with real coefficients such as $\{(x,y) : x + y \geq \sqrt{2}\}$ are not in $\mathcal{H}$ but are interpretable by $\mathcal{H}$ with depth 1. In general, the closure is related to the concept of universally measurable sets.

We start with the following lemma, which is derived from well-known results in measure theory (see the proof in Appendix E.2).

Lemma 9.1. *Let $\mathcal{X}$ be any domain and $\mathcal{H} \subseteq 2^{\mathcal{X}}$. Then,* $\operatorname{clos}(\sigma(\mathcal{H})) = \operatorname{clos}(\operatorname{Alg}(\mathcal{H}))$.

We now state the algebraic characterization of the concepts that are approximable by a given hypothesis class $\mathcal{H}$ on some domain $\mathcal{X}$.

Theorem 9.3 (Algebraic characterization of approximability)**.** *Let $\mathcal{X}$ be any domain and $\mathcal{H}$ any hypothesis class over $\mathcal{X}$. A concept $c \subseteq \mathcal{X}$ is approximable by $\mathcal{H}$ if and only if $c \in \operatorname{clos}(\sigma(\mathcal{H}))$.*

Proof. Suppose c is universally approximable by $\mathcal{H}$. Let $P \in \mathcal{P}(\mathcal{X})$ be any distribution. Then, for every $\varepsilon > 0$ there exists an ε-accurate $\mathcal{H}$-approximation $T \in \operatorname{Alg}(\mathcal{H})$ of c under P. Then $P(T \triangle c) = L_P(T, c) \leq \varepsilon$. Consider now the sequence $T_1, T_2, \ldots \in \operatorname{Alg}(\mathcal{H})$ such that, for each $n \in \mathbb{N}^+$, T_n is an ε_n-accurate $\mathcal{H}$-approximation of c under P with the choice $\varepsilon_n := 2^{-n}$. The sequence $(T_n)_{n\in\mathbb{N}^+}$ is such that $\lim_{n\to\infty} P(T_n \triangle c) \leq \lim_{n\to\infty} 2^{-n} = 0$, and thus $c \in \operatorname{clos}(\operatorname{Alg}(\mathcal{H})) = \operatorname{clos}(\sigma(\mathcal{H}))$, where the latter equality follows by Lemma 9.1.

Now suppose $c \in \operatorname{clos}(\sigma(\mathcal{H})) = \operatorname{clos}(\operatorname{Alg}(\mathcal{H}))$. Fix a distribution $P \in \mathcal{P}(\mathcal{X})$ and $\varepsilon > 0$. By definition of closure, and because $\operatorname{Alg}(\mathcal{H}) \equiv \mathcal{T}_{\mathcal{H}}$, there exists a sequence $T_1, T_2, \ldots \in \operatorname{Alg}(\mathcal{H})$ of trees such that $\lim_{n\to\infty} P(T_n \triangle c) = 0$, and thus there exists some $i \in \mathbb{N}^+$ such that $P(T_i \triangle c) \leq \varepsilon$. This implies that T_i is an ε-accurate $\mathcal{H}$-approximation of c under P with finite depth. As this holds for every P and every $\varepsilon > 0$, it follows that c is universally approximable by $\mathcal{H}$. □

Furthermore, we manage to prove an algebraic characterization for the concepts that are uniformly interpretable, given a VC class $\mathcal{H}$ on some domain $\mathcal{X}$.

Theorem 9.4 (Characterization of uniform interpretability for VC classes)**.** *Let $\mathcal{X}$ be any domain and let $\mathcal{H}$ be a VC hypothesis class over $\mathcal{X}$. A concept c is uniformly interpretable (at a constant depth) if and only if $c \in \bigcup_{d=1}^{\infty} \operatorname{clos}(\operatorname{Alg}_d(\mathcal{H}))$.*

Proof. Since $\mathrm{VC}(\mathcal{X}, \mathcal{H}) < \infty$, item (3) of Theorem 9.1 implies that there exists $d \in \mathbb{N}$ such that, for all $P \in \mathcal{P}(\mathcal{X})$ and all $\varepsilon > 0$, $\mathrm{depth}^{c}_{\mathcal{H}}(\varepsilon \mid P) \leq d$. Using an argument similar to the one used in the proof of Theorem 9.3, we then conclude that $c \in \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. Hence, the set of concepts that are uniformly interpretable by $\mathcal{H}$ is precisely $\bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. □

If the domain $\mathcal{X}$ is countable, then closure reduces to pointwise convergence and our algebraic characterization becomes simpler. This is formalized in the next theorem, whose proof relies on some technical lemmas and can be found in Appendix E.2. Namely, we show that $\mathrm{clos}(\sigma(\mathcal{H})) = \sigma(\mathcal{H})$ and $\bigcup_{d=1}^{\infty} \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H})) = \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.

Theorem 9.5. (Characterization for VC classes and countable domains) *Let $\mathcal{X}$ be any countable domain, let c be any concept, and let $\mathcal{H}$ be a VC hypothesis class over $\mathcal{X}$. Then:*

1. *c is approximable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H})$.*
2. *c is approximable but not interpretable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H}) \setminus \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*
3. *c is uniformly interpretable by $\mathcal{H}$ if and only if $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*

One way to interpret the above algebraic characterization of approximable and uniformly interpretable concepts consists of what follows, whenever $\mathcal{X}$ is countable and $\mathcal{H}$ is VC. A concept c approximable by $\mathcal{H}$ equivalently belongs to $\mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$, which alternatively means that there exists a sequence of finite $\mathcal{H}$-based trees that converges to c (or, alternatively, the trees in the sequence approximate increasingly well c with loss arbitrarily close to zero). On the other hand, the class of concepts that are uniformly interpretable by $\mathcal{H}$ corresponds to $\mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$, meaning that any such concept c is essentially equivalent to a *single* finite tree whose splitting functions can be individually approximated arbitrarily well by single elements of $\mathcal{H}$.

9.7 General Representations

Although shallow decision trees are the blueprint of interpretable models, our theory naturally extends to ways of measuring the complexity of elements in $\mathrm{Alg}(\mathcal{H})$ different from the tree depth. Next, we define a set of minimal conditions (satisfied, e.g., by tree depth) that a function must satisfy to be used as a complexity measure for $\mathrm{Alg}(\mathcal{H})$.

Definition 9.4. *Let $\mathcal{X}$ be any domain and $\mathcal{H}$ a hypothesis class over $\mathcal{X}$. A function $\Gamma \colon \mathrm{Alg}(\mathcal{H}) \to \mathbb{N}$ is a* graded complexity measure *if:*

1. *$\Gamma(f) = 0$ for all $f \in \mathcal{H}$,*
2. *$\Gamma(f_1 \cup f_2) \leq 1 + \Gamma(f_1) + \Gamma(f_2)$ for all $f_1, f_2 \in \mathrm{Alg}(\mathcal{H})$,*
3. *$\Gamma(f_1 \cap f_2) \leq 1 + \Gamma(f_1) + \Gamma(f_2)$ for all $f_1, f_2 \in \mathrm{Alg}(\mathcal{H})$, and*
4. *$\Gamma(\mathcal{X} \setminus f) \leq 1 + \Gamma(f)$ for all $f \in \mathrm{Alg}(\mathcal{H})$.*

The minimal complexity of an ε-accurate $\mathcal{H}$-interpretation of c under P is

$$\Gamma^{c}_{\mathcal{H}}(\varepsilon \mid P) \coloneqq \inf_{T \in \mathrm{Alg}(\mathcal{H}) \colon L_P(T,c) \leq \varepsilon} \Gamma(T) \; . \tag{9.12}$$

The definitions of approximability, interpretability, and uniform interpretability are readily generalized to an arbitrary graded complexity measure, by simply replacing $\mathrm{depth}(\cdot)$ with $\Gamma(\cdot)$. We can then prove the following extension of Theorem 9.1.

Theorem 9.6 (Interpretability trichotomy for general representations)**.** *Let $\mathcal{X}$ be any domain and let Γ be any graded complexity measure. Then, for every concept c and every VC hypothesis class $\mathcal{H}$ over $\mathcal{X}$ exactly one of the following cases holds:*

(1) c is not approximable by $\mathcal{H}$;

(2) c is approximable by $\mathcal{H}$ but not interpretable by $\mathcal{H}$;

(3) c is uniformly interpretable by $\mathcal{H}$ at constant Γ-complexity rate.

If $\mathrm{VC}(\mathcal{X}, \mathcal{H}) = \infty$ then all claims above hold true, but with (3) replaced by:

(3′) c is uniformly interpretable by $\mathcal{H}$ at a Γ-complexity rate $\mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ for some $d \in \mathbb{N}$.

Unlike Theorem 9.1, cases (2) and (3) might collapse for certain choices of Γ even when $\mathcal{H}$ is not a VC class. Indeed, according to our definition, Γ is not forced to grow at any specific rate, and thus $\Gamma(f)$ might be bounded by some constant uniformly over $\mathrm{Alg}(\mathcal{H})$. In an extreme case one might in fact set $\Gamma \equiv 0$, although clearly this would not yield any interesting result.

Proof of Theorem 9.6. The proof is similar to the proof of Theorem 9.1. Suppose (1) fails, so $\Gamma^c_{\mathcal{H}}(\varepsilon \mid P) < \infty$ for all $\varepsilon > 0$ and all distributions P. This implies that, for any fixed $\gamma \in (0, \frac{1}{2})$, exactly one of the following two cases holds:

(a) for every $k \in \mathbb{N}$ there exists a distribution P_k such that $\Gamma^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P_k\big) > k$;

(b) there exists $k \in \mathbb{N}$ such that $\Gamma^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq k$ for all distributions P.

Suppose (a) holds; we show this implies case (2) of the trichotomy. Choose any function $r\colon (0, 1] \to \mathbb{N}$. For every $n \in \mathbb{N}^+$, let $d_n := r(2^{-n}(\frac{1}{2} - \gamma))$, and consider the following distribution over $\mathcal{X}$:

$$P^* := \sum_{n \in \mathbb{N}^+} 2^{-n} \cdot P_{d_n} \; . \tag{9.13}$$

Since P_{d_n} appears in P^* with coefficient 2^{-n}, this implies that, for $\varepsilon_n := 2^{-n}(\frac{1}{2} - \gamma)$, any ε_n-accurate interpretation of c under P^* is $(\frac{1}{2} - \gamma)$-accurate under P_{d_n}, and thus

$$\Gamma^c_{\mathcal{H}}(\varepsilon_n \mid P^*) \geq \Gamma^c_{\mathcal{H}}\bigg(\frac{1}{2} - \gamma \mid P_{d_n}\bigg) > d_n = r(\varepsilon_n) \; . \tag{9.14}$$

Hence, $\Gamma^c_{\mathcal{H}}(\varepsilon_n \mid P^*) > r(\varepsilon_n)$ for all $n \in \mathbb{N}^+$.

Suppose now (b) holds; we show this implies case (3) of the trichotomy. Define the family $\mathcal{A}_k := \big\{A \in \mathrm{Alg}(\mathcal{H}) \,:\, \Gamma(A) \leq k\big\}$. Fix any $P \in \mathcal{P}(\mathcal{X})$ and $\varepsilon > 0$. Following the same argument as in the proof of case (3) in Theorem 9.1, there exists an $\mathcal{A}_k$-based decision tree T such that $L_P(T, c) \leq \varepsilon$ and $\mathrm{depth}(T) \leq d$ for some $d \in \mathbb{N}$ independent of P and ε. Now we rewrite T as an element of $\mathrm{Alg}(\mathcal{H})$. Let $A_v \in \mathcal{A}_k$ be the decision stump T used at $v \in \mathcal{V}(T)$ and, denoting by $\mathcal{L}(T)$ the set of leaves of T, let $\lambda_z \in \{0, 1\}$ be the label of the leaf $z \in \mathcal{L}(T)$ in T. For every $v \in \mathcal{V}(T)$, define

$$A^T_v := \begin{cases} \mathcal{X} & v \in \mathcal{L}(T),\ \lambda_v = 1 \\ \emptyset & v \in \mathcal{L}(T),\ \lambda_v = 0 \\ (A_v \cap A^T_u) \cup (\overline{A}_v \cap A^T_w) & v \notin \mathcal{L}(T) \end{cases} \tag{9.15}$$

where u and w are, respectively, the left and right child of v when $v \notin \mathcal{L}(T)$. Let $A := A_r^T$ where r is the root of T. Observe that A is equivalent to T, and that $A \in \mathrm{Alg}(\mathcal{H})$. Moreover, $\Gamma(A_v^T) \leq 4 + 2\Gamma(A_v) + \Gamma(A_u^T) + \Gamma(A_w^T)$ by the properties of Γ (see Definition 9.4). Therefore,

$$\Gamma(A) = \mathcal{O}\left(\sum_{v \in \mathcal{V}(T)} (\Gamma(A_v) + 1) \right) = (k+1) \times \mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(|\mathcal{V}(T)|) \ , \tag{9.16}$$

where we used the fact that $\Gamma(A_v) \leq k$ because $A_v \in \mathcal{A}_k$. To conclude the proof, note that the above bound on depth(T) implies $\mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(2^{\mathrm{depth}(T)}) = \mathcal{O}(2^d)$, where both d and the constants in the $\mathcal{O}(\cdot)$ notation depend neither on P nor on ε.

As for case $(3')$, assume again (b) holds. Then, Theorem 9.2 applied to the class $\mathcal{A}_k$ implies the existence of an $\mathcal{A}_k$-based decision tree T such that $L_P(T, c) \leq \varepsilon$ and depth$(T) \leq d \log \frac{1}{2\varepsilon}$ for all P and $\varepsilon > 0$, where $d = \frac{1}{2\gamma^2}$. Constructing again $A \in \mathrm{Alg}(\mathcal{H})$ equivalent to T as above and using the bound on depth(T), we have $\Gamma(A) = \mathcal{O}(|\mathcal{V}(T)|) = \mathcal{O}(2^{\mathrm{depth}(T)}) = \mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ where both d and the constants in the $\mathcal{O}(\cdot)$ notation are independent of P and ε. □

We remark that the $\mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ bound on the complexity rate for case $(3')$ is due to the generality of Γ; in Appendix E.3, we show a more specific condition on Γ that recovers the $\mathcal{O}(\log(1/\varepsilon))$ rate.

Bibliography

Jacob D. Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

Zeyuan Allen-Zhu, Sebastien Bubeck, and Yuanzhi Li. Make the minority great again: First-order regret bound for contextual bandits. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 186–194. PMLR, 2018.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

Noga Alon, Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. From bandits to experts: A tale of domination and independence. *Advances in Neural Information Processing Systems*, 26, 2013.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35. PMLR, 2015.

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

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*, volume 9. Cambridge University Press, 1999.

András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. *Mach. Learn.*, 30(1): 31–56, 1998. doi: 10.1023/A:1007454427662.

András Antos, Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. *Theoretical Computer Science*, 473:77–99, 2013.

Raman Arora, Teodor Vanislavov Marinov, and Mehryar Mohri. Bandits with feedback graphs and switching costs. In *Advances in Neural Information Processing Systems*, pages 10397–10407, 2019.

Patrick Assouad. Densité et dimension. *Annales de l'Institut Fourier*, 33(3):233–282, 1983. doi: 10.5802/aif.938.

Idan Attias and Steve Hanneke. Adversarially robust PAC learnability of real-valued functions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 1172–1199, 2023.

Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k-fold aggregations. *Journal of Machine Learning Research*, 25(144):1–29, 2024.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(1):7897–7927, 2022.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: PAC learning and online learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 44707–44739, 2023.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*, pages 150–165, 2007.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Minimax policies for combinatorial prediction games. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 107–132. PMLR, 2011.

Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3), 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Brenda S. Baker. Approximation algorithms for NP-complete problems on planar graphs. *Journal of the Association for Computing Machinery*, 41(1):153–180, 1994.

Peter L. Bartlett and Philip M. Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 392–401, 1995.

Peter L. Bartlett and Philip M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

Peter L. Bartlett, Philip M. Long, and Robert C. Williamson. Fat-shattering and the learnability of real-valued functions. *J. Comput. Syst. Sci.*, 52(3):434–452, 1996.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *FAT/ML Workshop 2017*, 2017. URL `https://arxiv.org/abs/1705.08504v6`.

Mikhail Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361, 2018.

Shai Ben-David. 2 notes on classes with Vapnik-Chervonenkis dimension 1. *arXiv preprint*, arXiv:1507.05307, 2015. URL `https://arxiv.org/abs/1507.05307`.

Shai Ben-David, Nicolò Cesa-Bianchi, and Philip M Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 333–340, 1992.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose H. Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Advances in Neural Information Processing Systems*, pages 11345–11354, 2019.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. No weighted-regret learning in adversarial bandits with delays. *Journal of Machine Learning Research*, 23, 2022.

Guy Blanc, Jane Lange, and Li-Yang Tan. Provably efficient, succinct, and precise explanations. *Advances in Neural Information Processing Systems*, 34:6129–6141, 2021.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. doi: 10.1145/76359.76371.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

Olivier Bousquet, Steve Hanneke, Shay Moran, Jonathan Shafer, and Ilya O. Tolstikhin. Fine-grained distribution-dependent learning curves. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 5890–5924. PMLR, 2023.

Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 648–668. PMLR, 2024a.

Marco Bressan, Emmanuel Esposito, and Maximilian Thiessen. Efficient algorithms for learning monophonic halfspaces in graphs. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 669–696. PMLR, 2024b.

Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, pages 382–391, 1993.

Nicolò Cesa-Bianchi, Khaled Eldowa, Emmanuel Esposito, and Julia Olkhovskaya. Improved regret bounds for bandits with expert advice. *arXiv preprint*, arXiv:2406.16802, 2024. URL `https://arxiv.org/abs/2406.16802`.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, 1997.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 605–622. PMLR, 2016.

Nicolò Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sébastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 465–481. PMLR, 2017.

Nicolò Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Delay and cooperation in nonstochastic bandits. *Journal of Machine Learning Research*, 20, 2019.

Seok-Ho Chang, Pamela C. Cosman, and Laurence B. Milstein. Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011. doi: 10.1109/TCOMM.2011.072011.100049.

Houshuang Chen, Zengfeng Huang, Shuai Li, and Chihao Zhang. Understanding bandits with graph feedback. In *Advances in Neural Information Processing Systems*, pages 24659–24669, 2021.

Houshuang Chen, Yuchen He, and Chihao Zhang. On interpolating experts and multi-armed bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6776–6802. PMLR, 2024.

Vasek Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979.

Alon Cohen, Tamir Hazan, and Tomer Koren. Online learning with feedback graphs without the graphs. In *International Conference on Machine Learning*, pages 811–819, 2016.

Alon Cohen, Tamir Hazan, and Tomer Koren. Tight bounds for bandit combinatorial optimization. In *Conference on Learning Theory*, pages 629–642. PMLR, 2017.

Roberto Colomboni, Emmanuel Esposito, and Andrea Paudice. An improved uniform convergence bound with fat-shattering dimension. *Information Processing Letters*, 188, 2025.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with sleeping experts and feedback graphs. In *International Conference on Machine Learning*, pages 1370–1378, 2019.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Ningshan Zhang. Online learning with dependent stochastic feedback graphs. In *International Conference on Machine Learning*, pages 2154–2163, 2020.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8, 1995.

Amit Daniely and Tom Helbertal. The price of bandit information in multiclass online classification. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 93–104. PMLR, 2013.

Christoph Dann, Chen-Yu Wei, and Julian Zimmert. A blackbox approach to best of both worlds in bandits and beyond. In *The Thirty Sixth Annual Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5503–5570. PMLR, 2023.

Sanjoy Dasgupta, Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. Explainable k-means and k-medians clustering. In *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria*, pages 12–18, 2020.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Richard M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6 (6):899–929, 1978. doi: 10.1214/aop/1176995384.

Richard M. Dudley. A course on empirical processes. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XII - 1982*, pages 1–142. Springer Berlin Heidelberg, 1984.

Gintare Karolina Dziugaite, Shai Ben-David, and Daniel M. Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint*, arXiv:2010.13764, 2020. URL `https://arxiv.org/abs/2010.13764`.

Stephen G. Eick. The two-armed bandit with delayed responses. *The Annals of Statistics*, 1988.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information-theoretic regret bounds for bandits with fixed expert advice. In *IEEE Information Theory Workshop*, pages 30–35. IEEE, 2023a.

Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 46122–46133. Curran Associates, Inc., 2023b.

Khaled Eldowa, Nicolò Cesa-Bianchi, Alberto Maria Metelli, and Marcello Restelli. Information capacity regret bounds for bandits with mediator feedback. *Journal of Machine Learning Research*, 25(353):1–36, 2024.

Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. What is interpretability? *Philosophy & Technology*, 34(4):833–862, 2021.

Liad Erez and Tomer Koren. Towards best-of-all-worlds online learning with feedback graphs. *Advances in Neural Information Processing Systems*, 34:28511–28521, 2021.

Emmanuel Esposito, Federico Fusco, Dirk van der Hoeven, and Nicolò Cesa-Bianchi. Learning on the edge: Online learning with stochastic feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 34776–34788, 2022.

Emmanuel Esposito, Saeed Masoudian, Hao Qiu, Dirk van der Hoeven, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Delayed bandits: When do intermediate observations help? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9374–9395. PMLR, 2023.

Zhili Feng and Po-Ling Loh. Online learning with graph-structured feedback against adaptive adversaries. In *IEEE International Symposium on Information Theory*, pages 931–935, 2018.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Pierre Gaillard, Gilles Stoltz, and Tim van Erven. A second-order bound with excess losses. In *Conference on Learning Theory*, pages 176–196, 2014.

Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of LIME. In *International conference on artificial intelligence and statistics*, pages 1287–1296, 2020. URL `https://arxiv.org/abs/2008.11092v3`.

Pouya M. Ghari and Yanning Shen. Online learning with probabilistic feedback. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4183–4187, 2022.

Pouya M. Ghari and Yanning Shen. Online learning with uncertain feedback graphs. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9636–9650, 2024. doi: 10.1109/TNNLS.2023.3235734.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

Lee-Ad Gottlieb, Eran Kaufman, Aryeh Kontorovich, and Gabriel Nivasch. Learning convex polyhedra with margin. *IEEE Transactions on Information Theory*, 68(3):1976–1984, 2022.

András György and Pooria Joulani. Adapting to delays and data in adversarial multi-armed bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021.

Magnús Már Halldórsson and Jaikumar Radhakrishnan. Greed is good: Approximating independent sets in sparse and bounded-degree graphs. *Algorithmica*, 18:145–163, 1997.

Paul R. Halmos. *Measure theory*, volume 18. Springer, 2013. doi: 10.1007/978-1-4684-9440-2. URL `https://doi.org/10.1007/978-1-4684-9440-2`.

Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129–2150, 2021. doi: 10.1214/20-AOS2029.

David Haussler, Jyrki Kivinen, and Manfred K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, 1998. doi: 10.1109/18.705569.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Dirk van der Hoeven and Nicolò Cesa-Bianchi. Nonstochastic bandits and experts with arm-dependent delays. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Dirk van der Hoeven, Tim van Erven, and Wojciech Kotłowski. The many faces of exponential weights in online learning. In *Conference on Learning Theory*, pages 2067–2092, 2018.

Dirk van der Hoeven, Federico Fusco, and Nicolò Cesa-Bianchi. Beyond bandit feedback in online multiclass classification. In *Advances in Neural Information Processing Systems*, pages 13280–13291, 2021.

Dirk van der Hoeven, Lukas Zierahn, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and mdps. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1285–1321. PMLR, 2023.

Lunjia Hu, Charlotte Peale, and Omer Reingold. Metric entropy duality and the sample complexity of outcome indistinguishability. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, pages 515–552, 2022.

Johan Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Mathematica*, 182:105–142, 1999.

Shinji Ito. On the minimax regret for contextual linear bandits and multi-armed bandits with expert advice. In *Advances in Neural Information Processing Systems*, 2024.

Shinji Ito, Daisuke Hatano, Hanna Sumita, Kei Takemura, Takuro Fukunaga, Naonori Kakimura, and Ken-Ichi Kawarabayashi. Improved regret bounds for bandit combinatorial optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, volume 35, pages 28631–28643, 2022.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. Improved best-of-both-worlds guarantees for multi-armed bandits: FTRL with general regularizers and multiple optimal arms. In *Advances in Neural Information Processing Systems*, volume 36, pages 30918–30978. Curran Associates, Inc., 2023.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, 2013.

Satyen Kale. Multiarmed bandits with limited expert advice. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 107–122. PMLR, 13–15 Jun 2014.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher, editors, *Proceedings of a Symposium on the Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

Michael J. Kearns and Yishay Mansour. On the boosting ability of top-down decision tree learning algorithms. *J. Comput. Syst. Sci.*, 58(1):109–128, 1999. doi: 10.1006/JCSS.1997.1543.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. *Machine learning*, 80(2):245–272, 2010.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, volume 27, pages 613–621, 2014.

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with Erdős-Rényi side-observation graphs. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016a.

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In *International Conference on Artificial Intelligence and Statistics*, pages 1186–1194, 2016b.

Wojciech Kotłowski. On minimaxity of follow the leader strategy in the stochastic setting. *Theoretical Computer Science*, 742:50–65, 2018. Algorithmic Learning Theory.

Joon Kwon and Vianney Perchet. Gains and losses are fundamentally different in regret minimization: The sparse case. *Journal of Machine Learning Research*, 17(227):1–32, 2016.

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

Jeffrey Li, Vaishnavh Nagarajan, Gregory Plumb, and Ameet Talwalkar. A learning theoretic perspective on local explainability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7aL-OtQrBWD.

Shuai Li, Wei Chen, Zheng Wen, and Kwong-Sak Leung. Stochastic online learning with probabilistic graph feedback. In *Proceeding of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 4675–4682, 2020.

Nick Littlestone. *Mistake bounds and logarithmic linear-threshold learning algorithms*. PhD thesis, University of California, Santa Cruz, 1990. UMI Order No: GAX89-26506.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Fang Liu, Swapna Buccapatnam, and Ness B. Shroff. Information directed sampling for stochastic bandits with graph feedback. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3643–3650, 2018.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1739–1776. PMLR, 2018.

Haipeng Luo, Hanghang Tong, Mengxiao Zhang, and Yuheng Zhang. Improved high-probability regret for adversarial bandits with time-varying feedback graphs. In *International Conference on Algorithmic Learning Theory*, pages 1074–1100. PMLR, 2023.

Timothy Arthur Mann, Sven Gowal, András György, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Learning from delayed outcomes via proxies with applications to recommender systems. In *International Conference on Machine Learning*, pages 4324–4332, 2019.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in Neural Information Processing Systems*, 24, 2011.

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

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. *arXiv preprint*, arXiv:2308.10675, 2023. URL `https://arxiv.org/abs/2308.10675`.

Saeed Masoudian, Julian Zimmert, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. *arXiv preprint*, arXiv:2308.10675, 2024. URL `https://arxiv.org/abs/2308.10675`.

H. Brendan McMahan and Matthew J. Streeter. Tighter bounds for multi-armed bandits with expert advice. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Shahar Mendelson and Gideon Schechtman. The shattering dimension of sets of linear functionals. *The Annals of Probability*, 32(3):1746–1770, 2004.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3): 21:1–21:10, 2016. doi: 10.1145/2890490.

Michal Moshkovitz, Yao-Yuan Yang, and Kamalika Chaudhuri. Connecting interpretability and robustness in decision trees through separation. In *International Conference on Machine Learning*, pages 7839–7849. PMLR, 2021.

Márton Naszódi. Approximating a convex body by a polytope using the epsilon-net theorem. *Discrete & Computational Geometry*, 61:686–693, 2019.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, 2010.

Gergely Neu, András György, Csaba Szepesvári, and András Antos. Online markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928. URL `http://eudml.org/doc/159291`.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint*, arXiv:1912.13213, 2019. URL `https://arxiv.org/abs/1912.13213`.

David Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions, 1986.

David Pollard. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 2, 1990.

Alexander Rakhlin and Karthik Sridharan. Statistical learning theory and sequential prediction, October 2014. URL `https://www.cs.cornell.edu/~sridharan/lecnotes.pdf`.

Anshuka Rangi and Massimo Franceschetti. Online learning with feedback graphs and switching costs. In *International Conference on Artificial Intelligence and Statistics*, pages 2435–2444, 2019.

Alon Resler and Yishay Mansour. Adversarial online learning with noise. In *International Conference on Machine Learning*, pages 5429–5437, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Conference on Learning Theory*, pages 3227–3249. PMLR, 2020.

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

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

Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164:603–648, 2006.

Shuichi Sakai, Mitsunori Togasaki, and Koichi Yamazaki. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126(2-3):313–322, 2003.

Yevgeny Seldin and Gábor Lugosi. A lower bound for multi-armed bandits with expert advice. In *The 13th European Workshop on Reinforcement Learning (EWRL)*, volume 2, page 7, 2016.

Yevgeny Seldin, Koby Crammer, and Peter Bartlett. Open problem: Adversarial multiarmed bandits with limited advice. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 1067–1072. PMLR, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010.

Richard Simon. Adaptive treatment assignment methods and clinical trials. *Biometrics*, 33, 1977.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2), 2019.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint*, arXiv:1904.07272, 2024. URL `https://arxiv.org/abs/1904.07272`.

Clifford Smyth. Reimer's inequality and Tardos' conjecture. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, pages 218–221, 2002.

Eiji Takimoto and Akira Maruoka. Top-down decision tree learning as information based boosting. *Theoretical Computer Science*, 292(2):447–464, 2003. ISSN 0304-3975. doi: 10.1016/S0304-3975(02)00181-0. Theoretical Aspects of Discovery Science.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

Gábor Tardos. Query complexity, or why is it difficult to separate $\mathrm{NP}^A \cap \mathrm{coNP}^A$ from P^A by random oracles A? *Combinatorica*, 9:385–392, 1989.

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

Tobias Sommer Thune, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Nonstochastic multiarmed bandits with unrestricted delays. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. ISSN 0001-0782. doi: 10.1145/1968.1972.

Vladimir N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, COLT '89, pages 3–21, 1989.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pages 11–30. Springer International Publishing, 2015.

Vijay V. Vazirani. *Approximation Algorithms*. Springer, 2001.

Claire Vernade, András György, and Timothy A. Mann. Non-stationary delayed bandits with intermediate observations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 2020.

Thibaut Vidal and Maximilian Schiffer. Born-again tree ensembles. In *International Conference on Machine Learning*, pages 9743–9753, 2020.

Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, 1990.

Nuri Mert Vural, Hakan Gokcesu, Kaan Gokcesu, and Suleyman S. Kozat. Minimax optimal algorithms for adversarial bandit problem with multiple plays. *IEEE Transactions on Signal Processing*, 67(16):4383–4398, 2019.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science*, pages 222–227, 1977.

Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting CNNs via decision trees. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6261–6270, 2019.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

Julian Zimmert and Yevgeny Seldin. An optimal algorithm for adversarial bandits with arbitrary delays. In *Proceedings on the International Conference on Artificial Intelligence and Statistics*, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3:103–128, 2007.

Appendix A

Proof Details for Chapter 3

A.1 Auxiliary Results

Lemma A.1. *If Algorithm 3.1 is run with $q \in (0,1)$, learning rate $\eta > 0$, and non-negative loss estimates that satisfy $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$ for all $t = 1, \dots, T$, then its regret satisfies*

$$R_T \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i \in V} p_t(i)^{2-q}\, \widehat{\ell}_t(i)^2\right] .$$

Proof. Let $i^* \in \arg\min_{i \in V} \sum_{t=1}^{T} \ell_t(i)$ be an action that minimizes the cumulative loss, and let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^*. Recall that for $t \in [T]$, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid I_1, \dots, I_{t-1}]$, and notice that p_t is measurable with respect to the σ-algebra generated by $I_1, \dots, I_{t-1}$. Hence, using that

$$\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i \in V} p_t(i)\ell_t(i) \qquad \text{and} \qquad \mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t ,$$

we have, via the tower rule and the linearity of expectation, that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t)\right] - \sum_{t=1}^{T} \ell_t(i^*) = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \ell_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t \rangle\right] ,$$

from which we can obtain the desired result by using Lemma A.2 (which holds even if the loss $\widehat{\ell}_t$ at each round $t \in [T]$ depends on the prediction p_t made at that round). □

Lemma A.2. *Let $q \in (0,1)$, $\eta > 0$, and $(y_t)_{t=1}^{T}$ be an arbitrary sequence of non-negative loss vectors in $\mathbb{R}^K$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_K and the q-Tsallis regularizer ψ_q over this sequence of losses. That is, $p_1 \in \arg\min_{p := \Delta_K} \psi_q(p)$, and for $t \in [T]$,*

$$p_{t+1} \coloneqq \arg\min_{p \in \Delta_K} \eta \sum_{s=1}^{t} \langle y_s, p \rangle + \psi_q(p) .$$

Then for any $u \in \Delta_K$,

$$\sum_{t=1}^{T} \langle p_t - u, y_t \rangle \le \frac{K^{1-q}}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i \in V} p_t(i)^{2-q}\, y_t(i)^2 .$$

Proof. By Theorem 28.5 in Lattimore and Szepesvári (2020), we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t - u, y_t \rangle &\le \frac{\psi_q(u) - \psi_q(p_1)}{\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) \\
&= \frac{K^{1-q} - 1}{(1-q)\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) \\
&\le \frac{K^{1-q}}{(1-q)\eta} + \sum_{t=1}^{T} \left(\langle p_t - p_{t+1}, y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \right) ,
\end{aligned}$$

where $D_{\psi_q}(\cdot, \cdot)$ is the Bregman divergence based on ψ_q. For bounding each summand in the second term, we follow a similar argument to that used in Theorem 30.2 in Lattimore and Szepesvári (2020). Namely, for each $i \in V$ and round $t \in [T]$, define $\overline{y}_t(i) \coloneqq \mathbb{I}\{p_{t+1}(i) \le p_t(i)\} y_t(i)$. We then have that

$$\begin{aligned}
\langle p_t - p_{t+1}, &y_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&\le \langle p_t - p_{t+1}, \overline{y}_t \rangle - \frac{1}{\eta} D_{\psi_q}(p_{t+1}, p_t) \\
&= \frac{1}{\eta} \langle p_t - p_{t+1}, \eta \overline{y}_t \rangle - \frac{1}{2\eta} \left\| p_{t+1} - p_t \right\|^2_{\nabla^2 \psi_q(z_t)} \\
&\le \frac{\eta}{2} \left\| \overline{y}_t \right\|^2_{(\nabla^2 \psi_q(z_t))^{-1}} \\
&= \frac{\eta}{2q} \sum_{i \in V} z_t(i)^{2-q} \, \overline{y}_t(i)^2 \\
&= \frac{\eta}{2q} \sum_{i \in V} \bigl(\gamma_t p_{t+1}(i) + (1 - \gamma_t) p_t(i) \bigr)^{2-q} \, \overline{y}_t(i)^2 \\
&\le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, \overline{y}_t(i)^2 + \gamma_t \frac{\eta}{2q} \sum_{i \in V} \bigl(p_{t+1}(i)^{2-q} - p_t(i)^{2-q} \bigr) \, \overline{y}_t(i)^2 \\
&\le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, \overline{y}_t(i)^2 \\
&\le \frac{\eta}{2q} \sum_{i \in V} p_t(i)^{2-q} \, y_t(i)^2 ,
\end{aligned}$$

where $z_t \coloneqq \gamma_t p_{t+1} + (1 - \gamma_t) p_t$ for some $\gamma_t \in [0, 1]$; the first inequality holds due to the non-negativity of the losses, the second inequality is an application of the Fenchel-Young inequality, the second equality holds since the Hessian of ψ_q is a diagonal matrix with $(\nabla^2 \psi_q(x))_{i,i} = q x(i)^{q-2}$, the third inequality is an application of Jensen's inequality (since $q \in (0, 1)$), and the fourth inequality holds since $\overline{y}_t(i) = 0$ for any i such that $p_{t+1}(i)^{2-q} > p_t(i)^{2-q}$. □

Lemma A.3. *Let a and b be positive integers such that $2 \le a \le b$, and let $n = \lceil \log_2 a \rceil$. Then,*

$$\sum_{r=0}^{n-1} \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)} \le \frac{\sqrt{2\pi} + 2\sqrt{2 - \ln 2}}{\ln 2} \sqrt{a \ln\left(\frac{e^2 b}{a}\right)} .$$

Proof. Since $n \le \log_2(2b)$ and $2^r \ln\bigl(e^2 b 2^{-r}\bigr)$ is monotonically increasing in r for $r \in [0, \log_2(eb)]$, we can bound the sum by an integral:

$$\sum_{r=0}^{n-1} \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)} \le \int_0^n \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)}\, \mathrm{d}r\ .$$

We proceed via a change of variable; let $x \coloneqq e^2 b 2^{-r}$, and note that $\mathrm{d}r = -\frac{\mathrm{d}x}{x \ln 2}$. We then have that

$$\begin{aligned}
\int_0^n \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)}\, \mathrm{d}r &= \sqrt{e^2 b} \int_0^n \sqrt{\frac{2^r}{e^2 b} \ln\bigl(e^2 b 2^{-r}\bigr)}\, \mathrm{d}r \\
&= -\frac{e\sqrt{b}}{\ln 2} \int_{e^2 b}^{e^2 b 2^{-n}} \sqrt{\frac{\ln x}{x^3}}\, \mathrm{d}x = \frac{e\sqrt{b}}{\ln 2} \int_{e^2 b 2^{-n}}^{e^2 b} \sqrt{\frac{\ln x}{x^3}}\, \mathrm{d}x \\
&= \frac{e\sqrt{b}}{\ln 2} \Bigl[-\sqrt{2\pi} \cdot \operatorname{erfc}\bigl(\sqrt{(\ln x)/2}\bigr) - 2\sqrt{(\ln x)/x} \Bigr]_{e^2 b 2^{-n}}^{e^2 b} \\
&\le \frac{e\sqrt{b}}{\ln 2} \left(\sqrt{2\pi} \cdot \operatorname{erfc}\bigl(\sqrt{\ln(e^2 b 2^{-n})/2}\bigr) + 2\sqrt{\frac{2^n \ln(e^2 b 2^{-n})}{e^2 b}} \right)\ ,
\end{aligned}$$

where $\operatorname{erfc}(x) \coloneqq 1 - \frac{2}{\sqrt{\pi}} \int_0^x \exp(-z^2)\, \mathrm{d}z$ is the complementary Gaussian error function, which is always positive. By Chang, Cosman, and Milstein (2011, Theorem 1), we have that $\operatorname{erfc}(x) \le \exp(-x^2)$. Consequently,

$$\begin{aligned}
\int_0^n \sqrt{2^r \ln\bigl(e^2 b 2^{-r}\bigr)}\, \mathrm{d}r &\le \frac{e\sqrt{b}}{\ln 2} \left(\sqrt{2\pi} \sqrt{\frac{2^n}{e^2 b}} + 2\sqrt{\frac{2^n \ln(e^2 b 2^{-n})}{e^2 b}} \right) \\
&= \frac{\sqrt{2^n}}{\ln 2} \left(\sqrt{2\pi} + 2\sqrt{\ln(e^2 b 2^{-n})} \right) \\
&\le \frac{\sqrt{2a}}{\ln 2} \left(\sqrt{2\pi} + 2\sqrt{\ln\left(\frac{e^2 b}{2a}\right)} \right) \\
&\le \frac{\sqrt{2\pi} + 2\sqrt{2 - \ln 2}}{\ln 2} \sqrt{a \ln\left(\frac{e^2 b}{a}\right)}\ ,
\end{aligned}$$

where in the second inequality we used once again the fact that $2^r \ln\bigl(e^2 b 2^{-r}\bigr)$ is monotonically increasing in r for $r \in [0, \log_2(eb)]$ to replace n with $\log_2(a) + 1$, and the last inequality holds since $b \ge a$. □

A.2 Proofs of Section 3.3

In this section, we provide the proof of Theorem 3.2, which is restated below.

Theorem 3.2. *Let $G_1, \dots, G_T$ be a sequence of undirected strongly observable feedback graphs, where each G_t has independence number $\alpha_t = \alpha$ for some common value $\alpha \in [K]$. If Algorithm 3.1 is run with input*

$$q = \frac{1}{2}\left(1 + \frac{\ln(K/\alpha)}{\sqrt{\ln^2(K/\alpha) + 4} + 2}\right) \in [1/2, 1) \qquad \textit{and} \qquad \eta = \frac{1}{3}\sqrt{\frac{2qK^{1-q}}{T(1-q)\alpha^q}}\ ,$$

and loss estimates (3.6)*, then its regret satisfies $R_T \le 6\sqrt{e\alpha T\,(2 + \ln(K/\alpha))}$.*

Proof. Let $i^* \in \arg\min_{i\in V} \sum_{t=1}^{T} \ell_t(i)$ and $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be its indicator vector. Whenever J_t is nonempty, let $j_t \in V$ be the only action such that $J_t = \{j_t\}$. Similarly to Zimmert and Lattimore (2019), let $z_t := \mathbb{I}\{J_t \neq \emptyset\}\,\mathbb{I}\{I_t \in N_t(j_t)\}\,\frac{1-\ell_t(j_t)}{1-p_t(j_t)}$ and define new losses $\widetilde{\ell}_t(i) := \widehat{\ell}_t(i) + z_t$ for each time step $t \in [T]$ and each action $i \in V$. Since $p_t, \mathbf{e}_{i^*} \in \Delta_K$, we have that $\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ for every $t \in [T]$. Then, using the fact that $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$, we get that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] ,$$

where the first equality holds via the same arguments made in the proof of Lemma A.1. If we consider the optimization step of Algorithm 3.1, computing the same inner product over the new losses $\widetilde{\ell}_1, \dots, \widetilde{\ell}_T$ for some $p \in \Delta_K$ gives

$$\left\langle \sum_{s=1}^{t} \widetilde{\ell}_s, p \right\rangle = \sum_{s=1}^{t} z_s + \left\langle \sum_{s=1}^{t} \widehat{\ell}_s, p \right\rangle ,$$

where the sum $\sum_{s=1}^{t} z_s$ is constant with respect to p. This implies that the objective functions in terms of either $(\widehat{\ell}_t)_{t\in[T]}$ and $(\widetilde{\ell}_t)_{t\in[T]}$, respectively, are minimized by the same probability distributions. However, notice that, unlike $(\widehat{\ell}_t)_{t\in[T]}$, the loss vectors in $(\widetilde{\ell}_t)_{t\in[T]}$ are always non-negative. Consequently, similar to the proof of Lemma A.1, we may apply Lemma A.2 to upper bound the regret of Algorithm 3.1 in terms of the losses $(\widetilde{\ell}_t)_{t\in[T]}$. Doing so gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \le \frac{K^{1-q}}{\eta(1-q)} + \frac{\eta}{2q} \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i\in V} p_t(i)^{2-q} \mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big]\right] . \tag{A.1}$$

We can bound the second term by observing that $\widetilde{\ell}_t(j_t) = 1$ whenever $J_t \neq \emptyset$. Therefore,

$$\begin{aligned}
\sum_{i\in V} p_t(i)^{2-q} \mathbb{E}_t\Big[\widetilde{\ell}_t(i)^2\Big] &\le 2 \sum_{i\in V\setminus J_t} p_t(i)^{2-q} \mathbb{E}_t\Big[\widehat{\ell}_t(i)^2\Big] + 2\mathbb{E}_t\left[z_t^2\right] \sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1 \\
&\le 2 \sum_{i\in V\setminus J_t} \frac{p_t(i)^{2-q}}{P_t(i)} + 2\mathbb{E}_t\left[z_t^2\right] \sum_{i\in V\setminus J_t} p_t(i)^{2-q} + 1 \\
&\le 2 \sum_{i\in V\setminus J_t} \frac{p_t(i)^{2-q}}{P_t(i)} + 3 ,
\end{aligned}$$

where the second inequality holds because $\mathbb{E}_t\big[\widehat{\ell}_t(i)^2\big] \le 1/P_t(i)$ for all $i \notin J_t$, and the third inequality follows from the fact that

$$\mathbb{E}_t\left[z_t^2\right] \sum_{i\in V\setminus J_t} p_t(i)^{2-q} = \mathbb{I}\{J_t \neq \emptyset\} \frac{\big(1-\ell_t(j_t)\big)^2}{1-p_t(j_t)} \sum_{i\in V\setminus J_t} p_t(i)^{2-q} \le 1 .$$

We can handle the remaining sum by separating it over nodes $i \in S_t$, which satisfy $P_t(i) = 1 - p_t(i)$ because of strong observability, and those in $\overline{S}_t = V \setminus S_t$. In the first case, any node $i \in S_t \setminus J_t$ has

$p_t(i) \le 1/2$ and thus

$$\sum_{i \in S_t \setminus J_t} \frac{p_t(i)^{2-q}}{P_t(i)} = \sum_{i \in S_t \setminus J_t} \frac{p_t(i)^{2-q}}{1 - p_t(i)} \le 2 \sum_{i \in S_t \setminus J_t} p_t(i)^{2-q} \le 2 \; .$$

while in the second case we have that $\sum_{i \in \overline{S}_t} p_t(i)^{2-q}/P_t(i) \le \alpha^q$ by Lemma 3.1 with $U = \overline{S}_t$ and $b = 1 - q$. Overall, we have shown that

$$\sum_{i \in V} p_t(i)^{2-q} \mathbb{E}_t\left[\widetilde{\ell}_t(i)^2\right] \le 2 \sum_{i \in \overline{S}_t} \frac{p_t(i)^{2-q}}{P_t(i)} + 7 \le 2\alpha^q + 7 \le 9\alpha^q \; . \tag{A.2}$$

Plugging back into (A.1), we obtain that

$$\begin{aligned} R_T &\le \frac{K^{1-q}}{\eta(1-q)} + \frac{9\eta}{2q}\alpha^q T \\ &= 3\sqrt{\frac{2K^{1-q}\alpha^q}{q(1-q)}T} \\ &\le 6\sqrt{e\alpha T\left(2 + \ln(K/\alpha)\right)} \; , \end{aligned}$$

where the equality is due to our choice of η, and the last inequality follows as in the proof of Theorem 3.1 together with our choice of q. □

A.3 Proofs of Section 3.4

In this section, we provide the proof of Theorem 3.3, which is restated below.

Theorem 3.3. *Let $C := 4\sqrt{6e}\frac{\sqrt{\pi}+\sqrt{4-2\ln 2}}{\ln 2}$. Then, the regret of Algorithm 3.2 satisfies*

$$R_T \le C\sqrt{\sum_{t=1}^{T} \alpha_t \left(2 + \ln\left(\frac{K}{\bar{\alpha}_T}\right)\right)} + \log_2 \bar{\alpha}_T \; .$$

Proof. Notice that if $\bar{\alpha}_T = 1$, the initial guess is correct and the algorithm will never restart. Moreover, since in this case we have that $\alpha_t = 1$ for all t, the theorem follows trivially from the regret bound of Theorem 3.2. Hence, we can assume for what follows that $\bar{\alpha}_T > 1$. Let $i^* \in \arg\min_{i \in [K]} \sum_{t=1}^{T} \ell_t(i)$ and $n = \lceil \log_2 \bar{\alpha}_T \rceil$. Note that the maximum value of r that the algorithm can reach is $n - 1$. To see this, observe that Lemma 3.1 implies that for any r and t, $H_t(q_r) \le \alpha_t^{q_r}$. Consequently, for any $t \ge T_r$,

$$\frac{1}{T} \sum_{s=T_r}^{t} H_s(q_r)^{1/q_r} \le \frac{1}{T} \sum_{s=T_r}^{t} \alpha_s \le \bar{\alpha}_T \le 2^n \; .$$

For $t \in [T]$, let r_t be the value of r at round t. Without loss of generality, we assume that r takes each value in $\{0, \dots, n-1\}$ for at least two rounds. Additionally, we define $T_n = T + 2$ for convenience. We start by decomposing the regret over the n intervals (each corresponding to a value of r in $\{0, \dots, n-1\}$) and bounding the instantaneous regret with 1 for each step in which we restart (i.e., at the last step of each but the last interval):

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\ &\le \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] + n - 1 \\ &\le \mathbb{E}\left[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] + \log_2 \bar{\alpha}_T \ . \end{aligned} \tag{A.3}$$

For what follows, let $\mathbf{e}_{i^*} \in \mathbb{R}^K$ be an indicator vector for i^* and let $\widetilde{\ell}_t$ be as defined in the proof of Theorem 3.2. Fix $r \in \{0, \dots, n-1\}$, we proceed by bounding the regret in the interval $[T_r, T_{r+1} - 2]$:

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\right\}\big(\ell_t(I_t) - \ell_t(i^*)\big)\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\right\}\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\right\}\langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \\ &= \mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle\right] \\ &\overset{(c)}{\le} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\right] \\ &\overset{(d)}{=} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\right\}\mathbb{E}_t\left[\sum_{i\in V} p_t(i)^{2-q_r}\widetilde{\ell}_t(i)^2\right]\right] \\ &\overset{(e)}{\le} \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\left\{r_t = r, \frac{1}{T}\sum_{s=T_{r_t}}^{t} H_s(q_{r_t})^{1/q_{r_t}} \le 2^{r_t+1}\right\}(2H_t(q_r) + 7)\right] \\ &= \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{\eta_r}{2q_r}\mathbb{E}\left[\sum_{t=T_r}^{T_{r+1}-2}(2H_t(q_r) + 7)\right] , \end{aligned} \tag{A.4}$$

where (a) follows since $\mathbb{E}_t\big[\ell_t(I_t)\big] = \sum_{i\in V} p_t(i)\ell_t(i)$, $\mathbb{E}_t\big[\widehat{\ell}_t\big] = \ell_t$, and the indicator at round t is measurable with respect to $\sigma(I_1, \dots, I_{t-1})$, that is, the σ-algebra generated by $I_1, \dots, I_{t-1}$; (b) follows since $\langle p_t - \mathbf{e}_{i^*}, \widehat{\ell}_t\rangle = \langle p_t - \mathbf{e}_{i^*}, \widetilde{\ell}_t\rangle$ holds by the definition of $\widetilde{\ell}_t$; (c) is an application of Lemma A.2, justifiable with the same argument leading to (A.1) in the proof of Theorem 3.2; (d) uses once again that the indicator at round t is measurable with respect to $\sigma(I_1, \dots, I_{t-1})$; finally, (e) follows via

(A.2). Define $T_{r:r+1} = T_{r+1} - T_r - 1$, and notice that

$$\begin{aligned}
\sum_{t=T_r}^{T_{r+1}-2} H_t(q_r) &= \frac{T_{r:r+1}}{T_{r:r+1}} \sum_{t=T_r}^{T_{r+1}-2} \big(H_t(q_r)^{1/q_r}\big)^{q_r} \\
&\le T_{r:r+1}\bigg(\frac{1}{T_{r:r+1}} \sum_{t=T_r}^{T_{r+1}-2} H_t(q_r)^{1/q_r}\bigg)^{q_r} \\
&\le T_{r:r+1}\bigg(\frac{T}{T_{r:r+1}} 2^{r+1}\bigg)^{q_r} \\
&\le 2T\big(2^r\big)^{q_r} \ ,
\end{aligned}$$

where the first inequality follows due to Jensen's inequality since $q_r \in (0,1)$, and the second follows from the restarting condition of Algorithm 3.2. Next, we plug this inequality back into (A.4), and then, similar to the proof of Theorem 3.2, we use the definitions of η_r and q_r and bound the resulting expression to get that

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{t=T_r}^{T_{r+1}-2} \big(\ell_t(I_t) - \ell_t(i^*)\big)\Bigg] &\le \frac{K^{1-q_r}}{\eta_r(1-q_r)} + \frac{11\eta_r}{2q_r} T\,(2^r)^{q_r} \\
&\le 2\sqrt{11eT2^r\Big(2+\ln\big(K2^{-r}\big)\Big)} \le 4\sqrt{3eT2^r \ln\big(e^2K2^{-r}\big)} \ .
\end{aligned}$$

We then sum this quantity over r and use Lemma A.3 to get that

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{r=0}^{n-1}\sum_{t=T_r}^{T_{r+1}-2} \big(\ell_t(I_t) - \ell_t(i^*)\big)\Bigg] &\le 4\sqrt{3eT}\sum_{r=0}^{n-1}\sqrt{2^r\ln\big(e^2K2^{-r}\big)} \\
&\le 4\sqrt{6e}\,\frac{\sqrt{\pi}+\sqrt{4-2\ln 2}}{\ln 2}\sqrt{\bar{\alpha}_T T\big(2+\ln\left(K/\bar{\alpha}_T\right)\big)} \ ,
\end{aligned}$$

which, together with (A.3), concludes the proof. □

A.4 Proof of the Lower Bound

In this section, we prove the lower bound provided in Section 3.5, which we restate below. As remarked before, our proof makes use of known techniques for proving lower bounds for the multitask bandit problem. In particular, parts of the proof are adapted from the proof of Theorem 7 in Eldowa et al. (2023a).

Theorem 3.4. *Pick any $K \ge 2$ and any α such that $2 \le \alpha \le K$. Then, for any algorithm and for all $T \ge \frac{\alpha \log_\alpha K}{4\log(4/3)}$, there exists a sequence of losses and feedback graphs $G_1, \ldots, G_T$ such that $\alpha(G_t) = \alpha$ for all $t \in [T]$ and*

$$R_T \ge \frac{1}{18\sqrt{2}}\sqrt{\alpha T \log_\alpha K}.$$

Proof. Once again, we define $M = \log_\alpha K$, which we assume for now to be an integer; we discuss in the end how to extend the proof to the case when it is not. The proof will be divided into five parts I–V. We begin by formalizing the class of environments described in Section 3.5 and stating two useful lemmas.

I. Preliminaries

We remind the reader that we identify each action in V with a vector $a \coloneqq \big(a(1), \dots, a(M)\big) \in [\alpha]^M$. We will focus on a set of M undirected graphs $\mathcal{G} \coloneqq \{G^i\}_{i=1}^M$, where G^i consists of α isolated cliques (with self-loops) $\{C_{i,j}\}_{j=1}^{\alpha}$ such that an action a belongs to clique $C_{i,j}$ if and only if $a(i) = j$. As remarked before, all these graphs have independence number α. For convenience, we also use actions in V as functions from $\mathcal{G}$ to $[\alpha]$, with $a(G^i) \coloneqq a(i)$.

An environment is identified by a function $\mu \colon [\alpha] \times \mathcal{G} \to [0,1]$ such that at every round t, after having drawn a graph G_t from the uniform distribution over $\mathcal{G}$ (denoted with $U_{\mathcal{G}}$), the environment latently draws for each $j \in [\alpha]$ and $G \in \mathcal{G}$, a Bernoulli random variable $\gamma_t(j; G)$ with mean $\mu(j; G_t)$. Subsequently, for defining the loss of action $a \in V$ at round t, we simply set $\ell_t(a) \coloneqq \gamma_t(a(G_t); G_t)$, whose expectation, conditioned on G_t, is $\mu(a(G_t); G_t)$. To simplify the notation, we use $\mu(a; G)$ as shorthand for $\mu(a(G); G)$ and $\gamma_t(a; G)$ as shorthand for $\gamma_t(a(G); G)$. Denote by A_t the action picked by the player at round t, which is chosen prior to observing G_t. We will focus on the following notion of stochastic regret, which we define for environment μ as:

$$\overline{R}_T(\mu) = \max_{a \in V} \mathbb{E}_\mu\left[\sum_{t=1}^{T} (\ell_t(A_t) - \ell_t(a))\right] ,$$

where $\mathbb{E}_\mu[\cdot]$ denotes the expectation with respect to the sequence of losses and graphs generated by environment μ, as well as the randomness in the choices of the player. We can use the tower rule to rewrite this expression as

$$\begin{aligned}
\overline{R}_T(\mu) &= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\ell_t(A_t) - \ell_t(a) \,\middle|\, G_t, A_t\right] \,\middle|\, A_t\right]\right] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mu(A_t; G_t) - \mu(a; G_t) \,\middle|\, A_t\right]\right] \\
&= \max_{a \in V} \sum_{t=1}^{T} \mathbb{E}_\mu\left[\sum_{i=1}^{M} U_{\mathcal{G}}(G^i)(\mu(A_t; G^i) - \mu(a; G^i))\right] \\
&= \max_{a \in V} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_\mu\left[\sum_{t=1}^{T} (\mu(A_t; G^i) - \mu(a; G^i))\right] . \qquad \text{(A.5)}
\end{aligned}$$

For a fixed algorithm, one can show via standard arguments that

$$\sup_{(\ell_t)_{t=1}^T, (G_t)_{t=1}^T} R_T \geq \sup_\mu \overline{R}_T(\mu) .$$

Hence, it suffices for our purposes to prove a lower bound for the right-hand side of this inequality.

In the following, we will have to be more precise about the probability measure with respect to which the expectation in (A.5) is defined. Let $\boldsymbol{\lambda}_t \in \{0,1\}^{K/\alpha}$ denote the vector of losses observed by the player at round t, which corresponds to the losses of the actions connected to A_t assuming that a systematic ordering of the actions makes it clear which coordinate of $\boldsymbol{\lambda}_t$ belongs to which action. Let $\mathbf{1}_{K/\alpha}$ and $\mathbf{0}_{K/\alpha}$ be the K/α dimensional* vectors of all ones and all zeros respectively. Clearly,

*Note that $K/\alpha = \alpha^{M-1}$ is an integer since $M (\geq 1)$ was assumed to be an integer.

we have that $\boldsymbol{\lambda}_t = \gamma_t(A_t; G_t)\mathbf{1}_{K/\alpha} = \ell_t(A_t)\mathbf{1}_{K/\alpha}$, which is a binary random variable taking values in $\{\mathbf{0}_{K/\alpha}, \mathbf{1}_{K/\alpha}\}$. Let $\mathbb{P}^\lambda_\mu$ be the probability distribution of $\boldsymbol{\lambda}_t$ in environment μ. Notice then that we have that

$$\mathbb{P}^\lambda_\mu(\gamma_t = \mathbf{1}_{K/\alpha} \,|\, G_t = G, A_t = a) = \mu(a; G) \ . \tag{A.6}$$

Let $H_t = (A_1, G_1, \boldsymbol{\lambda}_1, \dots, A_t, G_t, \boldsymbol{\lambda}_t) \in (V \times \mathcal{G} \times \{0,1\}^{K/\alpha})^t$ be the interaction trajectory after t steps. The policy π adopted by the player can be modelled as a sequence of probability kernels $\{\pi_t\}_{t=1}^T$ each mapping the trajectory so far to a distribution over the actions, i.e., A_t is sampled from $\pi_t(\cdot \,|\, H_{t-1})$. An environment μ and a policy π (implicit in the notation, and fixed throughout the rest of the proof) together define a distribution $\mathbb{P}_\mu$ over the set of possible trajectories of T steps such that

$$\mathbb{P}_\mu(H_T) \coloneqq \prod_{t=1}^T \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}^\lambda_\mu(\lambda_t \,|\, G_t, A_t) \ .$$

If P and Q are two distributions defined on the same space, let $D_{\mathrm{KL}}(P \,\|\, Q)$ and $\delta(P, Q)$ be the KL-divergence and the total variation distance respectively between P and Q. Furthermore, let $d(p \,\|\, q)$ be the KL-divergence between two Bernoulli random variables with means p and q. The following lemma provides an expression for the KL-divergence between two the probability distributions associated to two environments.

Lemma A.4. *For a fixed policy, let μ and μ' be two environments as described above. Then,*

$$D_{\mathrm{KL}}(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) = \frac{1}{M} \sum_{i=1}^M \sum_{a \in V} N_\mu(a; T) d\big(\mu(a; G^i) \,\|\, \mu'(a; G^i)\big) \ ,$$

where $N_\mu(a; T) \coloneqq \mathbb{E}_\mu\Big[\sum_{t=1}^T \mathbb{I}\{A_t = a\}\Big]$.

Proof. The proof is similar to that of Lemma 15.1 in Lattimore and Szepesvári (2020). Namely, we have in our case that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_\mu \,\|\, \mathbb{P}_{\mu'}) &= \mathbb{E}_\mu\left[\ln \frac{\mathbb{P}_\mu(H_T)}{\mathbb{P}_{\mu'}(H_T)}\right] \\
&= \mathbb{E}_\mu\left[\ln \frac{\prod_{t=1}^T \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}^\lambda_\mu(\lambda_t \,|\, G_t, A_t)}{\prod_{t=1}^T \pi_t(A_t \,|\, H_{t-1}) U_{\mathcal{G}}(G_t) \mathbb{P}^\lambda_{\mu'}(\lambda_t \,|\, G_t, A_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\ln \frac{\mathbb{P}^\lambda_\mu(\lambda_t \,|\, G_t, A_t)}{\mathbb{P}^\lambda_{\mu'}(\lambda_t \,|\, G_t, A_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\mathbb{E}_\mu\left[\ln \frac{\mathbb{P}^\lambda_\mu(\lambda_t \,|\, G_t, A_t)}{\mathbb{P}^\lambda_{\mu'}(\lambda_t \,|\, G_t, A_t)} \,\middle|\, G_t, A_t\right] \,\middle|\, A_t\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\mathbb{E}_\mu\left[D_{\mathrm{KL}}(\mathbb{P}^\lambda_\mu(\cdot \,|\, G_t, A_t) \,\|\, \mathbb{P}^\lambda_{\mu'}(\cdot \,|\, G_t, A_t)) \,\middle|\, A_t\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}_\mu\left[\sum_{i=1}^M U_{\mathcal{G}}(G^i) D_{\mathrm{KL}}(\mathbb{P}^\lambda_\mu(\cdot \,|\, G^i, A_t) \,\|\, \mathbb{P}^\lambda_{\mu'}(\cdot \,|\, G^i, A_t))\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{T} \mathbb{E}_\mu \Big[D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot \mid G^i, A_t) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \mid G^i, A_t)) \Big] \\
&= \frac{1}{M} \sum_{i=1}^{M} \sum_{a \in V} N_\mu(a; T) D_{\mathrm{KL}}(\mathbb{P}_\mu^\lambda(\cdot \mid G^i, a) \,\|\, \mathbb{P}_{\mu'}^\lambda(\cdot \mid G^i, a)) \\
&= \frac{1}{M} \sum_{i=1}^{M} \sum_{a \in V} N_\mu(a; T) d\big(\mu(a; G^i) \,\|\, \mu'(a; G^i)\big) \ ,
\end{aligned}
$$

where the last equality holds via (A.6). □

The following standard lemma, adapted from Lattimore and Szepesvári (2020), will be used in the sequel.

Lemma A.5. *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$. Let $a < b$ and $X : \Omega \to [a, b]$ be an $\mathcal{F}$-measurable random variable. Then,*

$$
\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b - a) \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \,\|\, Q)} \ .
$$

Proof. We have, by Exercise 14.4 in Lattimore and Szepesvári (2020), that

$$
\left| \int_\Omega X(\omega) dP(\omega) - \int_\Omega X(\omega) dQ(\omega) \right| \leq (b - a) \delta(P, Q) \ ,
$$

from which the lemma follows by applying Pinsker's inequality. □

II. Choosing the environments

We will construct a collection of environments $\{\mu_a\}_{a \in V}$, each associated to an action, such that for any $i \in [M]$ and $j \in [\alpha]$,

$$
\mu_a(j; G^i) \coloneqq \frac{1}{2} - \varepsilon \mathbb{I}\{a(i) = j\} \ ,
$$

where $0 < \varepsilon \leq \frac{1}{4}$ will be tuned later. In words, for a fixed graph, environment μ_a gives a slight advantage to actions that are connected to a in that graph, and thus agree with a in the corresponding game. Additionally, for every $a \in V$ and $i \in [M]$, we define environment μ_a^{-i} to be such that for any $s \in [M]$ and $j \in [\alpha]$,

$$
\mu_a^{-i}(j; G^s) \coloneqq \begin{cases} \frac{1}{2}, & \text{if } s = i \\ \mu_a(j; G^s), & \text{otherwise.} \end{cases}
$$

Similar to Eldowa et al. (2023a), we will define, for every $i \in [M]$, an equivalence relation $\sim_i$ on the arms such that

$$
a \sim_i a' \iff \forall s \in [M] \setminus \{i\}, a'(s) = a(s) \ ,
$$

for any $a, a' \in V$. This means that two arms are equivalent according to $\sim_i$ if and only if their choices of base actions coincide in all games that are different from i. Let $V/\sim_i$ be the set of equivalence classes of $\sim_i$. It is easy to see that $V/\sim_i$ contains exactly α^{M-1} equivalence classes, and that each class consists of α actions, each corresponding to a different choice of base action in game i. Notice then that for an equivalence class $W \in V/\sim_i$, all environments μ_a^{-i} with $a \in V$ are indeed identical. In the sequel, this environment will also be referred to as μ_W^{-i}.

III. Lower-bounding the regret of a single environment

Note that in environment μ_a, we have that $a = \arg\min_{a' \in V} \sum_{i=1}^M \mu_a(a'; G^i)$. Consequently, starting from (A.5) we get that

$$\begin{aligned}
\overline{R}_T(\mu_a) &= \sum_{i=1}^M \frac{1}{M} \mathbb{E}_{\mu_a}\left[\sum_{t=1}^T (\mu_a(A_t; G^i) - \mu_a(a; G^i))\right] \\
&= \sum_{i=1}^M \frac{1}{M} \mathbb{E}_{\mu_a}\left[\sum_{t=1}^T \left(\frac{1}{2} - \varepsilon \mathbb{I}\{A_t(i) = a(i)\} - \left(\frac{1}{2} - \varepsilon\right)\right)\right] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^M \mathbb{E}_{\mu_a}\left[\sum_{t=1}^T (1 - \mathbb{I}\{A_t(i) = a(i)\})\right] \\
&= \frac{\varepsilon}{M} \sum_{i=1}^M \big(T - N_{\mu_a}(i, a; T)\big) ,
\end{aligned}$$

where for environment μ, action a, and game i, $N_\mu(i, a; T) := \mathbb{E}_\mu\big[\sum_{t=1}^T \mathbb{I}\{A_t(i) = a(i)\}\big]$ is the expected number of times in environment μ that the action chosen by the policy agrees with action a in game i. Next, we use Lemma A.5 to obtain that

$$\overline{R}_T(\mu_a) \geq \frac{\varepsilon}{M} \sum_{i=1}^M \left(T - N_{\mu_a^{-i}}(i, a; T) - T\sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(\mathbb{P}_{\mu_a^{-i}} \,\|\, \mathbb{P}_{\mu_a}\big)}\right) . \tag{A.7}$$

For bounding the KL-divergence term, we start from Lemma A.4:

$$\begin{aligned}
D_{\mathrm{KL}}\big(\mathbb{P}_{\mu_a^{-i}} \,\|\, \mathbb{P}_{\mu_a}\big) &= \frac{1}{M} \sum_{s=1}^M \sum_{a' \in V} N_{\mu_a^{-i}}(a'; T) d\big(\mu_a^{-i}(a'; G^s) \,\|\, \mu_a(a'; G^s)\big) \\
&= \frac{1}{M} \sum_{a' \in V} N_{\mu_a^{-i}}(a'; T) d\big(\mu_a^{-i}(a'; G^i) \,\|\, \mu_a(a'; G^i)\big) \\
&= \frac{1}{M} \sum_{a' \in V} N_{\mu_a^{-i}}(a'; T) d\big(1/2 \,\|\, 1/2 - \varepsilon \mathbb{I}\{a'(i) = a(i)\}\big) \\
&= \frac{1}{M} \sum_{a' \in V} \mathbb{I}\{a'(i) = a(i)\} N_{\mu_a^{-i}}(a'; T) d\big(1/2 \,\|\, 1/2 - \varepsilon\big) \\
&\leq \frac{c\varepsilon^2}{M} \sum_{a' \in V} \mathbb{I}\{a'(i) = a(i)\} N_{\mu_a^{-i}}(a'; T) \\
&= \frac{c\varepsilon^2}{M} \sum_{a' \in V} \mathbb{I}\{a'(i) = a(i)\} \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^T \mathbb{I}\{A_t = a'\} \\
&= \frac{c\varepsilon^2}{M} \mathbb{E}_{\mu_a^{-i}} \sum_{t=1}^T \mathbb{I}\{A_t(i) = a(i)\} \\
&= \frac{c\varepsilon^2}{M} N_{\mu_a^{-i}}(i, a; T) ,
\end{aligned}$$

where the second equality holds since the two environments only differ in G^i, and the inequality holds for $\varepsilon \le \frac{1}{4}$ with $c = 8\log\frac{4}{3}$. Plugging back into (A.7) gets us that

$$\overline{R}_T(\mu_a) \ge \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M}N_{\mu_a^{-i}}(i,a;T)}\right) . \tag{A.8}$$

IV. Summing up

Fix $i \in [M]$. Notice that for $W \in V/\sim_i$,

$$\sum_{a\in W}\mathbb{I}\{A_t(i) = a(i)\} = 1$$

since each action in W corresponds to a different choice of base action in game i. Hence,

$$\begin{aligned}\frac{1}{\alpha^M}\sum_{a\in V}N_{\mu_a^{-i}}(i,a;T) &= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W}N_{\mu_a^{-i}}(i,a;T)\\ &= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\sum_{a\in W}N_{\mu_W^{-i}}(i,a;T)\\ &= \frac{1}{\alpha^M}\sum_{W\in V/\sim_i}\mathbb{E}_{\mu_W^{-i}}\left[\sum_{t=1}^{T}\sum_{a\in W}\mathbb{I}\{A_t(i) = a(i)\}\right]\\ &= \frac{1}{\alpha^M}\alpha^{M-1}T = \frac{T}{\alpha} .\end{aligned}$$

Using this together with (A.8) allows us to conclude that

$$\begin{aligned}\sup_\mu \overline{R}_T(\mu) &\ge \frac{1}{\alpha^M}\sum_{a\in V}\overline{R}_T(\mu_a)\\ &\ge \frac{1}{\alpha^M}\sum_{a\in V}\frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M}N_{\mu_a^{-i}}(i,a;T)}\right)\\ &\ge \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - \frac{1}{\alpha^M}\sum_{a\in V}N_{\mu_a^{-i}}(i,a;T) - \varepsilon T\sqrt{\frac{c}{2M\alpha^M}\sum_{a\in V}N_{\mu_a^{-i}}(i,a;T)}\right)\\ &= \frac{\varepsilon}{M}\sum_{i=1}^{M}\left(T - \frac{T}{\alpha} - \varepsilon T\sqrt{\frac{cT}{2M\alpha}}\right)\\ &= \varepsilon T\left(1 - \frac{1}{\alpha} - \varepsilon\sqrt{\frac{cT}{2M\alpha}}\right)\\ &\ge \varepsilon T\left(\frac{1}{2} - \varepsilon\sqrt{\frac{cT}{2M\alpha}}\right) ,\end{aligned}$$

where the third inequality holds due to the concavity of the square root, and the last inequality holds by our assumption that $\alpha \ge 2$. Setting $\varepsilon = \frac{1}{4}\sqrt{\frac{2M\alpha}{cT}}$ yields that

$$\sup_\mu \overline{R}_T(\mu) \ge \frac{1}{16}\sqrt{\frac{2}{c}}\cdot\sqrt{\alpha TM} \ge \frac{1}{18}\sqrt{\alpha TM} = \frac{1}{18}\sqrt{\alpha T\log_\alpha K} ,$$

whereas it holds that $\varepsilon \leq \frac{1}{4}$ thanks to the assumption made on T in the statement of the theorem.

V. The case when $\log_\alpha K$ is not an integer

If M is not an integer,[†] we can use the same construction as before for the first $\alpha^{\lfloor M \rfloor}$ actions and force the remaining actions to behave identically to some action in the construction. That is, we can designate a certain action such that, in all environments, all the excess actions receive the same loss as this action and are connected to it, to each other, and to every action that happens to share an edge with this designated action in a given graph (in other words, we are expanding the designated action into a clique). This way, the independence number of all the graphs in the construction is still α, and the excess actions do not provide any extra utility to the learner; playing one of them is exactly like playing the designated action, and the construction does not hide this from the player. We can then obtain the same bound as before but in terms of $\lfloor M \rfloor$, thus costing us an extra $1/\sqrt{2}$ factor to recover the desired bound (using that $\lfloor M \rfloor \geq M/2$). □

A.5 Comparison with Recent Work

In Chen et al. (2024), the authors consider a special case of the undirected feedback graph problem where the graph (fixed and known) is composed of α disjoint cliques with self-loops. For $j \in [\alpha]$, let m_j denote the number of actions in the j-th clique, implying that $\sum_{j=1}^{\alpha} m_j = K$ (the number of arms). For this problem, Chen et al. (2024, Theorem 4) provides a lower bound of order $\sqrt{T \sum_{j=1}^{\alpha} \ln(m_j + 1)}$. In particular, if the cliques are balanced (i.e., $m_1 = \cdots = m_\alpha = K/\alpha$), the lower bound becomes of order $\sqrt{\alpha T \ln(1 + K/\alpha)}$, thus matching the regret bound of Algorithm 3.1. This means that, for any value of $1 \leq \alpha \leq K$, there are feedback graphs on K nodes with independence number α such that no other algorithm can achieve a better minimax regret guarantee than that of our proposed algorithm.

We emphasize that this does not imply *graph-specific* minimax optimality. Indeed, as shown in Chen et al. (2024), when the cliques are unbalanced, the regret guarantee of our algorithm can be inferior to that of the algorithm they proposed, which matches the $\sqrt{T \sum_{j=1}^{\alpha} \ln(m_j + 1)}$ bound. However, beyond the disjoint cliques case, their algorithm requires computing a minimum clique cover for the given feedback graph G, which is known to be NP-hard (Karp, 1972). More importantly, their reliance on a clique cover leads to a dependence of the regret on the clique cover number $\theta(G)$ instead of the independence number $\alpha(G)$. One can argue that the ratio between $\theta(G)$ and $\alpha(G)$ can be $\Omega(K/(\ln K)^2)$ for most graphs on a sufficiently large number K of vertices (e.g., see Mannor and Shamir (2011, Section 6)). Finally, it is not clear how to generalize their approach to time-varying feedback graphs (informed or uninformed). Hence, despite the contributions of our work and those of Chen et al. (2024), the problem of characterizing the minimax regret rate at a graph-based granularity still calls for further investigation.

[†] Note that M is never smaller than 1 since $\alpha \leq K$.

A.6 Directed Strongly Observable Feedback Graphs

In this section, we consider the case of directed strongly observable graphs. For a directed graph $G = (V, E)$, recall that we define $N_G^{\text{in}}(i) = \{j \in V : (j, i) \in E\}$ to be the in-neighbourhood of node $i \in V$ in G, and $N_G^{\text{out}}(i) = \{j \in V : (i, j) \in E\}$ to be its out-neighbourhood. A directed graph G is strongly observable if for every $i \in V$, at least one of the following holds: $i \in N_G^{\text{in}}(i)$ or $j \in N_G^{\text{in}}(i)$ for all $j \neq i$. The independence number $\alpha(G)$ is still defined in the same manner as before; that is, the cardinality of the largest set of nodes such that no two nodes share an edge, regardless of orientation. The interaction protocol is the same as in the undirected case, except that, in each round $t \in [T]$, the learner only observes the losses of the actions in $N_{G_t}^{\text{out}}(I_t)$, which is the out-neighbourhood in graph G_t of the action I_t picked by the learner. As before, we will use $N_t^{\text{in}}(i)$ and $N_t^{\text{out}}(i)$ to denote $N_{G_t}^{\text{in}}(i)$ and $N_{G_t}^{\text{out}}(i)$ respectively. For this setting, a bound of $\mathcal{O}\big(\sqrt{\alpha T} \cdot \ln(KT)\big)$ was proven in Alon et al. (2015) for the Exp3.G algorithm. Later, Zimmert and Lattimore (2019) proved a bound of $\mathcal{O}\big(\sqrt{\alpha T \ln^3 K}\big)$ for OSMD with a variant of the q-Tsallis entropy regularizer where q was chosen as $1 - 1/(\ln K)$.

To use Algorithm 3.1 in the directed case, one can define loss estimates analogous to (3.6) by using the in-neighbourhood in place of the neighbourhood in the relevant quantities. Namely, let $S_t \coloneqq \big\{i \in V : i \notin N_t^{\text{in}}(i)\big\}$, $J_t \coloneqq \{i \in S_t : p_t(i) > 1/2\}$, and $P_t(i) \coloneqq \sum_{j \in N_t^{\text{in}}(i)} p_t(j)$. The loss estimates (again due to Zimmert and Lattimore (2019)) can then be given by

$$\widehat{\ell}_t(i) \coloneqq \begin{cases} \frac{\ell_t(i)}{P_t(i)} \mathbb{I}\left\{I_t \in N_t^{\text{in}}(i)\right\} & \text{if } i \in V \setminus J_t \\ \frac{\ell_t(i) - 1}{P_t(i)} \mathbb{I}\left\{I_t \in N_t^{\text{in}}(i)\right\} + 1 & \text{if } i \in J_t \,. \end{cases}$$

Algorithm 3.1 with these loss estimates can be analyzed in a similar manner to the proof of Theorem 3.2, with the major difference being the way that the variance term is handled for actions with self-loops. Namely, the relevant term is

$$\sum_{i \in V : i \in N_t^{\text{in}}(i)} \frac{p_t(i)^{2-q}}{\sum_{j \in N_t^{\text{in}}(i)} p_t(j)},$$

on which we elaborate more in the following.

Let $p \in \Delta_K$ and $\beta \in (0, 1/2)$ be such that $\min_{i \in V} p(i) \geq \beta$. We first consider the variance term given by the negative Shannon entropy regularizer. It is known (Alon et al., 2015) that such a variance term, restricted to nodes with a self-loop in the strongly observable feedback graph $G = (V, E)$, has an upper bound of the form

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} \leq 4\alpha(G) \ln\!\left(\frac{4K}{\alpha(G)\beta}\right) . \tag{A.9}$$

In addition to the fact that this variance bound has a linear dependence on the independence number $\alpha(G)$ of G, we observe that there is a logarithmic factor in K/α and $1/\beta$ given by the fact that we now consider directed graphs. The main problem is that, in general, we cannot hope to improve upon the above logarithmic factor as it can be shown to be unavoidable unless we manage to restrict the probability distributions we consider. Indeed, it is possible to show (Alon et al., 2017, Fact 4) that there exist probability distributions $p \in \Delta_K$ and directed strongly observable graphs G for which $\alpha(G) = 1$ and

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} = \frac{K+1}{2} = \frac{1}{2} \log_2\left(\frac{4}{\min_i p(i)}\right) = \alpha(G) \log^{\omega(1)}\left(\frac{K}{\alpha(G)}\right) .$$

A usual way to avoid this is to introduce some explicit exploration to the probability distributions in order to force a lower bound on the probabilities of all nodes, e.g., as in Exp3.G (Alon et al., 2015). This would bring the linear dependence on K down to α in the above bad case, while, on the other hand, introducing a $\ln(KT)$ factor which then worsens the overall dependence on the time horizon T.

Consider now the variance term given by the analysis of the q-FTRL algorithm. As already argued in Section 3.3, we can reuse the variance bound in (A.9) for the case of negative Shannon entropy because

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)^{2-q}}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} \le \sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)}{\sum_{j \in N_G^{\text{in}}(i)} p(j)}$$

for any $q \in (0,1)$, and such a bound is the best known so far for the general case of directed strongly observable graphs. However, we can be more clever in the way we utilize it. Similarly to the proof of Zimmert and Lattimore (2019, Theorem 14), we can gain an advantage from the adoption of q-FTRL by splitting the sum in the variance term into two sums according to some adequately chosen threshold β on the probabilities of the individual nodes. More precisely, by choosing $\beta \approx \exp\bigl(-\ln(K/\alpha)\ln K\bigr)$ and $q = 1 - 1/(\ln K)$, we can prove that

$$\sum_{i \in V : i \in N_G^{\text{in}}(i)} \frac{p(i)^{2-q}}{\sum_{j \in N_G^{\text{in}}(i)} p(j)} = \mathcal{O}\left(\alpha\left(1 + \ln\frac{K}{\alpha}\right)\ln K\right) .$$

We can further argue that, by following a similar analysis as in the proofs of Theorems 3.1 and 3.2, this variance bound would allow to show that the regret of q-FTRL is $\mathcal{O}\Bigl(\sqrt{\alpha T\bigl(1 + \ln(K/\alpha)\bigr)} \cdot \ln K\Bigr)$, where there is an additional $\ln K$ factor when compared to our regret bound in the undirected case (Theorem 3.2).

The presence of extra logarithmic factors is to be expected in the directed case, as many edges between distinct nodes might reduce the independence number of the graph, while providing information in one direction only. However, the undirected graph G' obtained from any directed strongly observable graph G by reciprocating edges between distinct nodes has the same independence number $\alpha(G') = \alpha(G)$ but the regret guarantee given by the more general analysis of q-FTRL would introduce a spurious $\ln K$ multiplicative factor. We remark that all the currently available upper bounds on the variance term (either with negative Shannon entropy or negative q-Tsallis entropy regularizers) do not exactly reflect the phenomenon of a gradually disappearing logarithmic factor when the graph is closer to being undirected (i.e., has fewer unreciprocated edges).

Taking these observations into account, we believe that it should be possible to achieve tighter guarantees that match our intuition, by improving the currently available tools. The bound on the variance term, for instance, is one part of the analysis that might be improvable. We might want to have a similar bound as (A.9) but with a sublinear dependence on α that varies according to the parameter q of the negative q-Tsallis entropy; e.g., ignoring logarithmic factors, we could expect it to become of order α^q as we managed to prove for the undirected case (Lemma 3.1). Doing so could allow a better tuning of q that might lead to improved logarithmic factors in the regret.

Appendix B

Missing Results from Chapter 4

B.1 Auxiliary Result

Lemma B.1. *Let $q \in (0,1)$, $b > 0$, $c > 1$, and $(y_t)_{t=1}^T$ be a sequence of loss vectors in $\mathbb{R}^N$ satisfying $y_t(i) \geq -b$ for all $t \in [T]$ and $i \in [N]$. Let $(p_t)_{t=1}^{T+1}$ be the predictions of FTRL with decision set Δ_N and the q-Tsallis regularizer ψ_q over this sequence of losses; that is, $p_1 := \arg\min_{p \in \Delta_N} \psi_q(p)$, and for $t \in [T]$,*

$$p_{t+1} := \operatorname*{arg\,min}_{p \in \Delta_N} \eta \sum_{s=1}^{t} \langle y_s, p \rangle + \psi_q(p)\,,$$

assuming the learning rate η satisfies $0 < \eta \leq \frac{q}{(1-q)b}\Big(1 - c^{\frac{q-1}{2-q}}\Big)$. Then for any $u \in \Delta_N$,

$$\sum_{t=1}^{T} \langle p_t - u, y_t \rangle \leq \frac{N^{1-q} - 1}{(1-q)\eta} + \frac{\eta c}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} p_t(i)^{2-q}\, y_t(i)^2\,.$$

Proof. Let $p'_{t+1} := \arg\min_{p \in \mathbb{R}^N_{\geq 0}} \langle p, y_t \rangle + D_{\psi_q}(p, p_t)$, where $D_{\psi_q}(\cdot,\cdot)$ denotes the Bregman divergence based on ψ_q. Via Lemma 7.14 in Orabona (2019) we have that

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t - u, y_t \rangle &\leq \frac{\psi_q(u) - \psi_q(p_1)}{\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2 \\
&\leq \frac{N^{1-q} - 1}{(1-q)\eta} + \frac{\eta}{2q} \sum_{t=1}^{T} \sum_{i=1}^{N} z_t(i)^{2-q}\, y_t(i)^2\,,
\end{aligned}$$

where z_t lies on the line segment between p_t and p'_{t+1}. A simple derivation shows that

$$p'_{t+1}(i) = p_t(i) \left(\frac{1}{1 + \eta \frac{1-q}{q} y_t(i) p_t(i)^{1-q}} \right)^{\frac{1}{1-q}}\,,$$

for each $i \in [N]$. On the other hand, it holds that

$$\eta \frac{1-q}{q} y_t(i) p_t(i)^{1-q} \geq -\eta \frac{1-q}{q} b p_t(i)^{1-q} \geq -\eta \frac{1-q}{q} b \geq c^{\frac{q-1}{2-q}} - 1\,,$$

where the first inequality uses that $y_t(i) \geq -b$ (and that $p_t(i), \eta > 0$), the second uses that $p_t(i) \leq 1$, and the third uses that $\eta \leq \frac{q}{(1-q)b}\left(1 - c^{\frac{q-1}{2-q}}\right)$. This entails that $p'_{t+1}(i) \leq c^{\frac{1}{2-q}} p_t(i)$, which implies that $z_t(i) \leq c^{\frac{1}{2-q}} p_t(i)$ concluding the proof. □

Appendix C

Proof Details for Chapter 5

C.1 On the Computation of the Optimal Probability Thresholds

The tasks of finding the independence number and (weak) domination number in a graph are notoriously NP-hard problems. In particular, while for the domination number, by a reduction to set cover, a simple greedy approach yields a logarithmic (in the number K of nodes) approximation (Vazirani, 2001), for the independence number it is known that even computing a $K^{1-\epsilon}$-approximation is hard, for any $\epsilon > 0$ (Håstad, 1999, Zuckerman, 2007).

Our algorithm OptimisticThenCommitGraph solves these computational aspects directly, whereas the hardness of finding α^* and δ^* may limit the applicability of EdgeCatcher in instances with a large and complex action space. In fact, the computation of the stopping function Φ involves finding the best thresholds ε_s^* and ε_w^*, defined in Equations (5.1) and (5.2), and therefore repeatedly solving NP-hard problems. In what follows, we present some observations that clarify to which extent (and at which cost) EdgeCatcher can still be implemented efficiently.

First, it is important to note that our algorithm is robust with respect to approximate knowledge of the topological parameters: the definition of Φ can be tweaked as to consider the approximation factor at the cost of having the same factor showing up in the regret bound (with the same order as the approximated graph parameter). This partly solves the problem for weakly observable graphs (as the $(\log(K) + 1)$-approximation only gives and extra $\mathrm{polylog}(K)$ in the regret) and for the classes of graphs where it is possible to efficiently compute good approximations of the independence number, e.g., planar graphs (Baker, 1994) or bounded-degree graphs (Halldórsson and Radhakrishnan, 1997).

Another approach consists in considering the fractional solutions of the independence and domination number linear programs. While for the former we obtain an approximation given by the integrality gap, for the latter we can show a tight dependence on the fractional weak domination number (thus improving the regret bound), as in Chen et al. (2021).

Furthermore, note that it is always possible to ignore the α and δ terms in the definition of Φ; it is not hard to see that such an approach yields a regret bound (ignoring polylog terms) of the type $\min\{\sqrt{(K/\varepsilon_1)T}, (K/\varepsilon_2)^{1/3}T^{2/3}\}$, where ε_1, respectively ε_2, is the largest ε such that $\mathrm{supp}\left([\mathcal{G}]_\varepsilon\right)$ is strongly, respectively weakly, observable. Although suboptimal, this drastic approach gives a regret bound with an optimal dependence on the T and ε terms (as $\varepsilon_s^* \le \varepsilon_1$ and $\varepsilon_w^* \le \varepsilon_2$).

Finally, we conclude by discussing how it is possible to drastically reduce the number of times that EdgeCatcher calls the routine to compute α and δ, at the cost of losing a small multiplicative factor in the regret. Crucially, we do not need to check the stopping condition involving Φ in every single round: it suffices to do so for a logarithmic number of times. Assume, in fact, to check the stopping condition in RoundRobin only when τ is a power of 2, i.e., $\tau = 2^b$ for some integer b. This single check covers all rounds τ' such that $\tau/2 = 2^{b-1} \leq \tau' \leq 2^b = \tau$. On the stochastic graph estimate $\widehat{\mathcal{G}}_\tau$ we can compute $\alpha_{\varepsilon_\tau}/\varepsilon_\tau$ and $\delta_{\varepsilon_\tau}/\varepsilon_\tau$, which are also 2-approximations for the best respective ratios on any thresholded graph corresponding to rounds of RoundRobin between $\tau/2$ and τ (note that such an approach would also improve the dependency of ε_τ and Δ on T in Theorems 5.2 and 5.3, and thus in the regret bound, from $\log(T)$ down to $\log(\log(T))$ due to an improved union bound).

C.2 Missing Results from Section 5.3

C.2.1 Proof of Theorem 5.2

Theorem 5.2. *If* RoundRobin *(Algorithm 5.1) is run on the stochastic feedback graph $\mathcal{G}$, then, with probability at least $1 - 1/T$, the estimate $\widehat{\mathcal{G}}_\tau$ is an ε_τ-good approximation of $\mathcal{G}$ simultaneously for all $\tau \leq \widehat{\tau}$, where $\widehat{\tau} \leq T/K$ is the index of the last iteration of the outer for-loop in Algorithm 5.1.*

Proof of Theorem 5.2. For all edges e and time steps $\tau \leq \widehat{\tau}$, we define the following two events: the event $\mathcal{E}_e^\tau \coloneqq \{\widehat{p}_e^\tau \geq \varepsilon_\tau\}$ that e belongs to the support of $\widehat{\mathcal{G}}_\tau$, and the event $\mathcal{F}_e^\tau \coloneqq \{|\widehat{p}_e^\tau - p_e| \leq p_e/2\}$ that $\widehat{p}_e^\tau$ is well estimated. For all $\tau \leq \widehat{\tau}$, we also define large and small edges in E according to their probabilities: $E_\tau^+ \coloneqq \{e \in V^2 : p_e \geq 2\varepsilon_\tau\}$ and $E_\tau^- \coloneqq \{e \in V^2 : p_e < \varepsilon_\tau/2\}$.

First, we look at the complementary event of $\mathcal{E}_e^\tau$ for any $\tau \leq \widehat{\tau}$ and $e \in E_\tau^+$. We have:

$$\mathbb{P}\left(\overline{\mathcal{E}_e^\tau}\right) = \mathbb{P}\left(\widehat{p}_e^\tau < \varepsilon_\tau\right) \leq \mathbb{P}\left(\widehat{p}_e^\tau \leq p_e/2\right) = \mathbb{P}\left(\widehat{p}_e^\tau - p_e \leq -p_e/2\right) \leq \mathrm{e}^{-\frac{\tau}{8}p_e} \leq \mathrm{e}^{-\frac{\tau}{4}\varepsilon_\tau} \leq \frac{1}{4KT^2} .$$

Note that in the first and second to last inequalities we used the fact that $p_e \geq 2\varepsilon_\tau$, in the last inequality the definition of ε_τ and the fact that $K \geq 2$, while in the second inequality we applied the Chernoff lower bound (multiplicative version, see Mitzenmacher and Upfal (2005, part 2 of Theorem 4.5)) on the estimator $\widehat{p}_e^\tau$.

If we call $\mathcal{E}$ the event corresponding to part 1 of Definition 5.1, we have the following:

$$\mathbb{P}\left(\mathcal{E}\right) = \mathbb{P}\left(\bigcap_{\tau \leq \widehat{\tau}} \bigcap_{e \in E_\tau^+} \mathcal{E}_e^\tau\right) \geq 1 - \sum_{\tau \leq \widehat{\tau}} \sum_{e \in E_\tau^+} \mathbb{P}\left(\widehat{p}_e^\tau < \varepsilon_\tau\right) \geq 1 - \sum_{\tau \leq \widehat{\tau}} \frac{|E_\tau^+|}{4KT^2} \geq 1 - \frac{1}{4T} , \tag{C.1}$$

where we used that $|E_\tau^+| \leq K^2$ for all $\tau \leq \widehat{\tau} \leq T/K$ with probability 1.

Next, we study the complementary event of $\mathcal{F}_e^\tau$ for $e \notin E_\tau^-$. For such e and any $\tau \leq \widehat{\tau}$, we can directly use the two-sided Chernoff bound (multiplicative version, as in Mitzenmacher and Upfal (2005, Corollary 4.6)) on the estimator $\widehat{p}_e^\tau$:

$$\mathbb{P}\left(\overline{\mathcal{F}_e^\tau}\right) = \mathbb{P}\left(|\widehat{p}_e^\tau - p_e| > \frac{1}{2}p_e\right) \leq 2\mathrm{e}^{-\frac{\tau}{12}p_e} \leq 2\mathrm{e}^{-\frac{\tau}{24}\varepsilon_\tau} \leq \frac{1}{2KT^2} .$$

Note that we used the definition of ε_τ and the facts that $2p_e \geq \varepsilon_\tau$ and $K, T \geq 2$. Now, if we call $\mathcal{F}$ the event corresponding to part 2 of Definition 5.1, we can proceed via union bounding as in

Equation (C.1) and get

$$\mathbb{P}\left(\mathcal{F}\right) = \mathbb{P}\left(\bigcap_{\tau\le\widehat{\tau}}\bigcap_{e\notin E_\tau^-}\mathcal{F}_e^\tau\right) \ge 1-\frac{1}{2T}\ . \tag{C.2}$$

As a third step, we get back to the $\mathcal{E}_e^\tau$ events, but we consider $e\in E_\tau^-$. For $\tau\le\widehat{\tau}$ and $e\in E_\tau^-$ we have:

$$\mathbb{P}\left(\mathcal{E}_e^\tau\right) = \mathbb{P}\left(\widehat{p}_e^\tau \ge \varepsilon_\tau\right) \le \mathbb{P}\left(\widehat{p}_e^\tau - p_e \ge \frac{1}{2}\varepsilon_\tau\right) = \mathbb{P}\left(\widehat{p}_e^\tau - p_e \ge xp_e\right),$$

where we used $p_e<\varepsilon_\tau/2$ and named $x=\varepsilon_\tau/(2p_e)>1$. At this point we can use the Chernoff upper bound (multiplicative version, see Mitzenmacher and Upfal (2005, part 1 of Theorem 4.4) with $\delta=x$) and obtain:

$$\mathbb{P}\left(\mathcal{E}_e^\tau\right) \le \mathbb{P}\left(\widehat{p}_e^\tau - p_e \ge xp_e\right) \le \left(\frac{\mathrm{e}^x}{(1+x)^{1+x}}\right)^{\tau p_e} \le \mathrm{e}^{-\frac{\tau}{3}xp_e} = \mathrm{e}^{-\frac{\tau}{6}\varepsilon_\tau} \le \frac{1}{4KT^2}\ .$$

The third inequality follows from $2x/(2+x)\le\ln(1+x)$ which holds for all positive x:

$$\frac{\mathrm{e}^x}{(1+x)^{1+x}} = \mathrm{e}^{x-(1+x)\ln(1+x)} \le \mathrm{e}^{-x^2/(2+x)} \le \mathrm{e}^{-x/3}, \quad \forall x\ge 1\ .$$

If we now call $\mathcal{C}$ the event described in part 3 of Definition 5.1, we get, using the bound on $\mathbb{P}\left(\mathcal{E}_e^\tau\right)$ and a union bound as in Equations (C.1) and (C.2):

$$\mathbb{P}\left(\mathcal{C}\right) = \mathbb{P}\left(\bigcap_{\tau\le\widehat{\tau}}\bigcap_{e\in E_\tau^-}\overline{\mathcal{E}_e^\tau}\right) \ge 1-\frac{1}{4T}\ . \tag{C.3}$$

The theorem then follows by a union bound on the complementary events of $\mathcal{E},\mathcal{F}$ and $\mathcal{C}$. □

C.2.2 Proof of Theorem 5.3

In order to prove the regret bound achieved by BlockReduction, we need to show that it is able to compute unbiased estimators for the average loss of observed actions within each time block. This property is guaranteed as long as the learner plays consistently a same action within each time block, and conditioned on the event that each action in the support out-neighborhood of the chosen action is observed at least once in the respective time block (depending on the realizations of the feedback graph).

Lemma C.1. *Let $G=\operatorname{supp}(\mathcal{G})$ and c_τ and $\widehat{c}_\tau$ defined as in Equations* (5.3) *and* (5.4)*. For each block B_τ, if the learner plays consistently action a, then for each $a'\in N_G^{\text{out}}(a)$ the estimators $\widehat{c}_\tau(a')$ are unbiased under $\mathcal{E}_{(a,a')}^\tau$:*

$$\mathbb{E}\left[\widehat{c}_\tau(a')\,\middle|\,\mathcal{E}_{(a,a')}^\tau\right] = c_\tau(a')\ , \quad \forall a'\in N_G^{\text{out}}(a)\ .$$

Proof. Recall that $\mathcal{E}_{(a,a')}^\tau$ is the event that the edge (a,a') in G is observed at least once in block B_τ. Substituting the definition (5.4) of the estimator, we can write

$$\mathbb{E}\left[\widehat{c}_\tau(a')\,\middle|\,\mathcal{E}_{(a,a')}^\tau\right] = \sum_{t\in B_\tau}\ell_t(a')\mathbb{E}\left[\frac{\mathbb{I}\{(a,a')\in E_t\}}{\Delta_{(a,a')}^\tau}\,\middle|\,\mathcal{E}_{(a,a')}^\tau\right]\ .$$

Now we just need to prove that the expectation in the right-hand side is equal to $1/\Delta$:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{I}\{(a,a') \in E_t\}}{\Delta^\tau_{(a,a')}} \,\middle|\, \mathcal{E}^\tau_{(a,a')}\right] &= \sum_{r=1}^{\Delta} \mathbb{E}\left[\frac{\mathbb{I}\{(a,a') \in E_t\}}{r} \,\middle|\, \Delta^\tau_{(a,a')} = r\right] \mathbb{P}\left(\Delta^\tau_{(a,a')} = r \,\middle|\, \mathcal{E}^\tau_{(a,a')}\right) \\
&= \sum_{r=1}^{\Delta} \frac{1}{r} \mathbb{P}\left((a,a') \in E_t \,\middle|\, \Delta^\tau_{(a,a')} = r\right) \mathbb{P}\left(\Delta^\tau_{(a,a')} = r \,\middle|\, \mathcal{E}^\tau_{(a,a')}\right) \\
&= \frac{1}{\Delta} \sum_{r=1}^{\Delta} \mathbb{P}\left(\Delta^\tau_{(a,a')} = r \,\middle|\, \mathcal{E}^\tau_{(a,a')}\right) = \frac{1}{\Delta} \;.
\end{aligned}$$

Note that in the third equality we used the fact that, conditioned on $\Delta^\tau_{(a,a')} = r > 0$, the r time steps when $(a,a') \in E_t$ are distributed uniformly at random in the Δ time steps. □

We can now prove the regret bound of BlockReduction in Theorem 5.3, which we restate below. Its regret depends on the performance of the algorithm $\mathcal{A}$ used on the meta-instance derived from the blocks reduction.

Theorem 5.3. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$, and let $\widehat{\mathcal{G}}$ be an ε-good approximation of $\mathcal{G}$. Let $\mathcal{A}$ be an algorithm for online learning with arbitrary deterministic feedback graph G with regret bound $R_N^{\mathcal{A}}(G)$ over any sequence of N losses in $[0,1]$. Then, the regret of* BlockReduction *(Algorithm 5.2) run with input $(T, \varepsilon/2, \widehat{\mathcal{G}}, \mathcal{A})$ is at most $\Delta R_N^{\mathcal{A}}\big(\mathrm{supp}(\widehat{\mathcal{G}})\big) + \Delta$, where $N := \lfloor T/\Delta \rfloor$ and $\Delta := \lceil \frac{4}{\varepsilon} \ln(KT) \rceil$.*

Proof of Theorem 5.3. Consider the partition of the T time steps into N blocks $B_1, \dots, B_N$ of equal size Δ and let $\mathcal{E}$ be the clean event, corresponding to all edges e in the graph $\widehat{G} := \mathrm{supp}\big(\widehat{\mathcal{G}}\big)$ being realized at least once in each block. Formally, denoting $G = (V, E)$, we let $\mathcal{E} := \bigcap_{\tau=1}^N \bigcap_{e \in E} \mathcal{E}^\tau_e$, where $\mathcal{E}^\tau_e$ are defined as in the proof of Lemma C.1. By Definition 5.1 (part 3), all the edges $e \in E$ have a probability p_e in $\mathcal{G}$ that is at least $\varepsilon/2$. Thus, it is immediate to verify that

$$\mathbb{P}\left(\mathcal{E}^\tau_e\right) = 1 - (1-p_e)^\Delta \geq 1 - \left(1 - \frac{\varepsilon}{2}\right)^\Delta \geq 1 - \mathrm{e}^{-\varepsilon\Delta/2} \geq 1 - \frac{1}{K^2T^2}$$

holds for any edge $e \in E$ using our choice of Δ. We show by union bound that the probability any of these edges never realizes in some block is

$$\mathbb{P}\left(\bigcup_{\tau \leq N} \bigcup_{e \in E} \overline{\mathcal{E}}^\tau_e\right) \leq \sum_{\tau \leq N} \sum_{e \in E} \mathbb{P}\left(\overline{\mathcal{E}}^\tau_e\right) \leq \frac{1}{T} \;,$$

where we used that there are at most K^2 directed edges (including self-loops) in $\widehat{G}$ and we substituted the chosen values of N and Δ.

We can then bound the overall regret R_T as follows:

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^T \ell_t(I_t) \,\middle|\, \mathcal{E}\right] - \min_k \sum_{t=1}^T \ell_t(k) + T \cdot \mathbb{P}\left(\overline{\mathcal{E}}\right) + (T - \Delta N) \;. \tag{C.4}$$

Note that the final term is an upper bound to the regret in the final time steps of the algorithm. We just showed that $\mathbb{P}\left(\overline{\mathcal{E}}\right)$ is smaller than $1/T$. This, together with the fact that $T - \Delta N$ is at most $\Delta - 1$, gives the additive Δ we have in the final statement.

We now focus on the remaining term, which corresponds to the regret conditioned on $\mathcal{E}$. It is equal to

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\ell_t(I_t)\,\middle|\,\mathcal{E}\right]-\min_k\sum_{t=1}^{T}\ell_t(k)&=\Delta\cdot\left(\mathbb{E}\left[\sum_{\tau=1}^{N}\sum_{t\in B_\tau}\frac{\ell_t(I_\tau)}{\Delta}\,\middle|\,\mathcal{E}\right]-\min_k\sum_{\tau=1}^{N}\sum_{t\in B_\tau}\frac{\ell_t(k)}{\Delta}\right)\\&=\Delta\cdot\left(\mathbb{E}\left[\sum_{\tau=1}^{N}c_\tau(I_\tau)\,\middle|\,\mathcal{E}\right]-\min_k\sum_{\tau=1}^{N}c_\tau(k)\right)\ ,\end{aligned}\tag{C.5}$$

where, we recall it, $c_\tau(i)$ is the average loss of action i in block B_τ. Indeed, our algorithm chooses the same action $I_t \coloneqq I_\tau$ for all time steps $t\in B_\tau$, and the decision is based on algorithm $\mathcal{A}$.

Consider now the loss estimates $\widehat{c}_1,\dots,\widehat{c}_N$ that we provide to algorithm $\mathcal{A}$. These estimates are such that $\mathbb{E}\left[\widehat{c}_\tau(i)\,|\,\mathcal{E}\right]=c_\tau(i)$ by Lemma C.1. Note that conditioning on $\mathcal{E}$ instead that on the single $\mathcal{E}_e^\tau$ does not affect the fact that the estimators are unbiased: this is due to the fact that the edge realizations are independent from the losses and the strategy of the learner.

Therefore, letting k^* be the action minimizing $c_1(k)+\cdots+c_T(k)$ over $k=1,\dots,K$,

$$\mathbb{E}\left[\sum_{\tau=1}^{N}c_\tau(I_\tau)\,\middle|\,\mathcal{E}\right]-\min_k\sum_{\tau=1}^{N}c_\tau(k)=\mathbb{E}\left[\sum_{\tau=1}^{N}\widehat{c}_\tau(I_\tau)-\sum_{\tau=1}^{N}\widehat{c}_\tau(k^*)\,\middle|\,\mathcal{E}\right]\le R_N^{\mathcal{A}}(\widehat{G})\ ,\tag{C.6}$$

where $R_N^{\mathcal{A}}(\widehat{G})$ is the regret bound of algorithm $\mathcal{A}$ given losses $\widehat{c}_1,\dots,\widehat{c}_N$ and feedback graph $\widehat{G}=\operatorname{supp}\bigl(\widehat{\mathcal{G}}\bigr)$. Finally, substituting Equations (C.5) and (C.6) into Equation (C.4) yields the desired bound. □

C.2.3 Proof of Corollary 5.1

Corollary 5.1. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$, and let $\widehat{\mathcal{G}}$ be an ε-good approximation of $\mathcal{G}$ for $\varepsilon\ge 1/T$ and with support $\widehat{G}$. The following statements hold:*

- *If $\widehat{G}$ is strongly observable with independence number α, then the regret of* BlockReduction *run with parameter $\varepsilon/2$ using* Exp3.G *for strongly observable graphs as base algorithm $\mathcal{A}$ satisfies: $R_T\le 4C_s\sqrt{(\alpha/\varepsilon)T}\cdot\ln^{3/2}(KT)$, where $C_s>0$ is a constant in the regret bound of $\mathcal{A}$.*
- *If $\widehat{G}$ is (weakly) observable with weak domination number δ, then the regret of* BlockReduction *run with parameter $\varepsilon/2$ using* Exp3.G *for weakly observable graphs as base algorithm $\mathcal{A}$ satisfies: $R_T\le 4C_w(\delta/\varepsilon)^{1/3}T^{2/3}\ln^{2/3}(KT)$, where $C_w>0$ is a constant in the regret bound of $\mathcal{A}$.*

Proof of Corollary 5.1. The statement follows from Theorem 5.3, the assumption on ε (which lets us safely handle the additive Δ term), and the fact that Exp3.G achieves regret $R_N^{\mathcal{A}}\le C_s\sqrt{\alpha N}\ln(KN)$ on strongly observable graphs, and regret $R_N^{\mathcal{A}}\le C_w(\delta\ln K)^{1/3}N^{2/3}$ on (weakly) observable graphs. □

C.2.4 Proof of Theorem 5.4

To prove Theorem 5.4 we first need two preliminary lemmata. In Lemma C.2 we present some generic properties of the stopping function $\Phi(\mathcal{G},T)$, while in Lemma C.3 we prove that $\Phi(\mathcal{G},T-\widehat{\tau}K)$ is indeed the regret obtained in BlockReduction after the stopping condition in RoundRobin is triggered.

Lemma C.2. *Let $\mathcal{G}$ be a stochastic feedback graph such that $\Phi(\mathcal{G}, T) \neq \infty$, and let ε^* be the threshold where the* arg min *in the definition of $\Phi(\mathcal{G}, T)$ is attained. Consider a run of the algorithm* EdgeCatcher *where* RoundRobin *does not fail while using the stopping function Φ defined in Equation* (5.5)*. We have the following:*

(i) $\Phi(\widehat{\mathcal{G}}_{\tau'}, T) \leq 2\Phi(\widehat{\mathcal{G}}_\tau, T)$, for all τ, τ' such that $\tau \leq \tau' \leq \widehat{\tau}$,

(ii) $\Phi(\widehat{\mathcal{G}}_\tau, T) \leq \sqrt{2}\Phi(\mathcal{G}, T)$ for all τ such that $120\ln(KT)/\varepsilon^* \leq \tau \leq \widehat{\tau}$ (if such τ exists),

where $\widehat{\tau} \leq \lfloor T/K \rfloor$ is the index of the last iteration of the outer for loop in Algorithm 5.1.

Proof. We consider a run of EdgeCatcher where RoundRobin does not fail. This means that all the $\widehat{\mathcal{G}}_\tau$ are ε_τ-good approximation of $\mathcal{G}$, for all $\tau \leq \widehat{\tau}$. Focus on the first part of the statement. All edges in $\mathrm{supp}\big(\widehat{\mathcal{G}}_\tau\big)$ are contained in $\mathrm{supp}\big(\widehat{\mathcal{G}}_{\tau'}\big)$ since RoundRobin does not fail. This implies that the observability regime only improves as τ increases. We have two cases: if the best threshold for $\widehat{\mathcal{G}}_\tau$ (say it corresponds to some edge probability in $\widehat{\mathcal{G}}_\tau$ without loss of generality) induces a thresholded stochastic feedback graph with strongly observable support $G = (V, E)$ and independence number α, we have that $\widehat{\mathcal{G}}_{\tau'}$ is strongly observable too; moreover, all the edges $e \in E$ are such that $|p_e - \widehat{p}_e^\tau| \leq p_e/2$ by Definition 5.1 (part 2); the same holds for τ': $|p_e - \widehat{p}_e^{\tau'}| \leq p_e/2$. Consider graph G with edge probabilities $\widehat{p}_e^{\tau'}$, respectively p_e and $\widehat{p}_e^\tau$ and let ε_1, respectively ε_2 and ε_3, be their smallest probability (restricting on the edges of G). We have that:

$$\min_{\varepsilon \in (0,1]} \left\{ \frac{\alpha\big([\widehat{\mathcal{G}}_{\tau'}]_\varepsilon\big)}{\varepsilon} : \mathrm{supp}\big([\widehat{\mathcal{G}}_{\tau'}]_\varepsilon\big) \text{ strongly observable} \right\} \leq \frac{\alpha}{\varepsilon_1} \leq 2\frac{\alpha}{\varepsilon_2} \leq 4\frac{\alpha}{\varepsilon_3}$$

$$= 4 \min_{\varepsilon \in (0,1]} \left\{ \frac{\alpha\big([\widehat{\mathcal{G}}_\tau]_\varepsilon\big)}{\varepsilon} : \mathrm{supp}\big([\widehat{\mathcal{G}}_\tau]_\varepsilon\big) \text{ strongly observable} \right\} ,$$

where the first inequality follows from suboptimality of graph G with threshold ε_1 for $\widehat{\mathcal{G}}_{\tau'}$, the second and the third inequality by the conditions on p_e, $\widehat{p}_e^{\tau'}$ and p_e^τ, and the last equality by definition of G and α. If we now substitute this inequality in the definition of Φ, we obtain that $2\Phi(\widehat{\mathcal{G}}_\tau, T) \geq \Phi(\widehat{\mathcal{G}}_{\tau'}, T)$. We can reason in the same exact way considering the (weakly) observable case and obtain $\sqrt[3]{4}\Phi(\widehat{\mathcal{G}}_\tau, T) \geq \Phi(\widehat{\mathcal{G}}_{\tau'}, T)$. Putting the two results together we conclude the proof of point (i).

We move our attention to the second part of the lemma. Because of Theorem 5.2 together with the lower bound on τ, it holds that $\widehat{\mathcal{G}}_\tau$ is an $\varepsilon^*/2$-good approximation of $\mathcal{G}$. This implies that all the edges in $\mathrm{supp}\left([\mathcal{G}]_{\varepsilon^*}\right)$ are contained in the support of $\widehat{\mathcal{G}}_\tau$ and that they are well approximated, as in parts 1 and 2 of Definition 5.1. We have two cases, according to the topology of the support corresponding to the threshold ε^* which guarantees the optimal regret for $\mathcal{G}$. First, consider the case that ε^* corresponds to a strongly observable structure in $\mathrm{supp}\left([\mathcal{G}]_{\varepsilon^*}\right)$ with independence number α^*; we have that

$$\min_{\varepsilon \in (0,1]} \left\{ \frac{\alpha\big([\widehat{\mathcal{G}}_\tau]_\varepsilon\big)}{\varepsilon} : \mathrm{supp}\big([\widehat{\mathcal{G}}_\tau]_\varepsilon\big) \text{ strongly observable} \right\} \leq 2\frac{\alpha^*}{\varepsilon^*}$$

$$= 2 \min_{\varepsilon \in (0,1]} \left\{ \frac{\alpha([\mathcal{G}]_\varepsilon)}{\varepsilon} : \mathrm{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ strongly observable} \right\} ,$$

where in the first inequality we used the suboptimality of threshold $\varepsilon^*/2$ for $\widehat{\mathcal{G}}_\tau$ and the fact that the independence number of $\alpha([\widehat{\mathcal{G}}_\tau]_{\varepsilon^*})$ is at most α^* (and the strong observability is maintained).

Then, we have that

$$\Phi(\widehat{\mathcal{G}}_\tau, T) \leq 4C_s\sqrt{2\frac{\alpha^*}{\varepsilon^*}T}\big(\ln(KT)\big)^{3/2} = \sqrt{2}\Phi(\mathcal{G}, T)\ ,$$

where the inequality follows naturally from the (possible) suboptimality of the choice of the strongly observable regime and the threshold $\varepsilon^*/2$ for $\widehat{\mathcal{G}}_\tau$. We can argue similarly for the case in which the optimal ε^* corresponds to the weakly observable regime in $\mathcal{G}$. In this case, for the same arguments as per the strongly observable regime, we have that

$$\min_{\varepsilon\in(0,1]}\left\{\frac{\delta([\widehat{\mathcal{G}}_\tau]_\varepsilon)}{\varepsilon} : \mathrm{supp}\big([\widehat{\mathcal{G}}_\tau]_\varepsilon\big) \text{ observable}\right\} \leq 2\frac{\delta^*}{\varepsilon^*} = 2\min_{\varepsilon\in(0,1]}\left\{\frac{\delta([\mathcal{G}]_\varepsilon)}{\varepsilon} : \mathrm{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ observable}\right\}\ .$$

Finally, similarly to the strongly observable case, it holds that

$$\Phi(\widehat{\mathcal{G}}_\tau, T) \leq 4C_w\left(2\frac{\delta^*}{\varepsilon^*}\big(\ln(KT)\big)^2\right)^{1/3}T^{2/3} = \sqrt[3]{2}\Phi(\mathcal{G}, T) \leq \sqrt{2}\Phi(\mathcal{G}, T)\ .$$

This concludes the proof. □

Lemma C.3. *Consider a run of* EdgeCatcher *(Algorithm 5.3). Assume that the invocation of* RoundRobin *returns a stochastic feedback graph $\widehat{\mathcal{G}}$ that is an $\widehat{\varepsilon}$-good approximation of $\mathcal{G}$ satisfying $\Phi(\widehat{\mathcal{G}}, T - \widehat{\tau}K) \leq \widehat{\tau}K$, where $\widehat{\tau}$ is the index of the last iteration of the outer for loop in Algorithm 5.1. Then, the regret experienced by the invocation of* BlockReduction *is at most $\Phi(\widehat{\mathcal{G}}, T - \widehat{\tau}K)$.*

Proof. Denote with $R_{T'}^{\mathrm{BR}}$ the worst-case regret experienced by BlockReduction in the final $T' = T - \widehat{\tau}K$ time steps, under the assumption on $\widehat{\mathcal{G}}$ in the statement, and let $\widehat{\varepsilon}^*$ be the best threshold as in Algorithm 5.3. We have two cases, according to $\widehat{\varepsilon}^*$ referring to strongly or (weakly) observable graphs. If $\widehat{\varepsilon}^* = \widehat{\varepsilon}^*_s$, then, by the part of Corollary 5.1 relative to strongly observable graphs, we have that

$$R_{T'}^{\mathrm{BR}} \leq 4C_s\sqrt{\frac{\widehat{\alpha}^*}{\widehat{\varepsilon}^*_s}T'}\big(\ln(KT')\big)^{3/2} = \Phi(\widehat{\mathcal{G}}_{\widehat{\varepsilon}^*}, T')\ .$$

If $\widehat{\varepsilon}^* = \widehat{\varepsilon}^*_w$, then we can apply the part of Corollary 5.1 relative to (weakly) observable graphs and obtain that

$$R_{T'}^{\mathrm{BR}} \leq 4C_w\left(\frac{\widehat{\delta}^*}{\widehat{\varepsilon}^*_w}\big(\ln(KT')\big)^2\right)^{1/3}(T')^{2/3} = \Phi(\widehat{\mathcal{G}}_{\widehat{\varepsilon}^*}, T')\ .$$

□

At this point, we have all the essential ingredients to prove the regret bound of EdgeCatcher as stated in Theorem 5.4. We rewrite the statement of Theorem 5.4 for convenience.

Theorem 5.4. *Consider the problem of online learning with stochastic feedback graph $\mathcal{G}$ on T time steps. If $\mathrm{supp}\left([\mathcal{G}]_{\varepsilon(K,T)}\right)$ is observable for $\varepsilon(K,T) := CK^3(\ln(KT))^2/T$ for a given constant $C > 0$, then there exists an algorithm whose regret R_T, ignoring logarithmic factors in K and T, satisfies $R_T \lesssim \min\left\{\sqrt{(\alpha^*/\varepsilon^*_s)T}, \left(\delta^*/\varepsilon^*_w\right)^{1/3}T^{2/3}\right\}$.*

Proof of Theorem 5.4. We condition the analysis on the clean event $\mathcal{E}$ that RoundRobin does not fail. Let $\widetilde{\varepsilon}$ be the largest ε such that $\operatorname{supp}\big([\mathcal{G}]_\varepsilon\big)$ is observable, and $\widetilde{\tau}$ be the smallest (random) integer such that $\operatorname{supp}\big(\widehat{\mathcal{G}}_{\widetilde{\tau}}\big)$ is observable for $\widehat{\mathcal{G}}_{\widetilde{\tau}}$ in RoundRobin. We have some immediate bound on these quantities. First, $\widetilde{\varepsilon} \ge \varepsilon(K,T)$, by the assumption on $\operatorname{supp}\big([\mathcal{G}]_{\varepsilon(K,T)}\big)$ being observable. Second, $\widetilde{\tau} \le \frac{120}{\widetilde{\varepsilon}}\ln(KT)$; this is due to the fact that, after $\tau = \lceil\frac{120}{\widetilde{\varepsilon}}\ln(KT)\rceil$ time steps, the estimated graph $\widehat{\mathcal{G}}_\tau$ is an $\widetilde{\varepsilon}/2$-good approximation of $\mathcal{G}$ and thus contains all the edges in $\operatorname{supp}\big([\mathcal{G}]_{\widetilde{\varepsilon}}\big)$ by Definition 5.1 (part 1) with $\varepsilon = \widetilde{\varepsilon}/2$, and because of the conditioning on $\mathcal{E}$. All in all, we can summarize these observations by noticing that

$$\frac{T}{2K} \ge 120\frac{\ln(KT)}{\varepsilon(K,T)} \ge 120\frac{\ln(KT)}{\widetilde{\varepsilon}} \ge \widetilde{\tau}\,,$$

where the first inequality is true as long as $\varepsilon(K,T) \ge 240K\ln(KT)/T$. Using point (i) of Lemma C.2 and the inequality we just showed, we observe that

$$\Phi(\widehat{\mathcal{G}}_{\lfloor\frac{T}{2K}\rfloor}, T) \le 2\Phi(\widehat{\mathcal{G}}_{\widetilde{\tau}}, T) \le 8C_w\Big(2\frac{KT^2}{\widetilde{\varepsilon}}\ln(KT)^2\Big)^{1/3} \le 8C_w\Big(2\frac{KT^2}{\varepsilon(K,T)}\ln(KT)^2\Big)^{1/3} \le \frac{T}{2}\,,$$

as long as $\varepsilon(K,T) \ge 2\cdot 16^3C_w^3K(\ln(KT))^2/T$. Note that in the previous chain of inequalities we considered the (possibly suboptimal) choice of the (weakly) observable structure of the graph with threshold $\widetilde{\varepsilon}$ and upper bound on δ given by K. The inequality we just showed implies that the stopping criterion in RoundRobin is triggered and thus we can apply Lemma C.3.

Now, let τ^* be the smallest τ such that $\Phi(\mathcal{G},T) = \Phi([\mathcal{G}]_{\varepsilon^*},T) \le \tau K$, being ε^* the optimal threshold for $\mathcal{G}$. In this second step, we want to show that $\widehat{\tau}$ is not too far away from τ^* for the interesting values of τ^*; namely, that $\widehat{\tau} \le 4\tau^*$ as long as $\Phi(\mathcal{G},T)$ is not $\widetilde{\Omega}(T)$.

First, consider the case that $\Phi(\mathcal{G},T)$ refers to the strongly observable regime in $\Phi([\mathcal{G}]_{\varepsilon^*},T)$. By minimality of τ^*, we have the following:

$$\tau^*K \ge \Phi(\mathcal{G},T) = 4C_s\sqrt{\frac{\alpha^*}{\varepsilon^*}T}\big(\ln(KT)\big)^{3/2} \ge \frac{1}{2}\tau^*K\,. \tag{C.7}$$

We now set the constant appearing in the definition of $\varepsilon(K,T)$ from the statement to be $C = 2\cdot16^3C_w^3$. With this choice, the previously stated requirements for $\varepsilon(K,T)$ are satisfied, while at the same time it holds that $\Phi(\mathcal{G},T) \le C_s^2T(\ln(KT))^2/(15K)$; this is immediate to verify by arguing that $\Phi(\mathcal{G},T)$ is at most the regret incurred by using the (possibly suboptimal, weakly) observable structure of $\mathcal{G}$ truncated at $\varepsilon(K,T)$. Then, from the second inequality of (C.7), it follows that $\tau^* \le 2C_s^2T(\ln(KT))^2/(15K^2)$. We can rewrite the first inequality of (C.7) as follows:

$$\varepsilon^* \ge 16C_s^2\frac{\alpha^*}{(K\tau^*)^2}T\big(\ln(KT)\big)^3 \ge 120\frac{\ln(KT)}{\tau^*}\,.$$

Consider now to what happens at the $\overline{\tau} = \lceil 120\ln(KT)/\varepsilon^*\rceil \le 4\tau^*$ iteration of RoundRobin. The estimated graph $\widehat{\mathcal{G}}_{\overline{\tau}}$ in that iteration is an $\varepsilon^*/2$-good approximation of $\mathcal{G}$, thus it contains all the edges of $\mathcal{G}$, with the probabilities correctly estimated up to a constant multiplicative factor, as detailed in Definition 5.1 (part 2). Thus,

$$\Phi(\widehat{\mathcal{G}}_{4\tau^*},T) \le 2\Phi(\widehat{\mathcal{G}}_{\overline{\tau}},T) \le 2\sqrt{2}\Phi(\mathcal{G},T) \le 4\tau^*K,$$

which implies that the stopping time $\widehat{\tau}$ is attained before $4\tau^*$. Note that the first inequality is due to point (i) of Lemma C.2, whereas the second inequality follows from point (ii) of Lemma C.2.

Similarly, we consider the case that $\Phi(\mathcal{G}, T)$ refers to the weakly observable regime in $\Phi([\mathcal{G}]_{\varepsilon^*}, T)$. By minimality of τ^*, we have the following:

$$\tau^* K \geq \Phi(\mathcal{G}, T) = 4C_w \left(\frac{\delta^*}{\varepsilon^*}\big(\ln(KT)\big)^2\right)^{1/3} T^{2/3} \geq \frac{1}{2}\tau^* K \; . \tag{C.8}$$

By the choice of $\varepsilon(K, T)$, we have that $\Phi(\mathcal{G}, T) \leq T\sqrt{2C_w^3 \ln(KT)/(15K)}$. Then, from the second inequality of (C.8), it follows that $\tau^* \leq T\sqrt{8C_w^3 \ln(KT)/(15K^3)}$. Consider now the first inequality, we can rewrite it to obtain:

$$\varepsilon^* \geq 64C_w^3 \frac{\delta^*}{(K\tau^*)^3}\big(T\ln(KT)\big)^2 \geq 120\frac{\ln(KT)}{\tau^*} \; .$$

We can now use the same argument as in the strongly observable case and conclude that $\widehat{\tau} \leq 4\tau^*$.

At this point, we are ready to show that our algorithm EdgeCatcher exhibits the desired regret bounds. We are conditioning on the good event $\mathcal{E}$; this happens with probability at least $1 - \frac{1}{T}$, so we just analyze this case, as the complementary of $\mathcal{E}$ yields at most an extra additive 1, in expectation, to the regret bound.

Recall that R_T is the worst-case regret; thus,

$$R_T \leq \widehat{\tau}K + \Phi(\widehat{\mathcal{G}}, T - \widehat{\tau}K) \leq 2\widehat{\tau}K \leq 8\tau^* K \leq 16\Phi(\mathcal{G}, T) \; ,$$

where in the first inequality we used the decomposition in regret before and after the commitment and the bound on Lemma C.3 (which is applicable given the conditioning on $\mathcal{E}$ and thus all the $\mathcal{G}_\tau$ are ε_τ-good approximations of $\mathcal{G}$), in the second one the definition of $\widehat{\tau}$, in the third one the fact that $\widehat{\tau} \leq 4\tau^*$, and in the last the definition of τ^* as minimal τ such that $\Phi(\mathcal{G}, T) \leq \tau K$. □

C.3 Proofs of Lower Bounds

The main idea in the lower bounds is that the adversary sets all edge probabilities equal to $\varepsilon \in (0, 1]$ in order to define a stochastic feedback graph $\mathcal{G}$ with a specific support G that satisfies adequate properties. This requires the attribution of additional power to the adversary because we allow it to choose the edge probabilities; nevertheless, this is fine from a worst-case perspective because it corresponds to choosing a particularly difficult instance among those that have certain characteristics. Doing so makes the edge between each (ordered) pair of nodes either realize independently at each round t with probability equal to ε, or never realize. Moreover, there exists a vertex that is at least marginally better than the other ones with respect to the expected loss. The learner only obtains information about the loss of the optimal node whenever it plays a node that is adjacent to it in $G = \mathrm{supp}(\mathcal{G})$ and the edge between the played node and the optimal node is realized. Since that edge is realized only with probability ε, it is significantly harder for the the learner to detect the optimal node, which allows the adversary to increase the size of the gaps between the optimal node and the suboptimal ones. More specifically, while in the deterministic setting playing once action a is enough to observe the loss incurred by a neighbouring action a', the learner will now need $1/\varepsilon$

time steps, in expectation, to observe the loss of a' if the edge (a, a') only realizes with probability ε. Further notice that, in the setting considered within the proofs of our lower bounds, the learner may even know the true distribution $\mathcal{G}$ and observe the realization of the entire feedback graph G_t at the end of each round t.

We start with a lower bound for the strongly observable case considering stochastic feedback graphs $\mathcal{G}$ with $\alpha(\mathcal{G}) > 1$. The following result can be recovered by adapting the proof of Alon et al. (2017, Theorem 5) that holds for any graph of interest (directed or undirected).

Theorem C.1. *Pick any directed or undirected graph $G = (V, E)$ with $\alpha(G) > 1$ and any $\varepsilon \in (0, 1]$. There exists a stochastic feedback graph $\mathcal{G}$ with* $\operatorname{supp}(\mathcal{G}) = G$ *and such that, for all $T \geq 0.0064\alpha([\mathcal{G}]_\varepsilon)^3/\varepsilon$ and for any possibly randomized algorithm $\mathcal{A}$, there exists a sequence $\ell_1, \dots, \ell_T$ of loss functions on which the expected regret of $\mathcal{A}$ with respect to the stochastic generation of $G_1, \dots, G_T \sim \mathcal{G}$ is at least $0.017\sqrt{\alpha([\mathcal{G}]_\varepsilon)T/\varepsilon}$.*

Proof. The structure of this proof follows the same rationale of the lower bound by Alon et al. (2017, Theorem 5) with additional considerations due to the stochasticity of the feedback graph. To prove the lower bound we will use Yao's minimax principle (Yao, 1977), which shows that it is sufficient to provide a probabilistic strategy for the adversary on which the expected regret of any deterministic algorithm is lower bounded.

We can assume that G has all self-loops. If G is missing some self-loops, we may add them for the sake of the lower bound: this only makes the problem easier for the learner. Also note that the addition of self-loops does not change the independence number of G. Now let $\mathcal{G}$ be such that $p(i, j) \in \{0, \varepsilon\}$ and $p(i, j) := \varepsilon$ if and only if $(i, j) \in E$, for all $i, j \in V$. Note that $\alpha(G) = \alpha(\mathcal{G})$ and $\mathcal{G} = [\mathcal{G}]_\varepsilon$. We also remark that the following lower bound for such a $\mathcal{G}$ will be a lower bound for the instance having a stochastic feedback graph obtained from the starting graph, without the addition of self-loops, by setting the realization probability of all its edges to ε. Without loss of generality, we order the nodes depending on an (arbitrary) independent set of G of size $\alpha(G)$ so that $1, 2, \dots, \alpha(G)$ are the nodes belonging to said independent set, and $\alpha(G) + 1, \dots, |V|$ correspond to all the other nodes in G.

We will use the following distribution of losses. We sample Z from some (later defined) distribution Q over the independent set chosen above. Conditioned on $Z = i$, the loss $\ell_t(j)$ is sampled from an independent Bernoulli distribution with mean $\frac{1}{2}$ if $j \neq i$ and $j \leq \alpha(G)$, it is sampled from an independent Bernoulli with mean $\frac{1}{2} - \beta$ if $j = i$ for some $\beta \in [0, \frac{1}{4}]$, and it is set to 1 otherwise.

We denote by T_i the number of times node i was chosen by the algorithm after T rounds and denote by $T_{\text{bad}} := \sum_{i > \alpha(G)} T_i$ the number of times the algorithm chooses an action not in the independent set. We use $\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot \mid Z = i]$ and $\mathbb{P}_i(\cdot) := \mathbb{P}(\cdot \mid Z = i)$ to denote the expectation and probability over $(G_1, \ell_1), \dots, (G_T, \ell_T)$ conditioned on $Z = i$, respectively. We denote by $\ell_t(I_t)$ the loss of algorithm A playing I_t in round t. We emphasize that the complete loss sequence and the (partial) loss sequence observed by the learner may differ depending not only on the actions of the learner but also on the realization of the edges in the feedback graph. This last observation will be used to lower bound the regret of the learner also in terms of ε, the probability of an edge realization.

We set $Q(i) := \frac{1}{\alpha(G)}$ if i is in the independent set and $Q(i) := 0$ otherwise. Following Alon et al. (2017, Equation (8)) we have, for any deterministic algorithm A, that

$$\max_{k \in V} \mathbb{E}\left[\sum_{t=1}^{T} \big(\ell_t(I_t) - \ell_t(k)\big)\right] \geq \beta\bigg(T - \frac{1}{\alpha(G)} \sum_{i \leq \alpha(G)} \mathbb{E}_i[T_i]\bigg) . \tag{C.9}$$

We now consider an auxiliary distribution $\mathbb{P}_0$, also over $(G_1, \ell_1), \ldots, (G_T, \ell_T)$, which is equivalent to the distribution $\mathbb{P}_i$ that we specified above, but with $\beta = 0$ for all nodes. We denote by $\mathbb{E}_0$ the corresponding expectation. We also denote by λ_t the feedback set at time t, composed by the realization G_t of the feedback graph together with the set of losses observed by the learner in round t, and by $\lambda^t \coloneqq (\lambda_1, \ldots, \lambda_t)$ the tuple of all feedback sets up to and including round t. Since the algorithm is deterministic, its action I_t in round t is fully determined by λ^{t-1}. Therefore, $\mathbb{E}_i[T_i \mid \lambda^T] = \mathbb{E}_0[T_i \mid \lambda^T]$. When λ^{t-1} is understood from the context, let $\mathbb{P}_{j,t} \coloneqq \mathbb{P}_j(\cdot \mid \lambda^{t-1})$ be the conditional probability measure of feedback sets λ_t at time t. We have that

$$\begin{aligned}
\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] &= \sum_{\lambda^T} \mathbb{P}_i(\lambda^T)\mathbb{E}_i[T_i \mid \lambda^T] - \sum_{\lambda^T} \mathbb{P}_0(\lambda^T)\mathbb{E}_0[T_i \mid \lambda^T] \\
&= \sum_{\lambda^T} \mathbb{P}_i(\lambda^T)\mathbb{E}_i[T_i \mid \lambda^T] - \sum_{\lambda^T} \mathbb{P}_0(\lambda^T)\mathbb{E}_i[T_i \mid \lambda^T] \\
&\leq T \sum_{\lambda^T : \mathbb{P}_i(\lambda^T) > \mathbb{P}_0(\lambda^T)} \big(\mathbb{P}_i(\lambda^T) - \mathbb{P}_0(\lambda^T)\big) .
\end{aligned}$$

By using Pinsker's inequality and the chain rule for the relative entropy, we can further observe that

$$\begin{aligned}
\sum_{\lambda^T : \mathbb{P}_i(\lambda^T) > \mathbb{P}_0(\lambda^T)} \big(\mathbb{P}_i(\lambda^T) - \mathbb{P}_0(\lambda^T)\big) &\leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0 \,\|\, \mathbb{P}_i)} \\
&= \sqrt{\frac{1}{2} \sum_{t=1}^{T} \sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\mathrm{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t})} ,
\end{aligned}$$

which, combined with the previous inequality, allows us to affirm that

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \leq \sqrt{\frac{1}{2} \sum_{t=1}^{T} \sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\mathrm{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t})} . \tag{C.10}$$

At this point, observe that $\operatorname{supp}(\mathcal{G}) = G = (V, E)$. Fix any λ^{t-1} and consider $D_{\mathrm{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t})$ where, we recall, $\mathbb{P}_{0,t}(\lambda_t) = \mathbb{P}_0(\lambda_t \mid \lambda^{t-1})$ and $\mathbb{P}_{i,t}(\lambda_t) = \mathbb{P}_i(\lambda_t \mid \lambda^{t-1})$. Recall that λ^{t-1} fully determines the node I_t picked by the algorithm in round t. If $(I_t, i) \notin E$, then $\mathbb{P}_{0,t}$ and $\mathbb{P}_{i,t}$ have the same distribution and the relative entropy term is 0. If $(I_t, i) \in E$, then the loss of node i in λ_t follows a Bernoulli distribution with mean $\frac{1}{2}$ under $\mathbb{P}_0$ and follows a Bernoulli distribution with mean $\frac{1}{2} - \beta$ under $\mathbb{P}_i$. Denote by $\mathcal{E}_t$ the event that edge (I_t, i) is realized in G_t. Note that $\mathbb{P}_0(\mathcal{E}_t) = \mathbb{P}_i(\mathcal{E}_t) = \varepsilon$. Using the log-sum inequality and the fact that the relative entropy between the two aforementioned Bernoulli distributions is given by $\frac{1}{2}\ln\big(\frac{1}{1-4\beta^2}\big)$, we can see that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t}) &= D_{\mathrm{KL}}\Big(\varepsilon\mathbb{P}_{0,t}(\cdot \mid \mathcal{E}_t) + (1-\varepsilon)\mathbb{P}_{0,t}(\cdot \mid \overline{\mathcal{E}_t}) \,\Big\|\, \varepsilon\mathbb{P}_{i,t}(\cdot \mid \mathcal{E}_t) + (1-\varepsilon)\mathbb{P}_{i,t}(\cdot \mid \overline{\mathcal{E}_t})\Big) \\
&= D_{\mathrm{KL}}\Big(\varepsilon\mathbb{P}_{0,t}(\cdot \mid \mathcal{E}_t) + (1-\varepsilon)\mathbb{P}_{0,t}(\cdot \mid \overline{\mathcal{E}_t}) \,\Big\|\, \varepsilon\mathbb{P}_{i,t}(\cdot \mid \mathcal{E}_t) + (1-\varepsilon)\mathbb{P}_{0,t}(\cdot \mid \overline{\mathcal{E}_t})\Big) \\
&\leq \varepsilon D_{\mathrm{KL}}\big(\mathbb{P}_{0,t}(\cdot \mid \mathcal{E}_t) \,\|\, \mathbb{P}_{i,t}(\cdot \mid \mathcal{E}_t)\big) + (1-\varepsilon) D_{\mathrm{KL}}\big(\mathbb{P}_{0,t}(\cdot \mid \overline{\mathcal{E}_t}) \,\|\, \mathbb{P}_{0,t}(\cdot \mid \overline{\mathcal{E}_t})\big)
\end{aligned}$$

$$= \varepsilon D_{\mathrm{KL}}\big(\mathbb{P}_{0,t}(\cdot \,|\, \mathcal{E}_t) \,\|\, \mathbb{P}_{i,t}(\cdot \,|\, \mathcal{E}_t)\big)$$
$$= -\frac{\varepsilon}{2} \ln\big(1 - 4\beta^2\big) \le 8\ln(4/3)\beta^2\varepsilon \ . \tag{C.11}$$

With this inequality, we may upper bound the sum in the right-hand side of (C.10) by considering, for each t, only the tuples λ^{t-1} for which $i \in N_G^{\mathrm{out}}(I_t)$ holds. Indeed, the KL divergence for any other possible λ^{t-1} is equal to 0 because the edge (I_t, i) never realizes (it is not in the support of $\mathcal{G}$, hence $p(I_t, i) = 0$). As a consequence,

$$\begin{aligned}\sum_{t=1}^{T} \sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\mathrm{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t}) &\le \sum_{t=1}^{T} \mathbb{P}_0(i \in N_G^{\mathrm{out}}(I_t)) 8\ln(4/3)\beta^2\varepsilon \\ &= 8\ln(4/3)\beta^2\varepsilon\mathbb{E}_0[|\{t : i \in N_G^{\mathrm{out}}(I_t)\}|] \\ &\le 8\ln(4/3)\beta^2\varepsilon\mathbb{E}_0[T_i + T_{\mathrm{bad}}] \ .\end{aligned} \tag{C.12}$$

We may claim that $\mathbb{E}_0[T_{\mathrm{bad}}] \le 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T$, because otherwise the expected regret under $\mathbb{P}_0$ would have been at least

$$\begin{aligned}\max_{k \in V} \mathbb{E}_0\left[\sum_{t=1}^{T} (\ell_t(I_t) - \ell_t(k))\right] &= \mathbb{E}_0\left[T_{\mathrm{bad}} + \frac{1}{2}\sum_{j \le \alpha(G)} T_j\right] - \frac{1}{2}T \\ &= \mathbb{E}_0\left[\frac{1}{2}T_{\mathrm{bad}} + \frac{1}{2}\left(T_{\mathrm{bad}} + \sum_{j \le \alpha(G)} T_j\right)\right] - \frac{1}{2}T \\ &= \mathbb{E}_0\left[\frac{1}{2}T_{\mathrm{bad}}\right] \\ &> 0.02\sqrt{\frac{\alpha(G)}{\varepsilon}}T \ .\end{aligned}$$

Combining Equations (C.10) and (C.12), and using that $\mathbb{E}_0[T_{\mathrm{bad}}] \le 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T$, we find that

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \le 2T\beta\sqrt{\varepsilon\ln(4/3)\mathbb{E}_0\left[T_i + 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T\right]} \ .$$

This implies that the regret can be further lower bounded, continuing from (C.9), by

$$\begin{aligned}&\beta\left(T - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_0[T_i] - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} 2T\beta\sqrt{\varepsilon\ln(4/3)\mathbb{E}_0\left[T_i + 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T\right]}\right) \\ &\ge \beta\left(T - \frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} \mathbb{E}_0[T_i] - 2T\beta\sqrt{\varepsilon\ln(4/3)\mathbb{E}_0\left[\frac{1}{\alpha(G)}\sum_{i=1}^{\alpha(G)} T_i + 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T\right]}\right) \\ &\ge \beta T\left(1 - \frac{1}{\alpha(G)} - 2\beta\sqrt{\varepsilon\ln(4/3)\left(\frac{T}{\alpha(G)} + 0.04\sqrt{\frac{\alpha(G)}{\varepsilon}}T\right)}\right) \ ,\end{aligned}$$

where the first inequality is Jensen's inequality for concave functions and the second inequality is due to the fact that $\sum_{i=1}^{\alpha(G)} \mathbb{E}_0[T_i] \leq T$ by definition of T_i. Since we assumed that $T \geq 0.0064\alpha(G)^3/\varepsilon$, we have that $0.04\sqrt{\frac{\alpha(G)}{\varepsilon}T} \leq \frac{T}{2\alpha(G)}$ and thus

$$\begin{aligned}\max_{k\in V} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(k))\right] &\geq \beta T\left(1 - \frac{1}{\alpha(G)} - 2\beta\sqrt{\frac{3}{2}\ln(4/3)\frac{\varepsilon T}{\alpha(G)}}\right) \\ &\geq \beta T\left(\frac{1}{2} - 2\beta\sqrt{\frac{3}{2}\ln(4/3)\frac{\varepsilon T}{\alpha(G)}}\right) ,\end{aligned}$$

where in the second inequality we used the assumption that $\alpha(G) \geq 2$. By setting $\beta = \frac{1}{33}\sqrt{\frac{\alpha(G)}{2\ln(4/3)\varepsilon T}} \in (0, \frac{1}{4}]$, we may complete the proof as

$$\max_{k\in V} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(k))\right] \geq \frac{1}{33}\left(\frac{1}{2} - \frac{\sqrt{3}}{33}\right)\sqrt{\frac{\alpha(G)T}{2\ln(4/3)\varepsilon}} \geq 0.017\sqrt{\frac{\alpha(G)}{\varepsilon}T} .$$

□

Given that this lower bound leaves the case $\alpha(\mathcal{G}) = 1$ uncovered, we provide an additional lower bound that considers any feedback graph. This new bound is tight up to logarithmic factors, for instance, in all cases where $\alpha(\mathcal{G})$ is constant.

Theorem C.2. *Pick any directed or undirected graph $G = (V, E)$ with $|V| = K \geq 2$ and any $\varepsilon \in (0, 1]$. There exists a stochastic feedback graph $\mathcal{G}$ with $\operatorname{supp}(\mathcal{G}) = G$ and such that, for all $T \geq 1/(2\varepsilon)$ and for any possibly randomized algorithm $\mathcal{A}$, there exists a sequence $\ell_1, \dots, \ell_T$ of loss functions on which the expected regret of $\mathcal{A}$ with respect to the stochastic generation of $G_1, \dots, G_T \sim \mathcal{G}$ is at least $\frac{1}{32}\sqrt{2T/\varepsilon}$.*

Proof. Following a similar rationale as in the proof of Theorem C.1, we can consider G to be the complete graph (with all self-loops) because the problem for it is easier than that with any other graph. In fact, adding edges never makes the problem harder to solve. Moreover, we can define $\mathcal{G}$ by setting all edge probabilities to ε so that $[\mathcal{G}]_\varepsilon = \mathcal{G}$ and $\operatorname{supp}(\mathcal{G}) = G$. We remark that the lower bound with such a $\mathcal{G}$ is also a lower bound for the instance obtained by considering the initial (possibly non-complete) graph and assigning realization probability ε to all its edges. Applying Yao's minimax principle allows us to reduce our current aim to proving a lower bound for the expected regret of any deterministic algorithm against a randomized adversary.

We can then construct the sequence of loss functions by defining their distribution. Let $v \in V$ be an arbitrary vertex, say, $v = 1$. Pick $Z \in \{-1, +1\}$ uniformly at random and define $\beta \coloneqq \frac{1}{4}(2\varepsilon T)^{-1/2} \in [0, \frac{1}{4}]$. Then, let the loss at any time t be independently $\ell_t(i) \sim \operatorname{Ber}\left(\frac{1}{2}\right)$ for $i \neq 1$ while $\ell_t(1) \sim \operatorname{Ber}\left(\frac{1}{2} - \beta Z\right)$. Define $\mathbb{P}_1(\cdot) \coloneqq \mathbb{P}(\cdot \mid Z = +1)$ and $\mathbb{P}_2(\cdot) \coloneqq \mathbb{P}(\cdot \mid Z = -1)$, as well as $\mathbb{E}_1[\cdot] \coloneqq \mathbb{E}[\cdot \mid Z = +1]$ and $\mathbb{E}_2[\cdot] \coloneqq \mathbb{E}[\cdot \mid Z = -1]$. We also define $\mathbb{P}_0(\cdot)$ and $\mathbb{E}_0[\cdot]$, obtained in an analogous manner as the previous ones by setting $\beta = 0$.

At this point, let T_1 be the number of times t that the algorithm selects vertex $I_t = 1$ after T rounds. Following a similar computation as in Equations (C.10) and (C.12), we first denote by

$\mathbb{P}_{j,t} \coloneqq \mathbb{P}_j(\cdot \mid \lambda^{t-1})$ the conditional probability over feedback sets λ_t, and notice that

$$\begin{aligned}\mathbb{E}_1[T_1] - \mathbb{E}_2[T_1] &\leq T\sqrt{\frac{1}{2}\sum_{t=1}^{T}\sum_{\lambda^{t-1}} \mathbb{P}_2(\lambda^{t-1}) D_{\mathrm{KL}}(\mathbb{P}_{2,t} \,\|\, \mathbb{P}_{1,t})} \\ &\leq T\sqrt{\varepsilon\beta T \ln\Bigl(1 + \frac{4\beta}{1-2\beta}\Bigr)} \\ &\leq 2\beta T\sqrt{2\varepsilon T}\ . \qquad \text{(C.13)}\end{aligned}$$

Conditioning on $Z = +1$, the algorithm incurs an expected instantaneous regret equal to β whenever it picks any vertex $i \neq 1$. Otherwise, conditioning on $Z = -1$, the algorithm incurs the same expected instantaneous regret each time it selects vertex 1. The expected regret thus becomes

$$\begin{aligned}\max_{k\in V} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(k))\right] &\geq \frac{1}{2}\mathbb{E}_1[\beta(T - T_1)] + \frac{1}{2}\mathbb{E}_2[\beta T_1] \\ &\geq \frac{\beta}{2}T - \frac{\beta}{2}(\mathbb{E}_1[T_1] - \mathbb{E}_2[T_1]) \\ &\geq \beta T\Bigl(\frac{1}{2} - \beta\sqrt{2\varepsilon T}\Bigr) = \frac{1}{4}\beta T = \frac{1}{32}\sqrt{\frac{2T}{\varepsilon}}\ ,\end{aligned}$$

where the third inequality follows by Equation (C.13), and we also use our choice of β. □

We can additionally prove further lower bounds for the weakly observable case. Here we also adapt the proof for the lower bound in the case of a deterministic feedback graph by having each edge realize only with probability $\varepsilon \in (0, 1]$ at each time step. We make the same considerations as in the previous lower bound for strongly observable graphs. In this case, however, we refer to Alon et al. (2015, Theorem 7). As in the case of deterministic feedback graph, we need the following combinatorial lemma.

Lemma C.4 (Alon et al. (2015, Lemma 8)). *Let $G = (V, E)$ be a directed graph over $|V| = n$ vertices, and let $W \subseteq V$ be a set of vertices whose minimal dominating set is of size k. Then, there exists an independent set $U \subseteq W$ of size $|U| \geq \frac{1}{50}k/\ln n$, such that any vertex of G dominates at most $\ln n$ vertices of U.*

We can then prove the desired lower bound which states what follows.

Theorem C.3. *Pick any directed or undirected, weakly observable graph $G = (V, E)$ with $|V| = K$ and $\delta(G) \geq 100 \ln K$, and any $\varepsilon \in (0, 1]$. There exists a stochastic feedback graph $\mathcal{G}$ with $\operatorname{supp}(\mathcal{G}) = G$ and such that, for all $T \geq 2K/(\varepsilon \ln K)$ and for any possibly randomized algorithm $\mathcal{A}$, there exists a sequence $\ell_1, \dots, \ell_T$ of loss functions on which the expected regret of $\mathcal{A}$ with respect to the stochastic generation of $G_1, \dots, G_T \sim \mathcal{G}$ is at least $\frac{1}{150}\bigl(\frac{\delta([\mathcal{G}]_\varepsilon)}{\varepsilon \ln^2 K}\bigr)^{1/3} T^{2/3}$.*

Proof. The proof follows the steps of the lower bound from Alon et al. (2015, Theorem 7). As in the previous lower bounds, we use Yao's minimax principle to infer that it suffices to design a probabilistic adversarial strategy that leads to a sufficiently large lower bound for the expected regret of any deterministic algorithm.

We consider any weakly observable $G = (V, E)$ having $|V| = K$ vertices and $\delta(G) \geq 100 \ln K$. Since the adversary may choose edge probabilities, it can pick them all equal to ε so that $\mathcal{G} = [\mathcal{G}]_\varepsilon$

and $\operatorname{supp}(\mathcal{G}) = G$. By Lemma C.4 we know that G contains an independent set U of size $|U| = m \geq \delta(G)/(50 \ln K)$ such that any $v \in V$ dominates no more than $\ln K$ vertices of U. We will denote actions in U as "good" actions, whereas all the others will be denoted as "bad" actions. Given our assumption on $\delta(G)$, we observe that $m \geq 2$. A further observation we can make is that $N_G^{\text{in}}(i) \subseteq V \setminus U$ for all $i \in U$ because U is independent, meaning that we need to pick a bad action in order to be able to observe the loss of any good action.

As similarly done in the proof of Theorem C.1, we sample Z from our "target" set U uniformly at random. This choice induces a distribution of the losses $\ell_t(i)$ for all t and all i independently. To be precise, given $\beta \coloneqq m^{1/3}(32\varepsilon T \ln K)^{-1/3} \in [0, \frac{1}{4}]$, the loss is $\ell_t(i) \sim \mathsf{Bern}(\frac{1}{2} - \beta)$ if $i = Z$, while it is $\ell_t(i) \sim \mathsf{Bern}(\frac{1}{2})$ if $i \in U$, $i \neq Z$. The loss is deterministically set to $\ell_t(i) \coloneqq 1$ for any other vertex $i \in V \setminus U$.

Taking up the same notation introduced in the proof of Theorem C.1, we denote by T_i the number of times action i is played by the deterministic algorithm after T rounds, while $T_{\text{bad}} \coloneqq \sum_{i \in V \setminus U} T_i$. In particular, I_t is the action chosen by the algorithm at time t. We also use $\mathbb{P}_i(\cdot) \coloneqq \mathbb{P}(\cdot \mid Z = i)$ and $\mathbb{E}_i[\cdot] \coloneqq \mathbb{E}[\cdot \mid Z = i]$ with a similar definition, including the auxiliary distribution $\mathbb{P}_0$ and the corresponding expectation $\mathbb{E}_0$ obtained by setting $\beta = 0$. Moreover, for each good action i we introduce $X_i \coloneqq \sum_{t=1}^T \mathbb{I}\{I_t \in N_G^{\text{in}}(i)\}$ to denote the number of times the algorithm picks a bad action from $N_G^{\text{in}}(i)$.

Notice that we can restrict our reasoning to algorithms that have $T_{\text{bad}} \leq \beta T$ (otherwise reducing to this case by only introducing a factor 3 in the regret bound), as similarly argued in the proof of Alon et al. (2015, Theorem 7). This implies that

$$\sum_{i \in U} X_i \leq T_{\text{bad}} \ln K \leq \beta T \ln K \tag{C.14}$$

since each $j \in V \setminus U$ dominates at most $\ln K$ vertices of U.

Recalling Equation (C.10), we are interested in bounding

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \leq T \sqrt{\frac{1}{2} \sum_{t=1}^T \sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\text{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t})} \;, \tag{C.15}$$

where $\mathbb{P}_{j,t} \coloneqq \mathbb{P}_j(\cdot \mid \lambda^{t-1})$ is the conditional probability over feedback sets λ_t. The KL divergence in the above sum is $D_{\text{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t}) \leq 8 \ln(4/3) \beta^2 \varepsilon$, where we use a similar reasoning as in Equation (C.11). As a consequence,

$$\begin{aligned}
\sum_{t=1}^T \sum_{\lambda^{t-1}} \mathbb{P}_0(\lambda^{t-1}) D_{\text{KL}}(\mathbb{P}_{0,t} \,\|\, \mathbb{P}_{i,t}) &\leq \sum_{t=1}^T \mathbb{P}_0(I_t \in N_G^{\text{in}}(i)) 8 \ln(4/3) \beta^2 \varepsilon \\
&\leq 4\beta^2 \varepsilon \mathbb{E}_0[|\{t : I_t \in N_G^{\text{in}}(i)\}|] \\
&= 4\beta^2 \varepsilon \mathbb{E}_0[X_i] \;,
\end{aligned}$$

which together with Equation (C.15) allows us to show that

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \leq \beta T \sqrt{2\varepsilon \mathbb{E}_0[X_i]} \;. \tag{C.16}$$

Let us now consider the expected regret for the deterministic algorithm at hand. We know that it must be at least

$$\max_{k\in V}\mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t)-\ell_t(k))\right]\geq\frac{1}{m}\sum_{i\in U}\mathbb{E}_i[\beta(T-T_i)]=\beta T-\frac{\beta}{m}\sum_{i\in U}\mathbb{E}_i[T_i]$$

because the algorithm incurs at least β regret each time it picks an action different from Z. By Equations (C.14) and (C.16), and using the concavity of the square root, the summation on the right-hand side is such that

$$\begin{aligned}\frac{1}{m}\sum_{i\in U}\mathbb{E}_i[T_i]&\leq\beta T\sqrt{\frac{2\varepsilon}{m}\sum_{i\in U}\mathbb{E}_0[X_i]}+\frac{1}{m}\mathbb{E}_0\left[\sum_{i\in U}T_i\right]\\&\leq T\sqrt{\frac{2\beta^3\varepsilon}{m}T\ln K}+\frac{T}{m}\\&=\frac{1}{4}T+\frac{1}{m}T\leq\frac{3}{4}T\ ,\end{aligned}\tag{C.17}$$

where the equality follows by our choice of β, whereas the last inequality holds because $m\geq 2$. Hence, the expected regret is

$$\max_{k\in V}\mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t)-\ell_t(k))\right]\geq\frac{\beta}{4}T=\frac{1}{4}\left(\frac{m}{32\varepsilon\ln K}\right)^{1/3}T^{2/3}\geq\frac{1}{50}\left(\frac{\delta(G)}{\varepsilon\ln^2 K}\right)^{1/3}T^{2/3}\,.$$

□

An additional theorem is required in order to cover the case $\delta(\mathcal{G})<100\ln K$. In the same way as in Alon et al. (2015), we follow a simple reasoning with generic weakly observable graphs. The following lower bound holds for weakly observable graphs of any size and is tight up to logarithmic factors for instances having $\delta(\mathcal{G})<100\ln K$.

Theorem C.4. *Pick any directed or undirected, weakly observable graph $G=(V,E)$ with $|V|\geq 2$ and any $\varepsilon\in(0,1]$. There exists a stochastic feedback graph $\mathcal{G}$ with* $\operatorname{supp}(\mathcal{G})=G$ *and such that, for all $T\geq 2\sqrt{2}/\varepsilon$ and for any possibly randomized algorithm $\mathcal{A}$, there exists a sequence $\ell_1,\dots,\ell_T$ of loss functions on which the expected regret of $\mathcal{A}$ with respect to the stochastic generation of $G_1,\dots,G_T\sim\mathcal{G}$ is at least $\frac{\sqrt{2}}{16}\varepsilon^{-1/3}T^{2/3}$.*

Proof. The proof follows a similar structure as that of Alon et al. (2015, Theorem 11). We consider the same instance constituted by a graph $G=(V,E)$ having $|V|\geq 3$ vertices, since it is the minimum number of vertices in order for G to be weakly observable. In fact, any graph with exactly 2 vertices is either unobservable or strongly observable. By definition, there exists a vertex in this graph with no self-loop and with at least one incoming edge missing from any of the remaining vertices. Without loss of generality, let $v=1$ be such a vertex and let $2\notin N^{\text{in}}_G(v)$ be one of the vertices without an edge towards v. We may consider the case where all edge probabilities are set to ε (implying that $\mathcal{G}=[\mathcal{G}]_\varepsilon$ and $\operatorname{supp}(\mathcal{G})=G$), given that we essentially assume the adversary can select them.

We can apply Yao's minimax principle, as usual, to reduce this problem to that of lower bounding the expected regret for any deterministic algorithm against a randomized adversary. Hence, we need to design a distribution for the loss functions $\ell_1,\dots,\ell_T$ provided to the algorithm. Let

$\beta := \frac{1}{2\sqrt{2}}(\varepsilon T)^{-1/3} \in [0, \frac{1}{4}]$ and pick $Z \in \{-1, +1\}$ uniformly at random. For all t, we choose the losses such that $\ell_t(1) \sim \mathrm{Ber}\,(1/2 - \beta Z)$, $\ell_t(2) \sim \mathrm{Ber}\,(1/2)$, and $\ell_t(j) := 1$ for all $j \neq 1$ independently. Similarly to the construction in the proof of Theorem C.3, we have "good" actions $\{1, 2\}$ incurring at most β expected instantaneous regret, while all remaining actions are "bad" since they incur at least 1/2 instantaneous regret in expectation.

We reuse the same definitions for T_i and X_i as in the proof of Theorem C.3 for any fixed deterministic algorithm. On the other hand, we let $\mathbb{P}_1(\cdot) := \mathbb{P}\left(\cdot \mid Z = +1\right)$ and $\mathbb{P}_2(\cdot) := \mathbb{P}\left(\cdot \mid Z = -1\right)$. We analogously define $\mathbb{E}_1[\cdot] := \mathbb{E}\left[\cdot \mid Z = +1\right]$ and $\mathbb{E}_2[\cdot] := \mathbb{E}\left[\cdot \mid Z = -1\right]$. Finally, we introduce $\mathbb{P}_0(\cdot)$ and $\mathbb{E}_0[\cdot]$ obtained as the previous ones by setting $Z = 0$.

Following the same rationale that led to Equation (C.16), we can show that

$$\mathbb{E}_i[T_i] - \mathbb{E}_0[T_i] \leq \beta T \sqrt{2\varepsilon \mathbb{E}_i[X_1]}$$

for $i \in \{1, 2\}$. This implies, via similar steps as in Equation (C.17), that

$$\frac{1}{2}\mathbb{E}_1[T_1] + \frac{1}{2}\mathbb{E}_2[T_2] \leq \beta T \sqrt{2\varepsilon \mathbb{E}\left[X_1\right]} + \frac{T}{2} \ . \tag{C.18}$$

Finally, if $\mathbb{E}\left[X_1\right] > \frac{1}{32}\beta^{-2}\varepsilon^{-1}$, the algorithm's expected regret becomes

$$\max_{k \in V} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(k))\right] \geq \frac{1}{2}\mathbb{E}\left[X_1\right] > \frac{1}{64}\beta^{-2}\varepsilon^{-1} = \frac{1}{8}\varepsilon^{-1/3}T^{2/3} \ ,$$

where the last equality holds by our choice of β. Otherwise, when $\mathbb{E}\left[X_1\right] \leq \frac{1}{32}\beta^{-2}\varepsilon^{-1}$, the right-hand side of Equation (C.18) is bounded by $\frac{3}{4}T$ and thus the regret must be

$$\max_{k \in V} \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(I_t) - \ell_t(k))\right] \geq \frac{1}{2}\mathbb{E}_1[\beta(T - T_1)] + \frac{1}{2}\mathbb{E}_2[\beta(T - T_2)] \geq \frac{\beta}{4}T = \frac{\sqrt{2}}{16}\varepsilon^{-1/3}T^{2/3} \ . \quad \square$$

C.4 Be Optimistic If You Can, Commit If You Must

In this section, we describe Algorithm C.1 and the analysis we use to obtain the results of Section 5.5. First of all, we briefly state the rationale for the design of this new algorithm. The main idea is similar in spirit to that of EdgeCatcher: Algorithm C.1 constantly updates the estimates for the edge probabilities of the underlying $\mathcal{G}$ and computes the best regret regime it can achieve. However, EdgeCatcher has to wait until it can determine the best regret regime before actually tackling the learning task.

Algorithm C.1: OptimisticThenCommitGraph (otcG)

```
Environment: stochastic feedback graph $\mathcal{G}$, sequence of losses $\ell_1, \ell_2, \dots, \ell_T$
Input: time horizon $T$ and actions $V = \{1, 2, \dots, K\}$
Initialize: sample $I_1$ uniformly at random, receive $G_1$
for $t = 2, \dots, T$ do
    if Equation (C.19) has never been true then                                ▷ optimistic phase
        $\widetilde{p}_t(j,i) \leftarrow \frac{1}{t-1} \sum_{s=1}^{t-1} \mathbb{I}\{(j,i) \in E_s\}$
        $\widehat{p}_t(j,i) \leftarrow \widetilde{p}_t(j,i) + \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1} \ln(3K^2T^2)} + \frac{3}{t-1} \ln(3K^2T^2)$
        $\widehat{\mathcal{G}}_t^{\mathrm{UCB}} \leftarrow \{\widehat{p}_t(j,i) : i, j \in V\}$
        compute $\theta_t$ and $\varepsilon_t^\theta$ as in Equation (C.26)
        $\widehat{\mathcal{G}}_t \leftarrow \{\widehat{p}_t(j,i)\mathbb{I}\{\widehat{p}_t(j,i) \geq \varepsilon_t^\theta\} : i, j \in V\}$ and $\widehat{G}_t \leftarrow \mathrm{supp}\big(\widehat{\mathcal{G}}_t\big)$
        compute $p_t^{\min} \leftarrow \min_i \min_{j \in N^{\mathrm{in}}_{\widehat{G}_t}(i)} \widehat{p}_t(j,i)$
        $\gamma_t \leftarrow \min\left\{\left(\min_{s \in [2,t]} t p_s^{\min}\right)^{-1/2}, \frac{1}{2}\right\}$
        $\eta_{t-1} \leftarrow \left(16/(\min_{s \in [2,t]}(p_s^{\min})^2) + 4t/(\min_{s \in [2,t]} p_s^{\min}) + \sum_{s=2}^{t-1} \theta_s(\widehat{\mathcal{G}}_s)\right)^{-1/2}$
        set $\psi_t$ to be the uniform distribution over $V$
        set $q_t(i) \propto \exp\left(\eta_{t-1} \sum_{s=2}^{t-1} \widetilde{\ell}_t(i)\right)$
        $\pi_t(i) \leftarrow (1 - \gamma_t) q_t(i) + \gamma_t \psi_t(i)$
    if Equation (C.19) is true for any $t' - 1 < t$ then                        ▷ commit phase
        set $t^\star$ to the first round $t' - 1$ in which Equation (C.19) is true
        set $\widetilde{\mathcal{G}} = \{\widetilde{p}(j,i) : i, j \in V\}$ as the stochastic graph with $\widetilde{p}(j,i) = \frac{1}{t^\star} \sum_{s=1}^{t^\star} \mathbb{I}\{(j,i) \in E_s\}$
        set $\widehat{\mathcal{G}} = \{\widetilde{p}(j,i)\mathbb{I}\{\widetilde{p}(j,i) \geq \varepsilon_{t^\star}\} : i, j \in V\}$ with $\varepsilon_{t^\star}$ as in Equation (C.29)
        set $[\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}} \leftarrow \{\widetilde{p}(j,i)\mathbb{I}\{\widetilde{p}(j,i) \geq \varepsilon^\star_{\delta,\sigma}\} : i, j \in V\}$ with $\varepsilon^\star_{\delta,\sigma}$ as in Equation (C.30)
        $\widetilde{p}_t(j,i) \leftarrow \frac{1}{t-1} \sum_{s=1}^{t-1} \mathbb{I}\{(j,i) \in E_s\}$
        $\widehat{p}_t(j,i) \leftarrow \widetilde{p}_t(j,i) + \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1} \ln(3K^2T^2)} + \frac{3}{t-1} \ln(3K^2T^2)$
        $\widehat{\mathcal{G}}_t^{\mathrm{UCB}} \leftarrow \{\widehat{p}_t(j,i) : i, j \in V\}$
        $\widehat{\mathcal{G}}_t \leftarrow \widehat{\mathcal{G}}_t^{\mathrm{UCB}}$ and $\widehat{G}_t \leftarrow \mathrm{supp}\big(\widehat{\mathcal{G}}_t\big)$
        $\gamma \leftarrow \min\left\{\left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(KT)\right)^{1/3} T^{-1/3}, \frac{1}{2}\right\}$
        $\eta \leftarrow \sqrt{\ln(K) \left(2T \left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})/\gamma + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)\right)^{-1}}$
        set $\psi_t$ according to (C.31)
        set $q_t(i) \propto \exp\left(\eta \sum_{s=t^\star+1}^{t-1} \widetilde{\ell}_t(i)\right)$
        $\pi_t(i) \leftarrow (1 - \gamma) q_t(i) + \gamma \psi_t(i)$
    sample $I_t \sim \pi_t$
    receive $G_t$ and $\{(i, \ell_t(i)) : i \in N^{\mathrm{out}}_{G_t}(I_t)\}$
    compute $\widetilde{\ell}_t(i)$ as in (5.6)
```

On the contrary, Algorithm C.1 begins by optimistically assuming that the best thresholded graph has a strongly observable support while simultaneously updating the edge probability estimates; this is made possible given the additional assumption on receiving the realized graph $G_t = (V, E_t) \sim \mathcal{G}$ together with the observed losses at the end of each round t. At any point in time, as soon as Algorithm C.1 finds that it can achieve a better regret regime by switching to the weakly observable one (by computing the optimal threshold on the current estimate for $\mathcal{G}$), it commits to weak observability. We can prove that this strategy is able to achieve the best possible regret over all thresholded feedback graphs, analogously to EdgeCatcher, but with a dependency on the improved graph-theoretic parameters introduced in Section 5.5.

Consequently, there are two regimes of Algorithm C.1. In the first regime, the algorithm works under the assumption that $\operatorname{supp}(\mathcal{G})$ is strongly observable; in the second regime, the algorithm works under the assumption that $\operatorname{supp}(\mathcal{G})$ is observable. The switch happens in round $t^\star + 1$, where $t^\star$ is the first round $t-1$ in which

$$\Psi_{t-1} \geq \Lambda_{t-1}, \tag{C.19}$$

is true. The term Ψ_t is an upper bound on the regret after the first t rounds, and is given by

$$\begin{aligned}\Psi_t \coloneqq \min\Bigg\{ t,\, & 2 + 11(\ln(3K^2T^2))^2 \max_{s\in[2,t]} \theta_s(\widehat{\mathcal{G}}_s) \\ & + \left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\sqrt{t \max_{s\in[2,t]} \theta_s(\widehat{\mathcal{G}}_s)}\Bigg\}\,,\end{aligned} \tag{C.20}$$

where $\widehat{\mathcal{G}}_t$ minimizes θ_t, which is defined in Equation (C.26). The term $\theta_t(\widehat{\mathcal{G}}_t)$ is an upper bound the second-order term in the regret bound of Exponential Weights. Crucially, the same term $\theta_t(\widehat{\mathcal{G}}_t)$ does not require us to compute a weighted independence number at each round: we can explicitly compute it in $O(K^4)$ time. Furthermore, in Lemma C.10 we show that, conditioning on the event $\mathcal{K}$, the term $\theta_t(\widehat{\mathcal{G}}_t)$ is upper bounded by the minimum thresholded weighted independence number of $\mathcal{G}$, which in turn is useful when bounding the regret. We recall that the event $\mathcal{K}$, introduced in Section 5.5, corresponds to the event that

$$|\widetilde{p}_t(j,i) - p(j,i)| \leq \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2), \quad \forall (j,i) \in V \times V$$

for all $t \geq 2$ simultaneously.

Similarly, Λ_t is an upper bound on the regret of Algorithm C.1 *if* it were to switch regime in round t and is given by

$$\Lambda_t = \min_{\varepsilon}\left\{ 41T^{2/3}\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\widehat{\mathcal{G}}_t]_\varepsilon)\right)^{1/3} + 41\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}_t]_\varepsilon)T}\right\}\,, \tag{C.21}$$

where $\widehat{\mathcal{G}}_t \coloneqq \{\widetilde{p}_t(j,i)\mathbb{I}\{\widetilde{p}_t(j,i) \geq 60\ln(KT)/t\} : i,j \in V\}$. In other words, Algorithm C.1 changes regime whenever it thinks that the regret of a (weakly) observable graph is smaller than the regret of a strongly observable graph. In the following, we prove that Ψ_t and Λ_t are indeed upper bounds on the regret, but first we state Lemma C.5, which is a central result in this section. More precisely, it provides an upper bound for the cost of not using the exact edge probabilities $p(j,i)$ but instead using upper confidence bound estimates $\widehat{p}_t(j,i)$. Note that the bound scales with $\bar{\pi}_t(i) \coloneqq \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)$. For $[\mathcal{G}]_\varepsilon$ having a strongly observable support, this is an important property of the bound since we require that $\bar{\pi}_t(i) \leq 1 - \pi_t(i)$ for vertices i without a self-loop in $\operatorname{supp}([\mathcal{G}]_\varepsilon)$ to ensure that we can bound the regret in terms of the weighted independence number.

Lemma C.5. *Define $\bar{\pi}_t(i) \coloneqq \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)$. For any distribution u over $[K]$ and $t^\star \leq T$, with* *estimator* (5.6) *we have that*

$$\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \le 2 + \sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right] + \mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right].$$

Proof. For $t > 1$, by the empirical Bernstein bound (Audibert, Munos, and Szepesvári, 2007, Theorem 1), with probability at least $1-\frac{1}{K^2T^2}$ we have that

$$\begin{aligned}\left|\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}\{(j,i)\in E_s\}-p(j,i)\right| &\le \sqrt{2\frac{\overline{\sigma}_t^2\ln(3K^2T^2)}{t-1}}+\frac{3}{t-1}\ln(3K^2T^2)\\ &\le \sqrt{\frac{2\widehat{p}_t(j,i)}{t-1}\ln(3K^2T^2)}+\frac{3}{t-1}\ln(3K^2T^2)\,,\end{aligned} \tag{C.22}$$

where we used the fact that

$$\overline{\sigma}_t^2=\frac{1}{t-1}\sum_{s'=1}^{t-1}\left(\mathbb{I}\{(j,i)\in E_{s'}\}-\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}\{(j,i)\in E_s\}\right)^2\le\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}\{(j,i)\in E_s\}\le\widehat{p}_t(j,i)\,.$$

Thus, by the union bound over K^2 edges and $t^\star$ rounds, we have that equation (C.22) holds for all edges and time steps $t \ge 2$ with probability at least $1-\frac{1}{T}$. This means that $\mathbb{P}(\mathcal{K}) \ge 1-\frac{1}{T}$ by definition of $\mathcal{K}$.

By using the tower rule and the fact that $\ell_t(i)\in[0,1]$, we can see that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right]\\ &=\mathbb{P}(\overline{\kappa})\,\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\,\middle|\,\overline{\kappa}\right]+(1-\mathbb{P}(\mathcal{K}))\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\,\middle|\,\mathcal{K}\right]\\ &\le\mathbb{P}(\overline{\kappa})\,T+\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\;\middle|\;\mathcal{K}\right]\\ &\le 2+\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\;\middle|\;\mathcal{K}\right]\;.\end{aligned} \tag{C.23}$$

Let $X_t \coloneqq \mathbb{I}\big\{i\in N^{\text{out}}_{G_t}(I_t)\wedge i\in N^{\text{out}}_{\widehat{G}_t}(I_t)\big\}$ be the indicator of the event that i belongs to both $N^{\text{out}}_{G_t}(I_t)$ and $N^{\text{out}}_{\widehat{G}_t}(I_t)$, and let $\xi_t(i)\coloneqq\widehat{P}_t(i)-P_t(i)=\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)}\pi_t(j)(\widehat{p}_t(j,i)-p(j,i))$. We continue by applying Lemma C.12 on the expectation in the right-hand side of Equation (C.23), obtaining that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\;\middle|\;\mathcal{K}\right]\\ &=\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]+\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right]\end{aligned}$$

$$\leq \mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right] + \mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}\pi_t(i)\xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right] ,$$

where the inequality is due to the fact that the loss is non-negative and the fact that $\xi_t(i) > 0$ because $\widehat{p}_t(j,i) - p(j,i) > 0$ is true, given $\mathcal{K}$. We already know that $\widehat{p}_t(j,i) \geq \widetilde{p}_t(j,i)$ by definition of $\widehat{p}_t(j,i)$. As long as $\mathcal{K}$ holds, we also know that $\widetilde{p}_t(j,i) - p(j,i) \leq \sqrt{\frac{2\widetilde{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)$ is true. Then, we can use all the above observations to demonstrate that the term $\xi_t(i)$ satisfies

$$\begin{aligned}
\xi_t(i) &= \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)(\widehat{p}_t(j,i) - p(j,i)) \\
&\leq \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\left(\sqrt{\frac{2\widehat{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)\right) \\
&\quad + \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\left(\widetilde{p}_t(j,i) - p(j,i)\right) \\
&\leq 2\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\left(\sqrt{\frac{2\widehat{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)\right) \qquad \text{(C.24)}
\end{aligned}$$

By the Cauchy-Schwarz inequality, it holds that

$$\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\sqrt{a_j} = \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \sqrt{\pi_t(j)}\sqrt{\pi_t(j)a_j} \leq \sqrt{\bar{\pi}_t(i)\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)a_j}$$

with $a_j \geq 0$ for all $j \in N^{\text{in}}_{\widehat{G}_t}(i)$, where we recall that $\bar{\pi}_t(i) = \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)$. We can use this property to further bound $\xi_t(i)$ in Equation (C.24) as

$$\begin{aligned}
\xi_t(i) &\leq 2\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(j)\left(\sqrt{\frac{2\widehat{p}_t(j,i)}{t-1}\ln(3K^2T^2)} + \frac{3}{t-1}\ln(3K^2T^2)\right) \\
&\leq 2\sqrt{2\bar{\pi}_t(i)\frac{\widehat{P}_t(i)\ln(3K^2T^2)}{t-1}} + \bar{\pi}_t(i)\frac{6\ln(3K^2T^2)}{t-1} .
\end{aligned}$$

At this point, we can use the inequality for $\xi_t(i)$ to show that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}\pi_t(i)\xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right] \\
&\quad\leq \mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sum_{i=1}^{K}\pi_t(i)\frac{X_t\ell_t(i)\sqrt{\bar{\pi}_t(i)}}{P_t(i)\sqrt{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right] \\
&\qquad + \sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\pi_t(i)\bar{\pi}_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right] \\
&\quad\leq \mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sum_{i=1}^{K}\pi_t(i)\sqrt{\frac{\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right] \qquad \text{(C.25)}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\bar{\pi}_t(i)\pi_t(i)}{\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right]\\
\le\;& \mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right]\\
&+\sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right],
\end{aligned}
$$

where in the second inequality we used the fact that $\ell_t(i)\le 1$ and that $\mathbb{E}_{t-1}[X_t]=P_t(i)$, while the final inequality is Jensen's inequality for concave functions.

By combining the above, we may complete the proof:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] &\le 2+\sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right]\\
+\mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right] &+\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]. \qquad\square
\end{aligned}
$$

C.4.1 Initial Regime of OTCG

To understand the initial regime of OTCG (Algorithm C.1), consider the following. Since the support of $\widehat{\mathcal{G}}_t^{\mathrm{UCB}}$ is the complete graph, there always exists a threshold ε for which $\mathrm{supp}\big([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon\big)$ is strongly observable. For ease of notation, given any stochastic feedback graph $\mathcal{G}$ with edge probabilities $p(j,i)$, we introduce

$$
P_t(i,\mathcal{G}) \coloneqq \sum_{j\in N^{\mathrm{in}}_{\mathrm{supp}(\mathcal{G})}(i)}\pi_t(j)p(j,i)\;.
$$

Denote by $\mathcal{S}$ the family of strongly observable graphs over vertices $V=[K]$; we can then define ε_t^θ as

$$
\begin{aligned}
\varepsilon_t^\theta &= \operatorname*{arg\,min}_{\varepsilon\,:\,\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}}\theta_t((\widehat{\mathcal{G}}_t^{\mathrm{UCB}})_\varepsilon)\\
&= \operatorname*{arg\,min}_{\varepsilon\,:\,\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}}\left(\frac{2}{\min_i\min_{j\in N^{\mathrm{in}}_{\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)}\widehat{p}_t(j,i)}+\sum_{i\in N^{\mathrm{in}}_{\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)}\frac{2\pi_t(i)}{P_t(i,(\widehat{\mathcal{G}}_t^{\mathrm{UCB}})_\varepsilon)}\right).
\end{aligned}
\tag{C.26}
$$

A crucial property of $\widehat{\mathcal{G}}_t$ (that is, $\widehat{\mathcal{G}}_t^{\mathrm{UCB}}$ thresholded at ε_t^θ) is that, if $\widehat{p}_t(j,i)\ge p(j,i)$ for all edges (j,i), by Lemma C.10 we have that

$$
\min_{\varepsilon\,:\,\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}}\theta_t([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)=\widetilde{O}\left(\min_{\varepsilon\,:\,\mathrm{supp}([\mathcal{G}]_\varepsilon)\in\mathcal{S}}\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\right),
$$

which is a property we will use when computing the final regret bound of Algorithm C.1. It also ensures that we can bound the cost of not knowing $p(j,i)$ in Lemma C.5 by $\min_{\varepsilon\,:\,\mathrm{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}}\theta_t([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)$, which is also important in computing the final regret bound of Algorithm C.1. We thus upper bound the regret of the initial regime of OTCG in terms of θ_t in what follows.

Lemma C.6. *For any distribution u over $[K]$, after $t^\star \le T$ rounds Algorithm C.1 guarantees*

$$\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right] \le 2 + 11(\ln(3K^2T^2))^2\,\mathbb{E}\left[\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)\;\middle|\;\mathcal{K}\right]$$
$$+\left(12\ln(K)+4\sqrt{2\ln(3K^2T^2)}\right)\mathbb{E}\left[\sqrt{t^\star\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)}\;\middle|\;\mathcal{K}\right]\;.$$

Proof. We start with an application of Lemma C.5:

$$\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \le 2+\sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\;\middle|\;\mathcal{K}\right]+\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]\;,$$

where, we recall it, $\bar{\pi}_t(i)=\sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)}\pi_t(j)$. Now, for i without a self-loop in $\widehat{G}_t$ we have that $\bar{\pi}_t(i)\le 1-\pi_t(i)$. Now, conditioning on $\mathcal{K}$, we may follow the reasoning surrounding Equation (C.28) to find that

$$\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\le\theta_t(\widehat{\mathcal{G}}_t)\;.$$

We now use $\sum_{t=1}^{T}\frac{1}{t}\le\ln(T)+1$, $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\le 2\sqrt{T}$, and the above inequality to obtain that

$$\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \le 2+\sum_{t=2}^{t^\star}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\theta_t(\widehat{\mathcal{G}}_t)\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=2}^{t^\star}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\theta_t(\widehat{\mathcal{G}}_t)}\;\middle|\;\mathcal{K}\right]+\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]$$
$$\le 2+6(\ln(3K^2T^2))^2\,\mathbb{E}\left[\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)\;\middle|\;\mathcal{K}\right]+\mathbb{E}\left[4\sqrt{2\ln(3K^2T^2)t^\star\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)}\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]\;.$$

By applying Lemma C.7, we can complete the proof:

$$\mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \le 2+6(\ln(3K^2T^2))^2\,\mathbb{E}\left[\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[4\sqrt{2\ln(3K^2T^2)t^\star\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)}\;\middle|\;\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[7\ln(K)\sqrt{\sum_{t=2}^{t^\star}\theta_t(\widehat{\mathcal{G}}_t)+\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)}\;\middle|\;\mathcal{K}\right]$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\max_{t\in[2,t^\star]} \frac{4\ln(K)}{p_t^{\min}} + 5\ln(K)\sqrt{\max_{t\in[2,t^\star]} \frac{t^\star}{p_t^{\min}}} \;\middle|\; \mathcal{K}\right] \\
&\le 2 + 11(\ln(3K^2T^2))^2\, \mathbb{E}\left[\max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t) \;\middle|\; \mathcal{K}\right] \\
&\quad + \left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\mathbb{E}\left[\sqrt{t^\star \max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t)} \;\middle|\; \mathcal{K}\right],
\end{aligned}
$$

where we used that $\frac{1}{p_t^{\min}} \le \theta_t(\widehat{\mathcal{G}}_t)$ for all $t \in [2, t^\star]$. □

In the proof of Lemma C.6 we make use of the following auxiliary result, which bounds the regret of π_t given $\mathcal{K}$.

Lemma C.7. *For any distribution u over $[K]$, after $t^\star \le T$ rounds Algorithm C.1 guarantees*

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(\pi_t(i) - u(i))\widetilde{\ell}_t(i) \;\middle|\; \mathcal{K}\right] &\le \mathbb{E}\left[7\ln(K)\sqrt{\sum_{t=2}^{t^\star}\theta_t(\widehat{\mathcal{G}}_t)} + \max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t) \;\middle|\; \mathcal{K}\right] \\
&\quad + \mathbb{E}\left[\max_{t\in[2,t^\star]} \frac{4\ln(K)}{p_t^{\min}} + 5\ln(K)\sqrt{\max_{t\in[2,t^\star]} \frac{t^\star}{p_t^{\min}}} \;\middle|\; \mathcal{K}\right].
\end{aligned}
$$

Proof. We want to apply Lemma C.11, which bounds the regret of Exponential Weights. Recall that Algorithm C.1 defines

$$
p_t^{\min} = \min_{i\in V} \min_{j\in N^{\text{in}}_{\text{supp}(\widehat{\mathcal{G}}_t)}(i)} \widehat{p}_t(j,i)
$$

as the minimum (positive) edge probability in $\widehat{\mathcal{G}}_t$. Observe that for any node i without a self-loop in $\text{supp}(\widehat{\mathcal{G}}_t)$ we have that

$$
\begin{aligned}
\widehat{P}_t(i) &= \sum_{j\ne i}\widehat{p}_t(j,i)\left((1-\gamma_t)q_t(i) + \frac{\gamma_t}{K}\right) \\
&\ge p_t^{\min}\sum_{j\ne i}\left((1-\gamma_t)q_t(i) + \frac{\gamma_t}{K}\right) \\
&= (1-\pi_t(i))p_t^{\min} \qquad \text{(C.27)} \\
&= \left(1-(1-\gamma_t)q_t(i) - \frac{\gamma_t}{K}\right)p_t^{\min} \\
&\ge \frac{\gamma_t}{2}p_t^{\min}\,.
\end{aligned}
$$

Using (C.27) and the definitions of η_{t-1} and γ_t, together with the fact that $\ell_t(i) \in [0,1]$, we can see that

$$
\eta_{t-1}\widetilde{\ell}_t(i) \le \eta_{t-1}\frac{1}{\widehat{P}_t(i)} \le \eta_{t-1}\frac{2}{\gamma_t p_t^{\min}} \le 1\,,
$$

where the last inequality is due to the fact that $\eta_{t-1} \le \frac{1}{2}\gamma_t p_t^{\min}$. Given event $\mathcal{K}$, since for any node i without a self-loop in $\text{supp}(\widehat{\mathcal{G}}_t)$ we have that $\eta_{t-1}\widetilde{\ell}_t(i) \le 1$, we may apply Lemma C.11 with

$S_t = S = \{i : i \notin N^{\text{in}}_{\text{supp}(\widehat{\mathcal{G}}_t)}(i)\}$ to obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(q_t(i) - u(i))\widetilde{\ell}_t(i) \;\middle|\; \mathcal{K}\right] \\
&\quad\le \mathbb{E}\left[\frac{\ln K}{\eta_{t^\star}} + \sum_{t=2}^{t^\star}\eta_{t-1}\left(\sum_{i\in S_t} q_t(i)(1-q_t(i))\widetilde{\ell}_t(i)^2 + \sum_{i\notin S_t} q_t(i)\widetilde{\ell}_t(i)^2\right) \;\middle|\; \mathcal{K}\right] .
\end{aligned}$$

We now bound

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in S_t} q_t(i)(1-q_t(i))\widetilde{\ell}_t(i)^2 \;\middle|\; \mathcal{K}\right] &= \mathbb{E}\left[\sum_{i\in S_t} q_t(i)(1-q_t(i))\frac{P_t(i)\ell_t(i)^2}{\widehat{P}_t(i)(P_t(i)+\xi_t(i))} \;\middle|\; \mathcal{K}\right] \\
&\le \mathbb{E}\left[\sum_{i\in S_t} q_t(i)\frac{(1-q_t(i))}{\widehat{P}_t(i)} \;\middle|\; \mathcal{K}\right] \\
&= \mathbb{E}\left[\sum_{i\in S_t} \frac{q_t(i)(1-q_t(i))}{P_t(i,\widehat{\mathcal{G}}_t)} \;\middle|\; \mathcal{K}\right] \\
&\le \mathbb{E}\left[\sum_{i\in S_t} \frac{2q_t(i)}{p_t^{\min}} \;\middle|\; \mathcal{K}\right] \le \mathbb{E}\left[\frac{2}{p_t^{\min}} \;\middle|\; \mathcal{K}\right] .
\end{aligned}$$

For $i \notin S_t$, since $\pi_t(i) \ge \frac{1}{2}q_t(i)$ and $\widehat{P}_t(i) - P_t(i) \ge 0$ given $\mathcal{K}$, we have that

$$\mathbb{E}\left[\sum_{i\notin S_t} q_t(i)\widetilde{\ell}_t(i)^2 \;\middle|\; \mathcal{K}\right] \le \mathbb{E}\left[\sum_{i\notin S_t} \frac{q_t(i)}{\widehat{P}_t(i)} \;\middle|\; \mathcal{K}\right] \le \mathbb{E}\left[\sum_{i\notin S_t} \frac{2\pi_t(i)}{P_t(i,\widehat{\mathcal{G}}_t)} \;\middle|\; \mathcal{K}\right] ,$$

which combined with the preceding inequality means that, given $\mathcal{K}$, we have that

$$\sum_{i\in S_t}\frac{q_t(i)(1-q_t(i))}{\widehat{P}_t(i)} + \sum_{i\notin S_t}\frac{q_t(i)}{\widehat{P}_t(i)} \le \frac{2}{p_t^{\min}} + \sum_{i\notin S_t}\frac{2\pi_t(i)}{P_t(i,\widehat{\mathcal{G}}_t)} = \theta_t(\widehat{\mathcal{G}}_t) . \tag{C.28}$$

Therefore, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K}(q_t(i) - u(i))\widetilde{\ell}_t(i) \;\middle|\; \mathcal{K}\right] \\
&\quad\le \mathbb{E}\left[\frac{\ln K}{\eta_{t^\star}} + \sum_{t=2}^{t^\star}\eta_{t-1}\left(\sum_{i\in S_t} \frac{q_t(i)(1-q_t(i))}{P_t(i,\widehat{\mathcal{G}}_t)} + \sum_{i\notin S_t} \frac{q_t(i)}{P_t(i,\widehat{\mathcal{G}}_t)}\right) \;\middle|\; \mathcal{K}\right] \\
&\quad\le \mathbb{E}\left[\frac{\ln K}{\eta_{t^\star}} + \sum_{t=2}^{t^\star}\eta_{t-1}\theta_t(\widehat{\mathcal{G}}_t) \;\middle|\; \mathcal{K}\right] .
\end{aligned}$$

Now, using a slightly modified version of (Gaillard, Stoltz, and Van Erven, 2014, Lemma 14) (replacing $|a_i| \le 1$ by $|a_i| \le \max_i |a_i|$) we can see that

$$\sum_{t=2}^{t^\star}\eta_{t-1}\theta_t(\widehat{\mathcal{G}}_t) \le \sum_{t=2}^{t^\star}\theta_t(\widehat{\mathcal{G}}_t)\sqrt{1 + \sum_{s=2}^{t-1}\theta_s(\widehat{\mathcal{G}}_s)}$$

$$\leq 3\sqrt{\sum_{t=2}^{t^\star} \theta_t(\widehat{\mathcal{G}}_t) + \max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t)} \; .$$

As a final step in this proof, we want to consider the distribution π_t the algorithm actually samples actions from instead of q_t. We can bound $\sum_{t=2}^{t^\star} \gamma_t \leq 2\sqrt{\max_{t\in[2,t^\star]} \frac{t^\star}{p_t^{\min}}}$ and

$$\frac{1}{\eta_{t^\star}} \leq \frac{4}{\min_{t\in[2,t^\star]} p_t^{\min}} + \sqrt{\frac{t^\star}{\min_{t\in[2,t^\star]} p_t^{\min}}} + \sqrt{\sum_{t=2}^{t^\star} \theta_t(\widehat{\mathcal{G}}_t)} \; .$$

Thus, combining the above we find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K} (\pi_t(i) - u(i))\widetilde{\ell}_t(i) \;\middle|\; \mathcal{K}\right] &\leq \mathbb{E}\left[\sum_{t=2}^{t^\star}\sum_{i=1}^{K} (q_t(i) - u(i))\widetilde{\ell}_t(i) + \sum_{t=2}^{t^\star} \gamma_t \;\middle|\; \mathcal{K}\right] \\
&\leq \mathbb{E}\left[7\ln(K)\sqrt{\sum_{t=2}^{t^\star} \theta_t(\widehat{\mathcal{G}}_t) + \max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t)} \;\middle|\; \mathcal{K}\right] \\
&\quad + \mathbb{E}\left[\max_{t\in[2,t^\star]} \frac{4\ln(K)}{p_t^{\min}} + 5\ln(K)\sqrt{\max_{t\in[2,t^\star]} \frac{t^\star}{p_t^{\min}}} \;\middle|\; \mathcal{K}\right] \; . \qquad \square
\end{aligned}$$

C.4.2 Regret After Round $t^\star$

With Lemma C.6 at hand, we can control the regret in the first $t^\star$ rounds. However, we also need to control the regret in the remaining rounds, which we show how to do here. Recall that $\widetilde{\mathcal{G}}$ is the graph with edge probabilities $\widetilde{p}(j,i) \coloneqq \frac{1}{t^\star}\sum_{s=1}^{t^\star} \mathbb{I}\{(j,i) \in E_s\}$. At the end of round $t^\star$ we have that $\widehat{\mathcal{G}} = [\widetilde{\mathcal{G}}]_{\varepsilon_{t^\star}}$ is an $\varepsilon_{t^\star}$-good approximation of $\mathcal{G}$ with high probability, where

$$\varepsilon_{t^\star} \coloneqq \frac{60\ln(KT)}{t^\star} \; . \tag{C.29}$$

We set

$$\varepsilon^\star_{\delta,\sigma} \coloneqq \mathop{\arg\min}_{\varepsilon\,:\,\mathrm{supp}\left([\widehat{\mathcal{G}}]_\varepsilon\right) \text{ observable}} \left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_\varepsilon)\ln(3K^2T^2)\right)^{1/3} T^{2/3} + \sqrt{\sigma([\widehat{\mathcal{G}}]_\varepsilon)T\ln(3K^2T^2)} \tag{C.30}$$

and define the corresponding stochastic graph by $[\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}} = \left\{\widetilde{p}(j,i)\mathbb{I}\{\widetilde{p}(j,i) \geq \varepsilon^\star_{\delta,\sigma}\} : i,j \in V\right\}$. We denote its support by $\widehat{G}^\star \coloneqq \mathrm{supp}\left([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}\right)$. We also require any estimated minimum weight weakly dominating set in round t, given by

$$D_t^\star \coloneqq \mathop{\arg\min}_{D\in\mathcal{D}(\widehat{G}^\star)} \sum_{i\in D} \frac{1}{\min_{j\in N^{\mathrm{out}}_{\widehat{G}^\star}(i)} \widehat{p}_t(i,j)} \; ,$$

where $\mathcal{D}(\widehat{G}^\star)$ corresponds to the family of weakly dominating sets in $\widehat{G}^\star$. We define

$$\psi_t(i) \propto \begin{cases} \left(\min_{j\in N^{\mathrm{out}}_{\widehat{G}^\star}(i)} \widehat{p}_t(i,j)\right)^{-1} & \text{for } i \in D_t^\star \\ 0 & \text{for } i \notin D_t^\star \end{cases} \tag{C.31}$$

to be the exploration distribution in round t. Note that this distribution is non-uniform over the weakly dominating set $D^\star_t$. This is because we want to ensure that the loss of each node is observed roughly equally often. If we were to sample uniformly at random, then this would not be possible because the probability that an edge realizes is not necessarily identical for all edges; however, note that the distribution is in fact uniform if the estimated edge probabilities are uniform.

Lemma C.8. *Suppose that $\widehat{\mathcal{G}}$ is an $\varepsilon_{t^\star}$-good approximation of $\mathcal{G}$. For any distribution u over $[K]$, Algorithm C.1 guarantees*

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\,\middle|\,\mathcal{K}\right]\\&\le 16\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)+5(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2))^{1/3}T^{2/3}+4\sqrt{\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})T\ln(K)}\,.\end{aligned}$$

Proof. Consider the set $S \coloneqq \{i : i \notin N^{\text{in}}_{\widehat{G}^\star}(i)\}$ of nodes without a self-loop in $\widehat{G}^\star$. Observe that for any node $i \in S$, given $\mathcal{K}$, we have that for some node $k \in D^\star_t$ with $t > t^\star$,

$$\begin{aligned}\widehat{P}_t(i) &= \sum_{j\neq i}\widehat{p}_t(j,i)\left((1-\gamma)q_t(i)+\gamma\psi_t(i)\right)\\&\ge \gamma\widehat{p}_t(k,i)\psi_t(k)\\&\ge \frac{\gamma}{\sum_{k\in D^\star_t}\left(\min_{j\in N^{\text{out}}_{\widehat{G}^\star}(k)}\widehat{p}_t(k,j)\right)^{-1}}\,.\end{aligned}$$

Observe that $\mathbb{E}\left[\widehat{p}_t(j,i)\mid\mathcal{K}\right]\ge p(j,i)\ge\frac{1}{2}\widetilde{p}(j,i)$ for all edges (j,i) in $\widehat{G}^\star$ by definition of $\varepsilon_{t^\star}$-good approximation. This implies that

$$\widehat{P}_t(i)\ge\frac{\gamma}{2\delta_{\mathsf{w}}(\widehat{\mathcal{G}}_{\varepsilon^\star_{\delta,\sigma}})} \tag{C.32}$$

holds for any node $i \in S$, conditioning on $\mathcal{K}$. We apply Lemma C.11 with $S_t = \emptyset$ to obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(q_t(i)-u(i))\widetilde{\ell}_t(i)\,\middle|\,\mathcal{K}\right]&\le\mathbb{E}\left[\frac{\ln K}{\eta}+\sum_{t=t^\star+1}^{T}\eta\sum_{i=1}^{K}q_t(i)\widetilde{\ell}_t(i)^2\,\middle|\,\mathcal{K}\right]\\&\le\mathbb{E}\left[\frac{\ln K}{\eta}+\sum_{t=t^\star+1}^{T}\eta\sum_{i=1}^{K}\frac{q_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right]\,,\end{aligned}$$

where we used the fact that $\widehat{P}_t(i)-P_t(i)\ge 0$, given $\mathcal{K}$. Recalling Equation (C.32) and using the fact that $\pi_t(i)\ge\frac{1}{2}q_t(i)$, we can see that

$$\mathbb{E}\left[\sum_{i\in S}\frac{q_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right]\le\mathbb{E}\left[\sum_{i\in S}\frac{2\pi_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right]\le\mathbb{E}\left[\frac{4\delta_{\mathsf{w}}(\widehat{\mathcal{G}}_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}\,\middle|\,\mathcal{K}\right]\,.$$

Considering the sum over $i \notin S$, we have

$$\mathbb{E}\left[\sum_{i\notin S}\frac{q_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right]\le\mathbb{E}\left[\sum_{i\notin S}\frac{2\pi_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right.$$

$$\leq \mathbb{E}\left[\sum_{i \notin S} \frac{2}{\widehat{p}_t(i,i)} \,\middle|\, \mathcal{K}\right] \leq \mathbb{E}\left[\sum_{i \notin S} \frac{4}{\widetilde{p}(i,i)} \,\middle|\, \mathcal{K}\right] \leq 4\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \;.$$

Thus, we have that

$$\mathbb{E}\left[\sum_{i=1}^{K} \frac{q_t(i)}{\widehat{P}_t(i)} \,\middle|\, \mathcal{K}\right] \leq 4\mathbb{E}\left[\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma} + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \,\middle|\, \mathcal{K}\right] \;, \tag{C.33}$$

which means that we can use $\eta \coloneqq \sqrt{\frac{\ln(K)}{4T}\big(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})/\gamma + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\big)^{-1}}$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=t^\star+1}^{T} \sum_{i=1}^{K} (\pi_t(i) - u(i))\widetilde{\ell}_t(i) \,\middle|\, \mathcal{K}\right] &\leq \mathbb{E}\left[\sum_{t=t^\star+1}^{T} \sum_{i=1}^{K} (q_t(i) - u(i))\widetilde{\ell}_t(i) \,\middle|\, \mathcal{K}\right] + \gamma T \\
&\leq \frac{\ln K}{\eta} + 4\eta T\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma} + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right) + \gamma T \\
&= 4\sqrt{T\ln(K)\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma} + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)} + \gamma T \;.
\end{aligned}$$

Now, observe that $T \leq 8\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(3K^2T^2)$ whenever the algorithm's parameter satisfies

$$\gamma = \min\left\{\big(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(3K^2T^2)\big)^{1/3} T^{-1/3}, \frac{1}{2}\right\} = \frac{1}{2} \;.$$

As a consequence,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=t^\star+1}^{T} \sum_{i=1}^{K} (\pi_t(i) - u(i))\widetilde{\ell}_t(i) \,\middle|\, \mathcal{K}\right] \\
&\leq 4\sqrt{T\ln(K)\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma} + \sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)} + \gamma T \\
&\leq 16\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(3K^2T^2) + 5(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(3K^2T^2))^{1/3}T^{2/3} + 4\sqrt{\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})T\ln(K)} \;,
\end{aligned}$$

which completes the proof. □

For the following lemma, we will use a simplifying assumption on T: we will assume that T is such that

$$\begin{aligned}
&2 + \big(37\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) + 12\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\big) \ln(3K^2T^2)^2 + 12\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})^{2/3} \big(\ln(3K^2T^2)\big)^{5/3} T^{1/3} \\
&\leq 28\left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}}) \ln(3K^2T^2)\right)^{1/3} T^{2/3} + 29\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})T} \;.
\end{aligned} \tag{C.34}$$

Lemma C.9. *Suppose that Equation* (C.34) *holds and that $\widehat{\mathcal{G}}$ is an $\varepsilon_{t^\star}$-good approximation of $\mathcal{G}$. For any distribution u over $[K]$, Algorithm C.1 guarantees*

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T} \sum_{i=1}^{K} (\pi_t(i) - u(i))\ell_t(i)\right]$$

$$\leq 41\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)^{1/3}T^{2/3} + 41\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})T}\,.$$

We also have that

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq$$
$$\min_{\varepsilon\geq 2\varepsilon_{t^\star}}\left\{82\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\mathcal{G}]_{\varepsilon})\right)^{1/3}T^{2/3} + 82\sqrt{\ln(3K^2T^2)\sigma([\mathcal{G}]_{\varepsilon})T}\ :\ \operatorname{supp}\left([\mathcal{G}]_{\varepsilon}\right)\ \text{observable}\right\}.$$

Proof. Following the proof of Lemma C.5, we can see that

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq 2 + \sum_{t=t^\star+1}^{T}\mathbb{E}\left[\frac{6\ln(3K^2T^2)}{t-1}\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}\,\middle|\,\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=t^\star+1}^{T}2\sqrt{2\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\sum_{i=1}^{K}\frac{\pi_t(i)\bar{\pi}_t(i)}{\widehat{P}_t(i)}}\,\middle|\,\mathcal{K}\right] + \mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\,\middle|\,\mathcal{K}\right].$$

Now, using the same reasoning that led to Equation (C.33), we have that

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq 2 + \sum_{t=t^\star+1}^{T}\mathbb{E}\left[\frac{12\ln(3K^2T^2)}{t-1}\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}+\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)\,\middle|\,\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=t^\star+1}^{T}4\sqrt{\frac{\ln(3K^2T^2)}{t-1}}\sqrt{\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}+\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}\,\middle|\,\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\,\middle|\,\mathcal{K}\right]$$
$$\leq 2+\mathbb{E}\left[12\ln(3K^2T^2)^2\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}+\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)\,\middle|\,\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[8\sqrt{T\ln(3K^2T^2)\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}+\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)}\,\middle|\,\mathcal{K}\right]$$
$$+\,\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\,\middle|\,\mathcal{K}\right],$$

where we used that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\leq 2\sqrt{T}$ and $\sum_{t=2}^{T}\frac{1}{t-1}\leq 1+\ln(T)\leq\ln(3K^2T^2)$ for $K,T\geq 2$. Following the final steps in the proof of Lemma C.8, we can show that

$$\mathbb{E}\left[8\sqrt{T\ln(3K^2T^2)\left(\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}+\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\right)}\,\middle|\,\mathcal{K}\right]$$
$$\leq 32\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2) + 8T^{2/3}\left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)\right)^{1/3} + 8\sqrt{T\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)}\,.$$

Hence, by applying Lemma C.8, we obtain that

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\widetilde{\ell}_t(i)\;\middle|\;\mathcal{K}\right]$$
$$\le 16\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)+5T^{2/3}(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2))^{1/3}+4\sqrt{T\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(K)}\;.$$

Finally, by definition of γ we notice that

$$12\ln(3K^2T^2)^2\frac{\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})}{\gamma}\le 24\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)^2+12T^{1/3}\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})^{2/3}\left(\ln(3K^2T^2)\right)^{5/3}\;.$$

Thus, combining the above we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right]\\&\le 2+37\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)^2+12T^{1/3}\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})^{2/3}\left(\ln(3K^2T^2)\right)^{5/3}\\&\quad+13T^{2/3}\left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)\right)^{1/3}+12\sqrt{T\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)}\\&\quad+12\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)^2\;.\end{aligned}$$

Since we assumed that Equation (C.34) holds, we can show that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right]\\&\le 41T^{2/3}\left(\delta_{\mathsf{w}}([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)\right)^{1/3}+41\sqrt{T\sigma([\widehat{\mathcal{G}}]_{\varepsilon^\star_{\delta,\sigma}})\ln(3K^2T^2)}\;,\end{aligned}$$

which is the first result in the statement. For the second result, recall that $\varepsilon^\star_{\delta,\sigma}$ is the minimizer of the above bound by its definition in Equation (C.30). Since $\widehat{\mathcal{G}}$ is an $\varepsilon_{t^\star}$-good approximation of $\mathcal{G}$, we conclude that

$$\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right]\le$$
$$\min_{\varepsilon\ge 2\varepsilon_{t^\star}}\left\{82T^{2/3}\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\right)^{1/3}+82\sqrt{\ln(3K^2T^2)\sigma([\mathcal{G}]_\varepsilon)T}\;:\;\mathrm{supp}\left([\mathcal{G}]_\varepsilon\right)\text{ observable}\right\}\;.$$

□

C.4.3 Regret After T Rounds

We now have all the intermediate results we need to prove the overall regret bound of Algorithm C.1.

Theorem C.5. *Suppose that* (C.34) *holds. Then, for any distribution u over $[K]$, Algorithm C.1 satisfies*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right]\le\min\Bigg\{T\;,$$

$$
\left. 6 + 2 \min_{\varepsilon\,:\,\mathrm{supp}([\mathcal{G}]_\varepsilon)\text{ strongly observable}} \left\{ 198\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)(\ln(2K^3T^2))^3 \right.\right.
$$

$$
\left.\left. + \left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\sqrt{18t^\star\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2)} \right\} ,\right.
$$

$$
\left. 4 + 164\ln(3K^2T^2) \min_{\varepsilon\,:\,\mathrm{supp}([\mathcal{G}]_\varepsilon)\text{ observable}} \left((\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon))^{1/3}T^{2/3} + \sqrt{\sigma([\mathcal{G}]_\varepsilon)T} \right) \right\} .
$$

Proof. Let us recall that in Equations (C.20) and (C.21) we define

$$
\Psi_{t^\star} = \min\left\{ t^\star,\, 2 + 11(\ln(3K^2T^2))^2 \max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t) \right.
$$

$$
\left. + \left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\sqrt{t^\star \max_{t\in[2,t^\star]} \theta_t(\widehat{\mathcal{G}}_t)} \right\}
$$

and

$$
\Lambda_{t^\star} = 41\left(\ln(3K^2T^2)\delta_{\mathsf{w}}(\widehat{\mathcal{G}}_{\varepsilon^\star_{\delta,\sigma}})\right)^{1/3} T^{2/3} + 41\sqrt{\ln(3K^2T^2)\sigma(\widehat{\mathcal{G}}_{\varepsilon^\star_{\delta,\sigma}})T} .
$$

Denote by $\mathcal{E}$ the event that $[\widetilde{\mathcal{G}}]_{\varepsilon_t}\{\widetilde{p}_t(j,i)\mathbb{I}\{\widetilde{p}_t(j,i) \geq 60\ln(KT)/t\} : i,j \in V\}$ is a ε_t-good approximation of $\mathcal{G}$ with $\varepsilon_t \coloneqq 60\ln(KT)/t$ for all $t \leq T$. By Lemma C.13, we have that $\mathcal{E}$ occurs with probability at least $1 - \frac{1}{T}$ and thus, for any $t^\star \in [1,T]$, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i) - u(i))\ell_t(i)\right] \\
&\quad\leq 1 + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i) - u(i))\ell_t(i) \;\middle|\; \mathcal{E}\right] \\
&\quad= 1 + \mathbb{E}\left[\sum_{t=1}^{t^\star}\sum_{i=1}^{K}(\pi_t(i) - u(i))\ell_t(i) \;\middle|\; \mathcal{E}\right] + \mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i) - u(i))\ell_t(i) \;\middle|\; \mathcal{E}\right] \\
&\quad\leq 1 + \mathbb{E}\left[\Psi_{t^\star} + \Lambda_{t^\star} \mid \mathcal{K}, \mathcal{E}\right] ,
\end{aligned}
$$

where the last inequality is due to Lemmas C.6 and C.9. We now consider two cases depending on whether Algorithm C.1 commits to the weakly observable regret regime at any time step or it never does so. In the first case, say Equation (C.19) never holds for any $t \in [2,T]$. We consequently have that

$$
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i) - u(i))\ell_t(i)\right] \leq 1 + \mathbb{E}\left[\min\left\{\Psi_{t^\star}, \Lambda_{t^\star}\right\} \mid \mathcal{K}, \mathcal{E}\right] .
$$

We first try to upper bound the conditional expectation of $\Lambda_{t^\star}$. By definition of ε-good approximation of $\mathcal{G}$, we have

$$
\mathbb{E}\left[\Lambda_{t^\star} \mid \mathcal{K}, \mathcal{E}\right] = \mathbb{E}\left[\min_{\varepsilon\in[0,1]}\left\{ 41\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)\right)^{1/3} T^{2/3} \right.\right.
$$

$$+ 41\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)T} \;:\; \operatorname{supp}\big([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon\big) \text{ observable}\Bigg\} \;\Bigg|\; \mathcal{K}, \mathcal{E}\Bigg]$$

$$\le 2\mathbb{E}\Bigg[\min_{\varepsilon\in[2\varepsilon_{t^\star},1]}\Bigg\{41\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)\right)^{1/3}T^{2/3}$$

$$+ 41\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)T} : \operatorname{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ observable}\Bigg\} \;\Bigg|\; \mathcal{K}, \mathcal{E}\Bigg] .$$

To cover the remaining thresholds in $[0, 2\varepsilon_{t^\star})$, we define $\varepsilon^\star_\Lambda := \max \mathcal{Q}$ as the largest threshold ε that minimizes

$$\mathcal{Q} := \operatorname*{arg\,min}_{\varepsilon\in[0,1]}\Bigg\{41\left(\ln(3K^2T^2)\delta_{\mathsf{w}}([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)\right)^{1/3}T^{2/3}$$

$$+ 41\sqrt{\ln(3K^2T^2)\sigma([\widehat{\mathcal{G}}_{t^\star}]_\varepsilon)T} \;:\; \operatorname{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ observable}\Bigg\} .$$

If $\varepsilon^\star_\Lambda < 2\varepsilon_{t^\star}$, meaning that $\varepsilon^\star_\Lambda$ as well as the other thresholds in $\mathcal{Q}$ do not belong to the already covered interval $[2\varepsilon_{t^\star}, 1]$, then $t^\star < \frac{120\ln(KT)}{\varepsilon^\star_\Lambda} = 120\ln(KT)t_{\varepsilon^\star_\Lambda}$ with $t_{\varepsilon^\star_\Lambda} := 1/\varepsilon^\star_\Lambda$. Thus, we must have that

$$t^\star \le 120\ln(KT)\left(\left(\delta_{\mathsf{w}}([\mathcal{G}]_{\varepsilon^\star_\Lambda})\right)^{1/3}t^{2/3}_{\varepsilon^\star_\Lambda} + \sqrt{\sigma([\mathcal{G}]_{\varepsilon^\star_\Lambda})t_{\varepsilon^\star_\Lambda}}\right)$$

$$\le \min_{\varepsilon\in[0,1]}\left\{120\ln(KT)\left((\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon))^{1/3}T^{2/3} + \sqrt{\sigma([\mathcal{G}]_\varepsilon)T}\right) \;:\; \operatorname{supp}\left([\mathcal{G}]_\varepsilon\right) \text{ observable}\right\} ,$$

where the first inequality is due to the fact that $\delta_{\mathsf{w}}([\mathcal{G}]_{\varepsilon^\star_\Lambda}) \ge t_{\varepsilon^\star_\Lambda}$ or $\sigma([\mathcal{G}]_{\varepsilon^\star_\Lambda}) \ge t_{\varepsilon^\star_\Lambda}$ or both are true because either $p(i,i) = \varepsilon^\star_\Lambda$ for some i such that $i \in N^{\text{in}}_{\operatorname{supp}([\mathcal{G}]_{\varepsilon^\star_\Lambda})}(i)$ or one of the minimum outgoing edge probabilities for a vertex in some minimum weight weakly dominating set is equal to $\varepsilon^\star_\Lambda$.

On the other hand, we also need to upper bound the conditional expectation of $\Psi_{t^\star}$. By Lemma C.10 and recalling the definition of α_{w} from Section 5.5, we have that

$$\mathbb{E}\left[\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t) \;\middle|\; \mathcal{K}\right] \le \mathbb{E}\left[\min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\text{ strongly observable}} 18\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2) \;\middle|\; \mathcal{K}\right] .$$

and thus

$$2 + \mathbb{E}\left[11(\ln(3K^2T^2))^2\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t) \;\middle|\; \mathcal{K}, \mathcal{E}\right]$$

$$+ \mathbb{E}\left[\left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\sqrt{t^\star\max_{t\in[2,t^\star]}\theta_t(\widehat{\mathcal{G}}_t)} \;\middle|\; \mathcal{K}, \mathcal{E}\right]$$

$$\le 2 + \min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\text{ strongly observable}}\Bigg\{198\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)(\ln(2K^3T^2))^3$$

$$+ \left(12\ln(K) + 4\sqrt{2\ln(3K^2T^2)}\right)\sqrt{18t^\star\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2)}\Bigg\} .$$

Since $120\ln(KT) \leq 82\ln(3K^2T^2)$, we can combine the above to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq \min\Bigg\{T\,, \\
&\quad 3+\min_{\varepsilon}\Bigg\{198\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)(\ln(2K^3T^2))^3 + \\
&\quad \Big(12\ln(K)+4\sqrt{2\ln(3K^2T^2)}\Big)\sqrt{18t^\star\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2)} \;:\; \operatorname{supp}([\mathcal{G}]_\varepsilon) \text{ strongly observable}\Bigg\}, \\
&\quad 1+\min_{\varepsilon}\Bigg\{82\ln(3K^2T^2)\left((\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon))^{1/3}T^{2/3}+\sqrt{\sigma([\mathcal{G}]_\varepsilon)T}\right) \;:\; \operatorname{supp}([\mathcal{G}]_\varepsilon) \text{ observable}\Bigg\}\Bigg\}\,.
\end{aligned}$$

In the second case, $t^\star$ is the first round in which Equation (C.19) holds. Therefore, we must have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq 1+2\mathbb{E}\left[\Psi_{t^\star}\,|\,\mathcal{K},\mathcal{E}\right]$$

and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \\
&\quad \leq 1+\mathbb{E}\left[\sum_{t=1}^{t^\star-1}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\,\middle|\,\mathcal{E}\right]+1+\mathbb{E}\left[\sum_{t=t^\star+1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\,\middle|\,\mathcal{E}\right] \\
&\quad \leq \mathbb{E}\left[\Psi_{t^\star-1}+\Lambda_{t^\star}\,|\,\mathcal{K},\mathcal{E}\right]+2 \\
&\quad \leq \mathbb{E}\left[\Lambda_{t^\star-1}+\Lambda_{t^\star}\,|\,\mathcal{K},\mathcal{E}\right]+2\,,
\end{aligned}$$

which combined give us

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq 1+\mathbb{E}\left[\min\left\{\Lambda_{t^\star-1}+\Lambda_{t^\star}+1, 2\Psi_{t^\star}\right\}\,|\,\mathcal{K},\mathcal{E}\right]\,.$$

Following the proof of the bound in the case where Equation (C.19) never holds for any $t \in [2,T]$, we can see that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K}(\pi_t(i)-u(i))\ell_t(i)\right] \leq \min\Bigg\{T\,, \\
&\quad 6+2\min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\text{ strongly observable}}\Bigg\{198\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)(\ln(2K^3T^2))^3 \\
&\qquad +\Big(12\ln(K)+4\sqrt{2\ln(3K^2T^2)}\Big)\sqrt{18t^\star\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2)}\Bigg\}\,, \\
&\quad 4+164\ln(3K^2T^2)\min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\text{ observable}}\left((\delta_{\mathsf{w}}([\mathcal{G}]_\varepsilon))^{1/3}T^{2/3}+\sqrt{\sigma([\mathcal{G}]_\varepsilon)T}\right)\Bigg\}\,,
\end{aligned}$$

which completes the proof. □

C.4.4 Auxiliary Lemmas for OTCG

In this section, we prove some results that are useful in the above regret analysis of OTCG (Algorithm C.1). Recall that $\mathcal{S}$ is the family of strongly observable graphs over vertices $V = [K]$.

Lemma C.10. *Suppose that there exists a threshold* ε *such that* $\operatorname{supp}([\mathcal{G}]_\varepsilon) \in \mathcal{S}$. *Then, we have that*

$$\mathbb{E}\left[\max_{t\in[2,t^\star]} \min_{\varepsilon\,:\,\operatorname{supp}\left([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon\right)\in\mathcal{S}} \theta_t([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon) \,\middle|\, \mathcal{K}\right] \le \mathbb{E}\left[\min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\in\mathcal{S}} 18\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2) \,\middle|\, \mathcal{K}\right]$$

Proof. Let us recall the definition of θ_t:

$$\theta_t([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon) = \frac{2}{\min_i \min_{j\in N^{\mathrm{in}}_{\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)} p(j,i)} + \sum_{i\in N^{\mathrm{in}}_{\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)} \frac{2\pi_t(i)}{P_t(i, [\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)} \,.$$

By definition of the weighted independence number (see Appendix C.5 for further details), we have that

$$\frac{2}{\min_i \min_{j\in N^{\mathrm{in}}_{\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)} p(j,i)} \le 2\alpha_{\mathsf{w}}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon) \,.$$

By Lemma C.16, we have that

$$2\sum_{i\in N^{\mathrm{in}}_{\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)}(i)} \frac{\pi_t(i)}{P_t(i, [\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)} \le 16\alpha_{\mathsf{w}}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\ln(2K^3T^2) \,,$$

where we used that $\gamma_t\psi_t(i) \ge \frac{1}{KT}$ and $\widehat{p}_t(j,i) \ge \frac{1}{T}$. Given $\mathcal{K}$, we have that $\widehat{p}_t(j,i) \ge p(j,i)$ and thus it holds that

$$\begin{aligned}
&\mathbb{E}\left[\max_{t\in[2,t^\star]} \min_{\varepsilon\,:\,\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}} \theta_t([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon) \,\middle|\, \mathcal{K}\right] \\
&\quad\le \mathbb{E}\left[\max_{t\in[2,t^\star]} \min_{\varepsilon\,:\,\operatorname{supp}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\in\mathcal{S}} 18\alpha_{\mathsf{w}}([\widehat{\mathcal{G}}_t^{\mathrm{UCB}}]_\varepsilon)\ln(2K^3T^2) \,\middle|\, \mathcal{K}\right] \\
&\quad\le \mathbb{E}\left[\min_{\varepsilon\,:\,\operatorname{supp}([\mathcal{G}]_\varepsilon)\in\mathcal{S}} 18\alpha_{\mathsf{w}}([\mathcal{G}]_\varepsilon)\ln(2K^3T^2) \,\middle|\, \mathcal{K}\right] \,. \qquad \square
\end{aligned}$$

The following result is a variant of the bound in Alon et al. (2015, Lemma 4) with a decreasing learning rate.

Lemma C.11. *Let* $q_1, \dots, q_T$ *be the probability vectors defined by* $q_t(i) \propto \exp(-\eta_{t-1}\sum_{s=1}^{t-1}\ell_s(i))$ *for a sequence of loss functions* $\ell_1, \dots, \ell_T$ *such that* $\ell_t(i) \ge 0$ *for all* t *and* i. *Let* $\eta_0 = \eta_1 \ge \ldots \ge \eta_T$. *For each* t, *let* S_t *be a subset of* $[K]$ *such that* $\eta_{t-1}\ell_t(i) \le 1$ *for all* $i \in S_t$. *Then, for any distribution* u *it holds that*

$$\sum_{t=1}^{T}\sum_{i=1}^{K}(q_t(i) - u(i))\ell_t(i) \le \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T}\eta_{t-1}\left(\sum_{i\in S_t} q_t(i)(1-q_t(i))\ell_t(i)^2 + \sum_{i\notin S_t} q_t(i)\ell_t(i)^2\right) .$$

Proof. The proof follows from a minor adaptation of the proof of Alon et al. (2015, Lemma 4). We start from Van der Hoeven, Van Erven, and Kotłowski (2018, Lemma 1):

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{K}&(q_t(i) - u(i))\ell_t(i) \\ &\leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T}\left(\sum_{i=1}^{K} q_t(i)\ell_t(i) + \frac{1}{\eta_{t-1}}\ln\left(\sum_{i=1}^{K} q_t(i)\exp(-\eta_{t-1}\ell_t(i))\right)\right) .\end{aligned} \tag{C.35}$$

Now, since $\ell_t(i) \geq 0$ we may use $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$ and $\ln(1-x) \leq -x$ for all $x < 1$ to show that

$$\begin{aligned}\frac{1}{\eta_{t-1}}\ln\left(\sum_{i=1}^{K} q_t(i)\exp(-\eta_{t-1}\ell_t(i))\right) &\leq \frac{1}{\eta_{t-1}}\ln\left(1 - \sum_{i=1}^{K} q_t(i)(\eta_{t-1}\ell_t(i) - \eta_{t-1}^2\ell_t(i)^2)\right) \\ &\leq -\sum_{i=1}^{K} q_t(i)(\ell_t(i) - \eta_{t-1}\ell_t(i)^2) .\end{aligned}$$

Combined with Equation (C.35), this gives us

$$\sum_{t=1}^{T}\sum_{i=1}^{K}(q_t(i) - u(i))\ell_t(i) \leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T}\sum_{i=1}^{K}\eta_{t-1}q_t(i)\ell_t(i)^2 .$$

We define $\bar{\ell}_t \coloneqq \sum_{i\in S_t} q_t(i)\ell_t(i)$. Since $\ell_t(i) \geq 0$ we have that $\eta_{t-1}(\ell_t(i) - \bar{\ell}_t) \geq -1$ by construction. Since adding the same $\bar{\ell}_t$ to each $\ell_t(i)$ on the r.h.s. of Equation (C.35) does not influence the regret we have

$$\sum_{t=1}^{T}\sum_{i=1}^{K}(q_t(i) - u(i))\ell_t(i) \leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T}\sum_{i=1}^{K}\eta_{t-1}q_t(i)(\ell_t(i) - \bar{\ell}_t)^2 .$$

To complete the proof we follow the proof of Alon et al. (2015, Lemma 4), which gives us

$$\sum_{t=1}^{T}\sum_{i=1}^{K}(q_t(i) - u(i))\ell_t(i) \leq \frac{\ln(K)}{\eta_T} + \sum_{t=1}^{T}\eta_{t-1}\left(\sum_{i\in S_t} q_t(i)(1 - q_t(i))\ell_t(i)^2 + \sum_{i\notin S_t} q_t(i)\ell_t(i)^2\right) .$$

□

Lemma C.12. *Let* $\xi_t(i) \coloneqq \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(i)(\widehat{p}_t(j,i) - p(j,i))$. *In any round* t*, we have that*

$$\begin{aligned}\sum_{i=1}^{K}&(\pi_t(i) - u(i))\widehat{\ell}_t(i) \\ &= \sum_{i=1}^{K}(\pi_t(i) - u(i))\widetilde{\ell}_t(i) + \sum_{i=1}^{K}(\pi_t(i) - u(i))\xi_t(i)\frac{\mathbb{I}\{i \in N^{\text{out}}_{G_t}(I_t) \wedge i \in N^{\text{out}}_{\widehat{G}_t}(I_t)\}\ell_t(i)}{P_t(i)\widehat{P}_t(i)} .\end{aligned}$$

Proof. Let $X_t \coloneqq \mathbb{I}\{i \in N^{\text{out}}_{G_t}(I_t) \wedge i \in N^{\text{out}}_{\widehat{G}_t}(I_t)\}$ and denote by

$$\xi_t(i) \coloneqq \widehat{P}_t(i) - P_t(i) = \sum_{j\in N^{\text{in}}_{\widehat{G}_t}(i)} \pi_t(i)(\widehat{p}_t(j,i) - p(j,i)) .$$

We have that

$$
\begin{aligned}
\widetilde{\ell}_t(i) = \frac{X_t\ell_t(i)}{\widehat{P}_t(i)} &= \frac{X_t\ell_t(i)}{P_t(i)+\xi_t(i)} \\
&= \frac{X_t\ell_t(i)(P_t(i)+\xi_t(i))}{P_t(i)(P_t(i)+\xi_t(i))} - \xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)(P_t(i)+\xi_t(i))} \\
&= \frac{X_t\ell_t(i)}{P_t(i)} - \xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)(P_t(i)+\xi_t(i))} \\
&= \widehat{\ell}_t(i) - \xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)} \; .
\end{aligned}
$$

Therefore, for any distribution u we have that

$$
\sum_{i=1}^K(\pi_t(i)-u(i))\widehat{\ell}_t(i) = \sum_{i=1}^K(\pi_t(i)-u(i))\widetilde{\ell}_t(i) + \sum_{i=1}^K(\pi_t(i)-u(i))\xi_t(i)\frac{X_t\ell_t(i)}{P_t(i)\widehat{P}_t(i)} \; ,
$$

which completes the proof. □

Lemma C.13. *Let $[\widetilde{\mathcal{G}}]_{\varepsilon_t} = \{\widetilde{p}_t(j,i)\mathbb{I}\{\widetilde{p}_t(j,i) \ge \varepsilon_t\} : i,j \in V\}$ and $\varepsilon_t := 60\ln(KT)/t$ for all $t \in [2,T]$. Then, with probability at least $1-1/T$, $[\widetilde{\mathcal{G}}]_{\varepsilon_t}$ is an ε_t-good approximation of $\mathcal{G}$ for all $t \in [2,T]$.*

Proof. Let $E_t^+ := \{(i,j) \in V^2 : p(i,j) \ge 2\varepsilon_t\}$ and $E_t^- := \{(i,j) \in V^2 : p(i,j) < \varepsilon_t/2\}$ be the two sets of edges as defined in the proof of Theorem 5.2. We let $\mathcal{E}_{(i,j)}^t := \{\widetilde{p}_t(i,j) \ge \varepsilon_t\}$ and $\mathcal{F}_{(i,j)}^t := \{|\widetilde{p}_t(i,j) - p(i,j)| \le p(i,j)/2\}$, for all $(i,j) \in V^2$ and all $t \in [2,T]$, be the events as similarly denoted in that same proof. We consequently define the events $\mathcal{E}$, $\mathcal{F}$, and $\mathcal{C}$ as

$$
\mathcal{E} = \bigcap_{t=1}^T \bigcap_{(i,j)\in E_t^+} \mathcal{E}_{(i,j)}^t \; , \qquad \mathcal{F} = \bigcap_{t=1}^T \bigcap_{(i,j)\notin E_t^-} \mathcal{F}_{(i,j)}^t \; , \qquad \mathcal{C} = \bigcap_{t=1}^T \bigcap_{(i,j)\in E_t^-} \overline{\mathcal{E}}_{(i,j)}^t \; .
$$

The following steps hold for all $K \ge 2$ and all $T \ge 2$.

We begin by observing that $\mathbb{P}\left(\widetilde{p}_t(i,j) < \varepsilon_t\right) \le \exp(-t\varepsilon_t/4) \le 1/(4K^2T^2)$ for all $t \in [2,T]$ and all $(i,j) \in E_t^+$, by a simple adaptation of the same argument in the proof of Theorem 5.2. Then,

$$
\mathbb{P}\left(\mathcal{E}\right) \ge 1 - \sum_{t=1}^T \frac{|E_t^+|}{4K^2T^2} \ge 1 - \frac{1}{4T} \; ,
$$

which follows from the fact that $|E_t^+| \le K^2$ for all $t \in [2,T]$. We can similarly argue that $\mathbb{P}\left(|\widetilde{p}_t(i,j) - p(i,j)| > p(i,j)/2\right) \le 2\exp(-t\varepsilon_t/24) \le 1/(2K^2T^2)$ for all $t \in [2,T]$ and all $(i,j) \notin E_t^-$; this implies that $\mathbb{P}\left(\mathcal{F}\right) \ge 1 - 1/(2T)$. Finally, we observe that $\mathbb{P}\left(\widetilde{p}_t(i,j) \ge \varepsilon_t\right) \le \exp(-t\varepsilon_t/6) \le 1/(4K^2T^2)$ for all $t \in [2,T]$ and all $(i,j) \in E_t^-$, hence $\mathbb{P}\left(\mathcal{C}\right) \ge 1 - 1/(4T)$. The statement follows by union bound over the complements of $\mathcal{E}$, $\mathcal{F}$, and $\mathcal{C}$. □

C.5 Weighted Independence Number

To improve the regret bounds in the case of strongly observable support, we need to introduce another graph-theoretic quantity: the *weighted independence number* $\alpha_{\mathsf{w}}(G,w)$, where $w \in \mathbb{R}_{>0}^K$ is a vector of positive weights assigned to the vertices of our strongly observable graph $G = (V,E)$ with

$V = [K]$. Let $w(U) \coloneqq \sum_{i \in U} w_i$ denote the weight of a subset of vertices $U \subseteq V$. Recall that the weighted independence number is defined as

$$\alpha_{\mathsf{w}}(G, w) \coloneqq \max_{S \in \mathcal{I}(G)} w(S) \ ,$$

that is, the weight of a maximum weight independent set. This set is chosen among all sets in the family $\mathcal{I}(G)$ of independent sets of G. It can be equivalently defined by the following integer linear program:

$$\begin{array}{rrcll} \alpha_{\mathsf{w}}(G, w) = \max\limits_{x} & \sum\limits_{i=1}^{K} w_i x_i & & & \\ \text{s.t.} & x_i + x_j & \leq & 1 & \forall (i,j) \in E, i \neq j \\ & x_i & \in & \{0,1\} & \forall i \in V \end{array}$$

We plan to define w according to our needs in what follows.

C.5.1 Undirected Graph

Let $\mathcal{G}$ be a stochastic feedback graph with edge probabilities $p(i,j)$ and such that its support $\operatorname{supp}(\mathcal{G}) = G = (V, E)$ is undirected and strongly observable. Moreover, let $N(i)$ be the neighborhood in G of any vertex $i \in V$ (excluding i) and let $C(i) \coloneqq N(i) \cup \{i\}$ be the extended neighborhood of i including vertex i itself.

We can use the edge probabilities from $\mathcal{G}$ to define a weight for each vertex i as

$$w_{\mathcal{G}}(i) \coloneqq w_i = \left(\frac{1}{|C(i)|} \sum_{j \in C(i)} p(j,i) \right)^{-1} .$$

This vertex weight is equal to the inverse of the arithmetic mean of the incident edge probabilities (including its self-loop). Note that the two probabilities $p(i,j)$ and $p(j,i)$ in the two directions of any undirected edge $(i,j) \in E$ need not be equal.

This definition allows us to upper bound the second-order term in the regret for vertices with self-loop (as similarly done in the analysis of Exp3.G (Alon et al., 2015)) in terms of the weighted independence number since we can reduce it to bounding

$$\sum_{i \in V} \frac{1}{\sum_{j \in C(i)} p(j,i)} = \sum_{i \in V} \frac{w_i}{|C(i)|} \ .$$

We thus require a weighted version of Turán's theorem, which is formulated in the lemma below. This result has already been proved (Sakai, Togasaki, and Yamazaki, 2003), but we nevertheless provide a proof for completeness.

Lemma C.14. *Let $G = (V, E)$ be an undirected graph with positive vertex weights w_i. Then,*

$$\sum_{i \in V} \frac{w_i}{|C(i)|} \leq \alpha_{\mathsf{w}}(G, w) \ .$$

Proof. Consider the following algorithm: as long as the graph is not empty, repeatedly choose a vertex j that minimizes $|C(j)|/w_j$ among all remaining vertices and remove it from the graph along with its neighborhood. Let $i_1, \dots, i_s$ be the sequence of s vertices picked by this algorithm, which form an independent set by construction. Additionally, let $G_1, \dots, G_{s+1}$ be the sequence of graphs generated by this iterative procedure, where $G_1 = G$ is the starting graph and G_{s+1} is the empty graph. We also let $C_r(i)$ denote the extended neighborhood over G_r of any $i \in V(G_r)$. Define

$$Q(H) \coloneqq \sum_{i \in V(H)} \frac{w_i}{|C(i)|} \qquad \forall H \subseteq G \ ,$$

as the quantity we are trying to bound for G and consider it over the graphs in the sequence generated by the procedure. It is strictly decreasing until reaching $Q(G_{s+1}) = 0$. In particular, at any step of the procedure it decreases by

$$Q(G_r) - Q(G_{r+1}) = \sum_{j \in C_r(i_r)} \frac{w_j}{|C(j)|} \le \sum_{j \in C_r(i_r)} \frac{w_{i_r}}{|C(i_r)|} = \frac{|C_r(i_r)|}{|C(i_r)|} w_{i_r} \le w_{i_r} \ ,$$

where the first inequality is due to the optimality of $|C(i_r)|/w_{i_r}$ at step r. We can use this inequality to bound $Q(G)$ by

$$Q(G) = \sum_{r=1}^{s} (Q(G_r) - Q(G_{r+1})) \le \sum_{r=1}^{s} w_{i_r} \le \max_{S \in \mathcal{I}(G)} w(S) = \alpha_{\mathsf{w}}(G, w) \ .$$

□

C.5.2 Directed Graph

Compared to the result in the previous section, we are more generally interested in directed graphs. We consider the case of directed, strongly observable support $\operatorname{supp}(\mathcal{G}) = G = (V, E)$ with $V = [K]$ and $(i,i) \in E$ for all $i \in V$. In the directed case, we distinguish the in-neighborhood $N^{\text{in}}(i)$ over G of a vertex $i \in V$ from its out-neighborhood $N^{\text{out}}(i)$. We use the convention that vertices with self-loops are not included in their neighborhoods, while all vertices are always included in their extended in-neighborhood $C^{\text{in}}(i) \coloneqq N^{\text{in}}(i) \cup \{i\}$ and out-neighborhood $C^{\text{out}}(i) \coloneqq N^{\text{out}}(i) \cup \{i\}$, respectively. We make this distinction to comply as much as possible with previous works providing analogous results (Alon et al., 2017), where the neighborhoods $N^{\text{in}}(i)$ and $N^{\text{out}}(i)$ did not include i even in the presence of the self-loop $(i,i) \in E$.

The weighted independence number is defined in the same way as per undirected graphs, ignoring the direction of edges for the independence condition. Here we define in two slightly different manners the vertex weights: let

$$w^{\text{in}}_{\mathcal{G}}(i) \coloneqq w^{\text{in}}_i = \left(\frac{1}{|C^{\text{in}}(i)|} \sum_{j \in C^{\text{in}}(i)} p(j,i) \right)^{-1} \tag{C.36}$$

be the inverse of the arithmetic mean of the incoming edge probabilities for i, and

$$w^{\text{out}}_{\mathcal{G}}(i) \coloneqq w^{\text{out}}_i = \left(\frac{1}{|C^{\text{out}}(i)|} \sum_{j \in C^{\text{out}}(i)} p(i,j) \right)^{-1} \tag{C.37}$$

the analogous over outgoing edges. These two different assignments of vertex weights induce two weighted independence numbers $\alpha_{\mathsf{w}}(G, w^{\text{in}})$ and $\alpha_{\mathsf{w}}(G, w^{\text{out}})$, respectively.

Then, we prove a lemma similar to (Alon et al., 2017, Lemma 13) in the weighted case. Note, however, that in this case the lemma is tightly related to the specific definitions of vertex weights we are adopting.

Lemma C.15. *Let $G = (V, E)$ be a directed graph with edge probabilities $p(i, j) \in [0, 1]$, and positive vertex weight vectors w^{in} and w^{out} as in Equations* (C.36) *and* (C.37)*, respectively. Then,*

$$\sum_{i \in V} \frac{w_i^{\text{in}}}{|C^{\text{in}}(i)|} \leq 3(\alpha_{\mathsf{w}}(G, w^{\text{in}}) + \alpha_{\mathsf{w}}(G, w^{\text{out}})) \ln(K + 1) \ .$$

Proof. We prove the statement by induction as in the proof of Alon et al. (2017, Lemma 13). Consider the following algorithm: as long as the graph is not empty, repeatedly choose the vertex j that maximizes $|C^{\text{in}}(j)|/w_j^{\text{in}}$ among all remaining vertices and remove it from the graph along with its incident edges. Let $i_1, \dots, i_K$ be the vertices in the order the algorithm picks them. Additionally, let $G_1, \dots, G_{K+1}$ be the sequence of graphs generated by this iterative procedure, where $G_1 = G$ is the original graph and G_{K+1} is the empty graph. We also let $C_r^{\text{in}}(i)$ denote the extended in-neighborhood over G_r of any $i \in V(G_r)$. Similarly to the proof of Lemma C.14, define

$$Q(H) \coloneqq \sum_{i \in V(H)} \frac{w_i^{\text{in}}}{|C^{\text{in}}(i)|} \qquad \forall H \subseteq G$$

as the quantity we want to bound for G, where the size of the in-neighborhood is always computed with respect to the starting graph G.

Define a new instance of the problem with graph $G' \coloneqq (V, E')$ as the undirected version of G, where the edge probabilities are defined as $p'(i, j) \coloneqq \frac{1}{2}p(i, j) + \frac{1}{2}p(j, i)$ for all $i, j \in V$ such that either $(i, j) \in E$ or $(j, i) \in E$. This new graph has $C(i) = C^{\text{in}}(i) \cup C^{\text{out}}(i)$. As a consequence, we can derive new vertex weights $w'_i \coloneqq \big(\frac{1}{|C(i)|} \sum_{j \in C(i)} p'(j, i)\big)^{-1}$. This instance is such that

$$\sum_{i \in V} \frac{|C(i)|}{w'_i} = \sum_{i \in V} \sum_{j \in C(i)} p'(j, i) = \sum_{i \in V} \sum_{j \in C^{\text{in}}(i)} p(j, i) = \sum_{i \in V} \frac{|C^{\text{in}}(i)|}{w_i^{\text{in}}} \ . \tag{C.38}$$

Furthermore, notice that the newly defined vertex weights satisfy

$$\begin{aligned} w'_i = \frac{|C(i)|}{\sum_{j \in C(i)} p'(j, i)} &\leq \frac{|C^{\text{in}}(i)|}{\sum_{j \in C(i)} p'(j, i)} + \frac{|C^{\text{out}}(i)|}{\sum_{j \in C(i)} p'(j, i)} \\ &\leq \frac{2|C^{\text{in}}(i)|}{\sum_{j \in C^{\text{in}}(i)} p(j, i)} + \frac{2|C^{\text{out}}(i)|}{\sum_{j \in C^{\text{out}}(i)} p(i, j)} \\ &= 2(w_i^{\text{in}} + w_i^{\text{out}}) \ . \end{aligned} \tag{C.39}$$

Consider now the first vertex i_1 chosen by the procedure we introduced before. The value it maximizes is lower bounded by

$$\begin{aligned} \max_{i \in V} \frac{|C^{\text{in}}(i)|}{w_i^{\text{in}}} &\geq \frac{1}{K} \sum_{i \in V} \frac{|C^{\text{in}}(i)|}{w_i^{\text{in}}} \\ &= \frac{1}{K} \sum_{i \in V} \frac{|C(i)|}{w'_i} & \text{by Equation (C.38)} \end{aligned}$$

$$
\begin{aligned}
&\geq \frac{K}{\sum_{i\in V}\frac{w_i'}{|C(i)|}} && \text{by Jensen's inequality}\\
&\geq \frac{K/2}{\sum_{i\in V}\frac{w_i^{\text{in}}}{|C(i)|}+\sum_{i\in V}\frac{w_i^{\text{out}}}{|C(i)|}} && \text{by Equation (C.39)}\\
&\geq \frac{K/2}{\alpha_{\mathsf{w}}(G,w^{\text{in}})+\alpha_{\mathsf{w}}(G,w^{\text{out}})}\;. && \text{by Lemma C.14 over } G' \qquad \text{(C.40)}
\end{aligned}
$$

We can use this fact to show an upper bound for the sum $Q(G)$ as

$$
\begin{aligned}
Q(G) = \sum_{i\in V}\frac{w_i^{\text{in}}}{|C^{\text{in}}(i)|} &= \frac{w_{i_1}^{\text{in}}}{|C^{\text{in}}(i_1)|} + \sum_{r=2}^{K}\frac{w_{i_r}^{\text{in}}}{|C^{\text{in}}(i_r)|}\\
&\leq \frac{2(\alpha_{\mathsf{w}}(G,w^{\text{in}})+\alpha_{\mathsf{w}}(G,w^{\text{out}}))}{K} + Q(G_2)\;. \qquad \text{by Equation (C.40)}
\end{aligned}
$$

As a last step, recursively repeat the same reasoning on $Q(G_2)$ and iterate it until reaching G_K to conclude that

$$
Q(G) \leq 2\sum_{r=1}^{K}\frac{\alpha_{\mathsf{w}}(G_r,w^{\text{in}})+\alpha_{\mathsf{w}}(G_r,w^{\text{out}})}{K-r+1} \leq 3(\alpha_{\mathsf{w}}(G,w^{\text{in}})+\alpha_{\mathsf{w}}(G,w^{\text{out}}))\ln(K+1)\;. \qquad \square
$$

We finally have all the tools required for demonstrating the next lemma. It essentially corresponds to Alon et al. (2015, Lemma 5) with the addition of edge probabilities. The main difference is that we show an upper bound in terms of two distinct independence numbers. They are both computed over the graph G with vertex weights defined in terms of the worst-case edge probabilities. To be specific, we have a first weight assignment w^- to vertices such that $w_{\mathcal{G}}^-(i) := w_i^- = \left(\min_{j\in C^{\text{in}}(i)} p(j,i)\right)^{-1}$ is the reciprocal of the minimum incoming edge probability for vertex i. The second assignment w^+, instead, assigns weight $w_{\mathcal{G}}^+(i) := w_i^+ = \left(\min_{j\in C^{\text{out}}(i)} p(i,j)\right)^{-1}$ equal to the inverse of the minimum outgoing edge probability for i.

Lemma C.16. *Let $G = (V,E)$ be a directed graph with $|V| = K \geq 2$ and edge probabilities $p(i,j)$, and such that $(i,i)\in E$ for all $i\in V$. Let $z_i\in\mathbb{R}_+$ be a positive weight assigned to each $i\in V$. Assume that $\sum_{i\in V} z_i \leq 1$ and that $z_i\geq\beta$ for all $i\in V$, given some constant $\beta\in(0,\frac{1}{2}]$. Then,*

$$
\sum_{i\in V}\frac{z_i}{\sum_{j\in C^{\text{in}}(i)} z_j p(j,i)} \leq 6(\alpha_{\mathsf{w}}(G,w^-)+\alpha_{\mathsf{w}}(G,w^+))\ln\left(\frac{2K^2}{\beta\rho}\right)\;,
$$

where $\rho := \min_{i\in V}\sum_{j\in C^{\text{in}}(i)} p(j,i) > 0$.

Proof. The structure of this proof is similar to that of Alon et al. (2015, Lemma 5). Define a discretization of $z_1,\dots,z_K$ such that $(m_i-1)/M \leq z_i \leq m_i/M$ for positive integers $m_1,\dots,m_K$ and $M := \left\lceil\frac{2K}{\beta\rho}\right\rceil$. The discretized values are such that, for all $i\in V$,

$$
\sum_{j\in C^{\text{in}}(i)} m_j p(j,i) \geq M\sum_{j\in C^{\text{in}}(i)} z_j p(j,i) \geq \frac{2K}{\beta\rho}\beta\sum_{j\in C^{\text{in}}(i)} p(j,i) \geq 2K \geq 2|C^{\text{in}}(i)|\;, \qquad \text{(C.41)}
$$

where the first inequality holds because $z_j \leq m_j/M$, the second follows by definition of M and by the assumption on z_j, whereas the third is due to the definition of ρ. Then, the sum of interest becomes

$$\begin{aligned}
\sum_{i \in V} \frac{z_i}{\sum_{j \in C^{\mathrm{in}}(i)} z_j p(j,i)} &\le \sum_{i \in V} \frac{m_i}{M \sum_{j \in C^{\mathrm{in}}(i)} z_j p(j,i)} && \text{since } z_i \le m_i/M \\
&\le \sum_{i \in V} \frac{m_i}{\sum_{j \in C^{\mathrm{in}}(i)} m_j p(j,i) - |C^{\mathrm{in}}(i)|} && \text{since } M z_j \ge m_j - 1 \\
&\le 2 \sum_{i \in V} \frac{m_i}{\sum_{j \in C^{\mathrm{in}}(i)} m_j p(j,i)} && \text{by Equation (C.41).} \qquad \text{(C.42)}
\end{aligned}$$

Now build a new directed graph $G' \coloneqq (V', E')$ derived (as in the proof of Alon et al. (2015, Lemma 5)) from graph G by replacing each node $i \in V$ with a clique K_i of size m_i and all its edges having probability $p(i,i)$. Additionally add an edge from any $i' \in K_i$ to any $j' \in K_j$ having edge probability $p(i,j)$ if and only if $(i,j) \in E$. As a consequence, the right-hand side of Equation (C.42) is equal to

$$2 \sum_{i \in V'} \frac{1}{\sum_{j \in C^{\mathrm{in}}_{G'}(i)} p(j,i)} \; .$$

Observe that the independent sets in G are preserved in G': any independent set $S = \{i : i \in V'\} \in \mathcal{I}(G')$ in G' has a corresponding one $\{i : i' \in S, i' \in K_i\}$ in G with same cardinality and weight, assuming that the weight of $i' \in K_i$ in G' is equal to the weight of $i \in V$ according to the weight assignment in G. We can reduce this latter sum to the same form as in Lemma C.15 by assigning vertex weights

$$w^{\mathrm{in}}_{i'} \coloneqq \left(\sum_{j \in C^{\mathrm{in}}(i)} \frac{m_j}{\sum_{k \in C^{\mathrm{in}}(i)} m_k} p(j,i) \right)^{-1}, \qquad w^{\mathrm{out}}_{i'} \coloneqq \left(\sum_{j \in C^{\mathrm{out}}(i)} \frac{m_j}{\sum_{k \in C^{\mathrm{out}}(i)} m_k} p(i,j) \right)^{-1},$$

to each vertex $i' \in K_i$, for all $i \in V$. Indeed, the previous sum becomes

$$\begin{aligned}
\sum_{i \in V'} \frac{1}{\sum_{j \in C^{\mathrm{in}}_{G'}(i)} p(j,i)} &= \sum_{i \in V'} \frac{w^{\mathrm{in}}_i}{|C^{\mathrm{in}}_{G'}(i)|} \\
&\le 3(\alpha_{\mathsf{w}}(G', w^{\mathrm{in}}) + \alpha_{\mathsf{w}}(G', w^{\mathrm{out}})) \ln(|V'| + 1) && \text{by Lemma C.15} \\
&\le 3(\alpha_{\mathsf{w}}(G, w^-) + \alpha_{\mathsf{w}}(G, w^+)) \ln(|V'| + 1) \; ,
\end{aligned}$$

where the last inequality follows from the fact that $w^{\mathrm{in}}_{i'} \le w^-_i$ and $w^{\mathrm{out}}_{i'} \le w^+_i$ for all $i \in V$ and all $i' \in K_i$.

We conclude the proof by observing that this newly constructed graph also has

$$1 + |V'| = 1 + \sum_{i \in V} m_i \le 1 + \sum_{i \in V} (M z_i + 1) \le K + M + 1 \le 2K \left(1 + \frac{1}{\beta\rho}\right) \le \frac{2K^2}{\beta\rho}$$

vertices, where the final inequality holds because $\beta\rho \le K/2$ by definition, and we used the fact that $K \ge 2$. □

Appendix D

Proof Details for Chapter 6

D.1 Auxiliary Results

Lemma D.1. *Consider any algorithm that picks actions $(A_t)_{t\in[T]}$ in the adversarial delayed bandits problem with intermediate feedback with arbitrary action-state mappings $(s_t)_{t\in[T]}$ and i.i.d. loss vectors $(\ell_t)_{t\in[T]}$. Then, for any given $\delta \in (0,1)$,*

$$R_T - \mathcal{R}_T \le \sqrt{2T\ln(2/\delta)} \qquad \textit{and} \qquad \mathcal{R}_T - R_T \le \sqrt{2T\ln(2K/\delta)}$$

individually hold with probability at least $1-\delta$.

Proof. First, observe that we can relate the two notions of regret as

$$R_T = \mathcal{R}_T + \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) + \underbrace{\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) - \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a))}_{(\triangle)} \ .$$

By Azuma-Hoeffding inequality, we can show that each side of

$$-\sqrt{\frac{T}{2}\ln\frac{1}{\delta'}} \le \sum_{t=1}^{T}\big(\theta(S_t) - \ell_t(S_t)\big) \le \sqrt{\frac{T}{2}\ln\frac{1}{\delta'}} \tag{D.1}$$

holds with probability at least $1-\delta'$. Now, define

$$a_\ell^* \in \arg\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\ell_t(s_t(a)) \qquad \text{and} \qquad a_\theta^* \in \arg\min_{a\in\mathcal{A}}\sum_{t=1}^{T}\theta(s_t(a)) \ .$$

On the one hand, observe that

$$(\triangle) \le \sum_{t=1}^{T}\ell_t(s_t(a_\theta^*)) - \sum_{t=1}^{T}\theta(s_t(a_\theta^*)) \le \sqrt{\frac{T}{2}\ln\frac{1}{\delta'}} \ ,$$

where the last inequality holds with probability at least $1-\delta'$ by Azuma-Hoeffding inequality. On the other hand, we can show that

$$(\triangle) \geq \sum_{t=1}^{T} \ell_t(s_t(a^*_\ell)) - \sum_{t=1}^{T} \theta(s_t(a^*_\ell)) =: (\diamond) \ .$$

However, in this case a^*_ℓ depends on the entire sequence $\ell_1, \dots, \ell_T$. We thus need to use a union bound in order to show that

$$\mathbb{P}\left((\diamond) \leq -\sqrt{\frac{T}{2} \ln \frac{K}{\delta'}}\right) \leq \sum_{a \in \mathcal{A}} \mathbb{P}\left(\sum_{t=1}^{T} \ell_t(s_t(a)) - \sum_{t=1}^{T} \theta(s_t(a)) \leq -\sqrt{\frac{T}{2} \ln \frac{K}{\delta'}}\right) \leq \delta' \ ,$$

where the last inequality follows by Azuma-Hoeffding inequality. We conclude the proof by setting $\delta' = \delta/2$. □

Lemma D.2. *The estimates* $(\widehat{\theta}_t)_{t=1}^T$ *defined in Equation* (6.3) *are such that* $|\widehat{\theta}_t(s) - \theta(s)| \leq \frac{1}{2}\varepsilon_t(s)$ *simultaneously holds for all* $t \in [T]$ *and all* $s \in \mathcal{S}$ *with probability at least* $1 - \delta/2$.

Proof. In a similar way as in Vernade et al. (2020), define $X_m(s)$ to be the empirical mean estimate for $\theta(s)$ which uses the first $m \in [T]$ observed losses corresponding to state $s \in \mathcal{S}$. Notice that $\widehat{\theta}_t(s) = X_{N'_t(s)}(s)$, while we define $\varepsilon'_m(s) := \sqrt{\frac{2}{m} \ln \frac{4ST}{\delta}}$ so that $\varepsilon_t(s) = \varepsilon'_{N'_t(s)}(s)$. We can additionally observe that $\mathbb{E}\left[X_m(s)\right] = \theta(s)$. Then, we can use Azuma-Hoeffding inequality to show that

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{s \in \mathcal{S}} \bigcap_{t \in [T]} \left\{|\widehat{\theta}_t(s) - \theta(s)| \leq \frac{1}{2}\varepsilon_t(s)\right\}\right) &\geq \mathbb{P}\left(\bigcap_{s \in \mathcal{S}} \bigcap_{m \in [T]} \left\{|X_m(s) - \theta(s)| \leq \frac{1}{2}\varepsilon'_m(s)\right\}\right) \\
&\geq 1 - 2\sum_{s \in \mathcal{S}} \sum_{m=1}^{T} e^{-\frac{1}{2}\varepsilon'_m(s)^2 m} \\
&= 1 - \frac{\delta}{2} \ ,
\end{aligned}$$

where we also used a union bound in the second inequality. □

Lemma D.3. *Consider any algorithm that picks actions* $(A_t)_{t \in [T]}$ *in the BIO setting with adversarial action-state mappings* $(s_t)_{t \in [T]}$ *and stochastic loss vectors* $(\ell_t)_{t \in [T]}$. *Assume that the losses for any fixed state are i.i.d., whereas pairs of losses* $\ell_j(s), \ell_{j'}(s')$ *of distinct states* $s \neq s'$ *might be correlated when* $j > j'$ *and* $j - j' \leq d_{j'}$. *Then, it holds that* $\mathbb{E}\left[R_T\right] \leq \mathbb{E}\left[\mathcal{R}_T\right]$, *where the expectation is with respect to the stochasticity of the losses and the randomness of the algorithm.*

Proof. We know that $\mathbb{E}\left[\ell_t(s_t(a))\right] = \theta(s_t(a))$ for any fixed $a \in \mathcal{A}$ and all $t \in [T]$. We further observe that

$$\mathbb{E}\left[\ell_t(S_t)\right] = \mathbb{E}\Big[\mathbb{E}\left[\ell_t(s_t(A_t)) \mid A_t\right]\Big] = \mathbb{E}\left[\theta(S_t)\right]$$

holds for all $t \in [T]$, as A_t is independent of losses that can be correlated with ℓ_t. Now, define

$$a^*_\ell \in \operatorname*{arg\,min}_{a \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(s_t(a)) \qquad \text{and} \qquad a^*_\theta \in \operatorname*{arg\,min}_{a \in \mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a)) \ .$$

Then, we conclude the proof by showing that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_T\right] &= \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(S_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a_\ell^*))\right] \\ &\geq \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(S_t)\right] - \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(s_t(a_\theta^*))\right] = \sum_{t=1}^{T} \mathbb{E}\left[\theta(S_t)\right] - \sum_{t=1}^{T} \theta(s_t(a_\theta^*)) = \mathbb{E}\left[R_T\right] \ .\end{aligned}$$

□

D.2 High-Probability Regret Bound

D.2.1 Total Effective Delay Bound

Lemma 6.1 (Total effective delay). *If* `MetaBIO` *is run with any algorithm $\mathcal{B}$ on delays $(d_t)_{t\in[T]}$, then its total effective delay is $\widetilde{\mathcal{D}}_T \leq \mathcal{D}_\Phi$.*

Proof of Lemma 6.1. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s \coloneqq \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s. Denote by t_s the last time step $t \in \mathcal{T}_s$ such that $N_t(s) < \sigma_t$ and let $\mathcal{C}_s \coloneqq \{t \in \mathcal{T}_s : t \leq t_s\}$ be those rounds in $\mathcal{T}_s$ that come no later than t_s. According to the choice of t_s, all the rounds in $\mathcal{T}_s$ for which learner waits for the respective delayed loss, must belong to $\mathcal{C}_s$, while the learner incurs $\widetilde{d}_t = 0$ delay for rounds $t \in \mathcal{T}_s \setminus \mathcal{C}_s$. Now we partition $\mathcal{C}_s$ into two sets: the observed set $\mathcal{C}_s^{\text{obs}} \coloneqq \{t \in \mathcal{C}_s : t + d_t \leq t_s\}$ and the outstanding set $\mathcal{C}_s^{\text{out}} \coloneqq \{t \in \mathcal{C}_s : t + d_t > t_s\}$. From the choice of t_s, we can see that the number of rounds in $\mathcal{C}_s^{\text{obs}}$ is

$$|\mathcal{C}_s^{\text{obs}}| \leq N_{t_s}(s) < \sigma_{t_s} \leq \sigma_{\max} \ ,$$

and the number of rounds in C_s^{out} is

$$|\mathcal{C}_s^{\text{out}}| \leq \sigma_{t_s} \leq \sigma_{\max} \ .$$

Therefore, we have $|\mathcal{C}_s| \leq 2\sigma_{\max}$. So if we define $\mathcal{C}_{\text{all}} \coloneqq \bigcup_{s\in\mathcal{S}} \mathcal{C}_s$, then $|\mathcal{C}_{\text{all}}| \leq \min\{2S\sigma_{\max}, T\} = |\Phi|$. This also implies that

$$\sum_{t=1}^{T} \widetilde{d}_t \leq \sum_{t\in\mathcal{C}_{\text{all}}} d_t \leq \sum_{t\in\Phi} d_t$$

by definition of Φ. □

D.2.2 Improved Regret for `DAda-Exp3` for Fixed δ

We follow the analysis of Theorem 4.1 in György and Joulani (2021, Appendix A) and our goal is to use the knowledge of $\delta \in (0,1)$ to tune the learning rates $(\eta_t)_{t\in[T]}$ and the implicit exploration terms $(\gamma_t)_{t\in[T]}$, accordingly. Let $d_1, \ldots, d_T$ be the sequence of delays perceived by `DAda-Exp3`, and let $D_T \coloneqq \sum_{t=1}^{T} d_t$ be its total delay. Furthermore, let σ_t be the number of outstanding observations of `DAda-Exp3` at the beginning of round $t \in [T]$. Suppose that we take $\gamma_t = c\eta_t$ with $c > 0$ for all $t \in [T]$, then following the same analysis as in György and Joulani (2021, Appendix A), we end up with the following regret bound that holds with probability at least $1-2\delta'$ for any $\delta' \in (0, 1/2)$:

$$\begin{aligned}\mathcal{R}_T &\leq \frac{\ln K}{\eta_T} + \sum_{t=1}^{T} \eta_t(\sigma_t + (c+1)K) + \frac{\ln(K/\delta')}{2c\eta_T} + \frac{\sigma_{\max} + c + 1}{2c} \ln(1/\delta') \\ &= \frac{1}{\eta_T}\left(\ln K + \frac{\ln(K/\delta')}{2c}\right) + \sum_{t=1}^{T} \eta_t(\sigma_{t-1} + (c+1)K) + \frac{\sigma_{\max}+1}{2c}\ln(1/\delta') + \frac{\ln(1/\delta')}{2} \,.\end{aligned}$$

Therefore, by taking $\eta_t^{-1} = \sqrt{\frac{(c+1)Kt+\sum_{j=1}^{t}\sigma_j}{2\ln(K)+\frac{1}{c}\ln(K/\delta')}}$, we get the following bound with probability at least $1-2\delta'$:

$$\mathcal{R}_T \leq 2\sqrt{\left((c+1)KT + \sum_{t=1}^{T}\sigma_t\right)\left(2\ln(K) + \frac{\ln(K/\delta')}{c}\right)} + \frac{\sigma_{\max}+1}{2c}\ln(1/\delta') + \frac{\ln(1/\delta')}{2} \,.$$

We know that $\sum_{t=1}^{T}\sigma_t = D_T$ by definition of σ_t. Then, we can set $c = 1$ to obtain that the regret $\mathcal{R}_T$ (as per the original notion of regret used in György and Joulani (2021)) is

$$\mathcal{R}_T \leq 2\sqrt{2KT\left(3\ln(K) + \ln\left(1/\delta'\right)\right)} + 2\sqrt{D_T\left(3\ln(K) + \ln\left(1/\delta'\right)\right)} + \frac{\sigma_{\max}+2}{2}\ln(1/\delta') \tag{D.2}$$

with probability at least $1-2\delta'$.

From Lemma D.1, we have that

$$R_T \leq \mathcal{R}_T + \sqrt{2T\ln(2/\delta')} \tag{D.3}$$

holds with probability at least $1-\delta'$. So, combining Equations (D.2) and (D.3), and setting $\delta \coloneqq 3\delta'$, we can upper bound our notion of regret R_T as

$$R_T \leq 2\sqrt{2KT\left(3\ln(K) + \ln\left(3/\delta\right)\right)} + \sqrt{2T\ln(6/\delta)} + 2\sqrt{D_T\left(3\ln(K) + \ln\left(3/\delta\right)\right)} + \frac{\sigma_{\max}+2}{2}\ln(3/\delta) \tag{D.4}$$

with probability at least $1-\delta$.

D.2.3 Reduction to `DAda-Exp3` via `MetaBIO`

Based on the reduction via `MetaBIO`, we require that $\mathcal{B}$ guarantee a regret bound

$$\widehat{\mathcal{R}}_T^{\mathcal{B}} = \sum_{t=1}^{T}\widetilde{\theta}_t(S_t) - \min_{a\in\mathcal{A}}\sum_{t=1}^{T}\widetilde{\theta}_t(s_t(a)) \tag{D.5}$$

that holds with high probability when the losses experienced by $\mathcal{B}$ are of the form $\widetilde{\theta}_t\big(s_t(a)\big)$. Note that, even though the action-state mappings $s_1, \dots, s_T$ are unknown to the learner, we can provide those losses as long as $\mathcal{B}$ requires bandit feedback only. Indeed, we can compute $\widetilde{\theta}_t(S_t)$ defined in Equations (6.1) and (6.3), while we cannot determine $s_t(a)$ for all actions $a \in \mathcal{A}$ that are not A_t. As mentioned in Section 6.4, in this work we consider `DAda-Exp3` (György and Joulani, 2021) as algorithm $\mathcal{B}$ used by `MetaBIO`. In what follows, we refer to this specific choice for the algorithm $\mathcal{B}$.

The analysis of `DAda-Exp3` for the high-probability bound (Theorem 6.1) is such that most steps only require that the loss of each action is bounded in $[0,1]$. Then, those steps apply for any such sequence of loss vectors. However, the crucial part of that analysis that requires attention is the application of Lemma 1 from Neu (2015). We restate it below for reference.

Before that, we introduce the notation required for stating the result. We consider a learner choosing actions $A_1,\dots,A_T$ according to probability distributions $p_1,\dots,p_T$ over actions. We denote by $\mathcal{F}_{t-1}$ the observation history of the learner until the beginning of round t. The result uses importance-weighted estimates for the losses $\ell_1,\dots,\ell_T$ with implicit exploration, where the implicit exploration parameter is $\gamma_t \geq 0$ for each time t. These loss estimates are defined as

$$\widetilde{\ell}_t(a) = \frac{\mathbb{I}\{A_t = a\}}{p_t(a)+\gamma_t}\ell_t(a) \qquad \forall t \in [T], \forall a \in \mathcal{A} \ . \tag{D.6}$$

Lemma D.4 (Neu (2015, Lemma 1)). *Let γ_t and $\alpha_t(a)$ be nonnegative $\mathcal{F}_{t-1}$-measurable random variables such that $\alpha_t(a) \leq 2\gamma_t$, for all $t \in [T]$ and all $a \in \mathcal{A}$. Let $\widetilde{\ell}_t(a)$ be as in* (D.6). *Then,*

$$\sum_{t=1}^{T}\sum_{a=1}^{K} \alpha_t(a)\big(\widetilde{\ell}_t(a) - \ell_t(a)\big) \leq \ln(1/\delta)$$

holds with probability at least $1-\delta$ for any $\delta \in (0,1)$.

In our case, we require an analogous result that work when loss vectors correspond with our estimates $\widetilde{\theta}_1,\dots,\widetilde{\theta}_T$. However, these estimate have a dependency with the past actions chosen by the learner. This requires some nontrivial changes in the proof of Neu (2015, Lemma 1).

Before that, we introduce some crucial definitions for this proof. Let $\rho(t) \coloneqq t + d_t$ be the arrival time for the realized loss $\ell_t(S_t)$ of the state S_t observed at time $t \in [T]$. Let $\widetilde{\rho}(t) \coloneqq t + \widetilde{d}_t$ be instead the arrival time perceived by algorithm $\mathcal{B}$ relative to its choice of A_t at time t, i.e., when $\mathcal{B}$ receives $\widetilde{\theta}_t(S_t)$. This also means that $\widetilde{\theta}_t(S_t)$ is only defined at time $\widetilde{\rho}(t) \leq \rho(t)$.

Let $\pi\colon [T] \to [T]$ be the permutation of $[T]$ that orders rounds according to their value of $\widetilde{\rho}$. In other words, π satisfies the following property:

$$\pi(r) < \pi(t) \qquad \Longleftrightarrow \qquad \widetilde{\rho}(r) < \widetilde{\rho}(t) \lor (\widetilde{\rho}(r) = \widetilde{\rho}(t) \land r < t) \qquad \forall r,t \in [T] \ . \tag{D.7}$$

This permutation allows us to sort rounds according to the order in which `MetaBIO` feeds $\mathcal{B}$ with a respective estimate for the mean loss. In particular, the r-th round in this order corresponds with the round $t_r \coloneqq \pi^{-1}(r)$, for any $r \in [T]$. Hence, we can equivalently define the round t_r as the round such that its estimate $\widetilde{\theta}_{t_r}(S_{t_r})$ for the mean loss $\theta(S_{t_r})$ is the r-th estimate received by $\mathcal{B}$.

Define

$$\mathcal{F}_r \coloneqq \{(j, A_j, S_j, \ell_j(S_j)) \mid j \in [T], \pi(j) \leq r\} \qquad \forall r \in [T] \tag{D.8}$$

as the information observed by $\mathcal{B}$ by the end to the time step when we feed it the estimate relative to round t_r. Note that this defines a filtration, as $\mathcal{F}_{r-1} \subseteq \mathcal{F}_r$ for all $r \in [T]$, which has some desirable properties thanks to the ordering π we consider. In particular, we have that $\widetilde{d}_{t_r}, \varepsilon_{t_r}, p_{t_r}, N'_{t_r}$ are $\mathcal{F}_{r-1}$-measurable random variables by the way we define them. This property is also due to the fact that N_{t_r} and $\mathcal{L}'_{t_r}$ are determined when conditioning on $\mathcal{F}_{r-1}$. Moreover, we are now interested in the following importance-weighted loss estimates with implicit exploration:

$$\widetilde{\ell}_t(a) \coloneqq \frac{\mathbb{I}\{A_t = a\}}{p_t(a) + \gamma_t}\widetilde{\theta}_t(s_t(a)) \qquad \forall t \in [T], \forall a \in \mathcal{A} \, . \tag{D.9}$$

Corollary D.1. *Let γ_{t_r} and $\alpha_{t_r}(a)$ be non-negative $\mathcal{F}_{r-1}$-measurable random variables such that $\alpha_{t_r}(a) \le 2\gamma_{t_r}$, for all $r \in [T]$ and all $a \in \mathcal{A}$. Let $\widetilde{\ell}_t(a)$ be as in* (D.9)*. Then,*

$$\sum_{t=1}^{T}\sum_{a=1}^{K} \alpha_t(a)\big(\widetilde{\ell}_t(a) - \widetilde{\theta}_t(s_t(a))\big) \le \ln\left(1/\delta\right)$$

holds with probability at least $1-\delta$ for any $\delta \in (0,1)$.

Proof. We follow the proof of Neu (2015, Lemma 1) by considering any realization $\ell_1, \dots, \ell_T$ of the losses. The main difference is that, when defining the supermartingale as in the original proof, we need to consider the terms of the sum in the order denoted by π instead of the increasing order of t. For this reason, we rewrite the sum from the statement by following the order given by π:

$$\sum_{r=1}^{T}\sum_{a=1}^{K} \alpha_{t_r}(a)\big(\widetilde{\ell}_{t_r}(a) - \widetilde{\theta}_{t_r}(s_{t_r}(a))\big) \, .$$

At this point, we need prove that $\mathbb{E}\big[\widetilde{\ell}_{t_r}(a) \,\big|\, \mathcal{F}_{r-1}\big] \le \widetilde{\theta}_{t_r}(s_{t_r}(a))$, where we recall that $t_r = \pi^{-1}(r)$. Also recall that ε_{t_r}, p_{t_r} and γ_{t_r} are $\mathcal{F}_{r-1}$-measurable. This property allows us to prove the inequality with the conditional expectation of $\widehat{\theta}_t$ instead of the one with the actual optimistic estimates $\widetilde{\theta}_t$, by the definition of the latter. In other words, we now need to prove that $\mathbb{E}\big[\widehat{\ell}_{t_r}(a) \,\big|\, \mathcal{F}_{r-1}\big] \le \widehat{\theta}_{t_r}(s_{t_r}(a))$, where $\widehat{\ell}_t(a) = \frac{\mathbb{I}\{A_t=a\}}{p_t(a)+\gamma_t}\widehat{\theta}_t(s_t(a))$.

We can consider two cases depending on whether $\widetilde{d}_{t_r} < d_{t_r}$ is true or not (and, thus, we are in the case $\widetilde{d}_{t_r} = d_{t_r}$). In the first case, note that the realized losses used for computing $\widehat{\theta}_{t_r}(s_{t_r}(a))$ correspond to time steps in $\mathcal{L}'_{t_r}(s_{t_r}(a))$, for which there is a corresponding tuple in $\mathcal{F}_{r-1}$. Therefore, we have that $\widehat{\theta}_{t_r}(s_{t_r}(a))$ is $\mathcal{F}_{r-1}$-measurable, and we can show that

$$\mathbb{E}\left[\widehat{\ell}_{t_r}(a)\mathbb{I}\Big\{\widetilde{d}_{t_r} < d_{t_r}\Big\} \,\Big|\, \mathcal{F}_{r-1}\right] = \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \,\middle|\, \mathcal{F}_{r-1}\right] \frac{\mathbb{I}\Big\{\widetilde{d}_{t_r} < d_{t_r}\Big\}}{N'_{t_r}(s_{t_r}(a))} \sum_{j \in \mathcal{L}'_{t_r}(s_{t_r}(a))} \ell_j(s_{t_r}(a)) \, .$$

In the second case, we have that $\widetilde{d}_{t_r} = d_{t_r}$, which implies that $t_r \in \mathcal{L}'_{t_r}(s_{t_r}(a))$ in the case $A_{t_r} = a$. This means that we have a corresponding tuple in $\mathcal{F}_{r-1}$ only for rounds in $\mathcal{L}'_{t_r}(s_{t_r}(a)) \setminus \{t_r\}$. Nonetheless, this does not pose an issue since we have the indicator $\mathbb{I}\{A_{t_r} = a\}$, and thus $S_{t_r} = s_t(a)$. Indeed, we have that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\ell}_{t_r}(a)\mathbb{I}\Big\{\widetilde{d}_{t_r} = d_{t_r}\Big\} \,\Big|\, \mathcal{F}_{r-1}\right] &= \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \cdot \frac{\mathbb{I}\Big\{\widetilde{d}_{t_r} = d_{t_r}\Big\}}{N'_{t_r}(s_{t_r}(a))} \sum_{j \in \mathcal{L}'_{t_r}(s_{t_r}(a))} \ell_j(s_{t_r}(a)) \,\middle|\, \mathcal{F}_{r-1}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \,\middle|\, \mathcal{F}_{r-1}\right] \frac{\mathbb{I}\Big\{\widetilde{d}_{t_r} = d_{t_r}\Big\}}{N'_{t_r}(s_{t_r}(a))} \sum_{\substack{j \in \mathcal{L}'_{t_r}(s_{t_r}(a)) \\ j \neq t_r}} \ell_j(s_{t_r}(a))
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \,\middle|\, \mathcal{F}_{r-1}\right] \frac{\mathbb{I}\left\{\widetilde{d}_{t_r} = d_{t_r}\right\}}{N'_{t_r}(s_{t_r}(a))} \ell_{t_r}(s_{t_r}(a)) \\
&= \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \,\middle|\, \mathcal{F}_{r-1}\right] \frac{\mathbb{I}\left\{\widetilde{d}_{t_r} = d_{t_r}\right\}}{N'_{t_r}(s_{t_r}(a))} \sum_{j \in \mathcal{L}'_{t_r}(s_{t_r}(a))} \ell_j(s_{t_r}(a))
\end{aligned}
$$

and therefore the inequality

$$
\mathbb{E}\left[\widehat{\ell}_{t_r}(a) \,\middle|\, \mathcal{F}_{r-1}\right] = \mathbb{E}\left[\frac{\mathbb{I}\{A_{t_r} = a\}}{p_{t_r}(a) + \gamma_{t_r}} \,\middle|\, \mathcal{F}_{r-1}\right] \widehat{\theta}_{t_r}(s_{t_r}(a)) \le \widehat{\theta}_{t_r}(s_{t_r}(a))
$$

is true because $\mathbb{I}\{\widetilde{d}_t < d_t\} + \mathbb{I}\{\widetilde{d}_t = d_t\} = 1$ for all $t \in [T]$, and by definition of $\widehat{\theta}_t$.

As already mentioned, this is equivalent to proving that $\mathbb{E}\left[\widetilde{\ell}_{t_r}(a) \,\middle|\, \mathcal{F}_{r-1}\right] \le \widetilde{\theta}_{t_r}(s_{t_r}(a))$ holds. By using a notation similar to the original proof, if we define $\widetilde{\lambda}_r \coloneqq \sum_{a=1}^K \alpha_{t_r}(a)\widetilde{\ell}_{t_r}(a)$ and $\lambda_r \coloneqq \sum_{a=1}^K \alpha_{t_r}(a)\widetilde{\theta}_{t_r}(s_{t_r}(a))$, the process $(Z_r)_{r\in[T]}$ with $Z_r \coloneqq \exp\left(\sum_{j=1}^r \left(\widetilde{\lambda}_j - \lambda_j\right)\right)$ is a supermartingale with respect to $(\mathcal{F}_r)_{r\in[T]}$ which has the same properties as in the proof of Neu (2015, Lemma 1). This concludes the current proof by following a similar reasoning as in the original one. □

Thanks to this result, we can conclude that the adoption of `DAda-Exp3` for the reduction via `MetaBIO` can guarantee a high-probability regret bound on $\widehat{\mathcal{R}}_T^{\mathcal{B}}$ as stated in Theorem 6.1, but with total delay $\widetilde{\mathcal{D}}_T = \sum_{t=1}^T \widetilde{d}_t$ instead of $\mathcal{D}_T$.

D.2.4 Regret of `MetaBIO`

By Lemma D.2, we have that

$$
R_T \le \sum_{t=1}^T \widetilde{\theta}_t(S_t) - \min_{a\in\mathcal{A}} \sum_{t=1}^T \widetilde{\theta}_t(s_t(a)) + \sum_{t=1}^T \varepsilon_t(S_t) = \widehat{\mathcal{R}}_T^{\mathcal{B}} + \sum_{t=1}^T \varepsilon_t(S_t) \tag{D.10}
$$

with probability at least $1 - \delta/2$, where $\widehat{R}_T^{\mathcal{B}}$ (Equation (D.5)) is the regret of algorithm $\mathcal{B}$ when fed with $(\widetilde{\theta}_t \circ s_t)_{t\in[T]}$ as losses.

Lemma D.5. *Conditioning on the event as stated in Lemma D.2, the sum of errors suffered from* `MetaBIO` *by using the loss estimates* $(\widetilde{\theta}_t)_{t\in[T]}$ *from Equations* (6.1) *and* (6.3) *is*

$$
\sum_{t=1}^T \varepsilon_t(S_t) \le \left(4 + 2\sqrt{2}\right)\sqrt{ST \ln \frac{4ST}{\delta}} \;.
$$

Proof. First, observe that we can rewrite the sum of errors as

$$
\sum_{t=1}^T \varepsilon_t(S_t) = \sum_{t=1}^T \varepsilon_t(S_t)\mathbb{I}\left\{\widetilde{d}_t < d_t\right\} + \sum_{t=1}^T \varepsilon_t(S_t)\mathbb{I}\left\{\widetilde{d}_t = d_t\right\} .
$$

We now provide an upper bound for the first sum of errors. For any $s \in \mathcal{S}$, we define $\mathcal{T}_s \coloneqq \{t \in [T] : S_t = s\}$ to be the set of all rounds when the state observed by the learner corresponds to s.

We can bound it as

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{I}\{\widetilde{d}_t < d_t\} &= \sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s} \varepsilon_t(s)\mathbb{I}\{\widetilde{d}_t < d_t\} \\
&= \sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s}\sqrt{\frac{1}{N'_t(s)}}\mathbb{I}\{\widetilde{d}_t < d_t\} \\
&\le 2\sqrt{\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s}\sqrt{\frac{1}{M_t(s)}}\mathbb{I}\{\widetilde{d}_t < d_t\} && \left(\text{because } N'_t(s) \ge \tfrac{1}{2}M_t(s)\right) \\
&\le 4\sqrt{\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sqrt{M_T(s)} && (\text{since } M_t(s) \text{ is increasing over } \mathcal{T}_s) \\
&\le 4\sqrt{ST\ln\frac{4ST}{\delta}}\ ,
\end{aligned}$$

where the second inequality holds because $N'_t(S_t) = N_t(S_t) \ge \frac{1}{2}M_t(S_t)$ when $\widetilde{d}_t < d_t$ since $M_t(S_t) \le N_t(S_t) + \sigma_t$, while the last one follows by Jensen's inequality and the fact that $\sum_{s\in\mathcal{S}} M_T(s) = T$.

As a last step, we provide an upper bound to the second sum. Let $J_s \coloneqq \{r \in \mathcal{T}_s : \widetilde{d}_r = d_r\}$ and notice that $|J_s| \le |\mathcal{T}_s| = M_T(s)$. Observe that $\rho(t) = \widetilde{\rho}(t)$ for each round t such that $\widetilde{d}_t = d_t$, and thus by Equation (D.7) we have that

$$\pi(r) < \pi(t) \iff \rho(r) < \rho(t) \lor (\rho(r) = \rho(t) \land r < t)$$

for all $r, t \in [T]$ such that $\widetilde{d}_r = d_r$ and $\widetilde{d}_t = d_t$. Define $\nu_s : J_s \to \big[|J_s|\big]$ by

$$\nu_s(t) \coloneqq |\{r \in J_s \,:\, \pi(r) \le \pi(t)\}| \qquad \forall t \in J_s\ .$$

Observe that $\nu_s(t) \le N'_t(s) = |\mathcal{L}'_t(s)|$ for all $s \in \mathcal{S}$ and all $t \in J_s$. This is due to the fact that $\nu_s(t)$ counts a subset of $\mathcal{L}'_t(s)$; to be precise, we have that $\nu_s(t) = |\mathcal{L}'_t(s) \cap J_s|$. Moreover, notice that the condition $\pi(r) \le \pi(t)$ defines a total order over J_s. Hence, $\nu_s(t)$ counts the number of elements of J_s preceding $t \in J_s$ (including t itself) in this total order. This implies that ν_s is a bijection between J_s and $\big[|J_s|\big]$. Then, using a similar reasoning as before, we show that

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t(S_t)\mathbb{I}\{\widetilde{d}_t = d_t\} &= \sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{T}_s}\sqrt{\frac{1}{N'_t(s)}}\mathbb{I}\{\widetilde{d}_t = d_t\} \\
&= \sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sum_{t\in J_s}\sqrt{\frac{1}{N'_t(s)}} && (\text{by definition of } J_s) \\
&\le \sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sum_{t\in J_s}\sqrt{\frac{1}{\nu_s(t)}} && (\text{since } \nu_s(t) \le N'_t(s) \text{ for } t \in J_s) \\
&\le 2\sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sqrt{|J_s|} && (\text{since } \nu_s(t) \text{ is bijective}) \\
&\le 2\sqrt{2\ln\frac{4ST}{\delta}}\sum_{s\in\mathcal{S}}\sqrt{M_T(s)} && (\text{since } |J_s| \le M_T(s))
\end{aligned}$$

$$\leq 2\sqrt{2ST \ln \frac{4ST}{\delta}} \,. \qquad \text{(by Jensen's inequality)}$$

□

Theorem 6.2. *Let $\delta \in (0,1)$. If we run* `MetaBIO` *using* `DAda-Exp3`*, then the regret of* `MetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$R_T \leq 2\sqrt{2KTC_{K,3\delta}} + 7\sqrt{ST \ln \frac{4ST}{\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,3\delta}} + \frac{\sigma_{\max} + 2}{2} \ln \frac{4}{\delta} \tag{6.8}$$

with probability at least $1 - \delta$.

Proof of Theorem 6.2. By Equation (D.10), the regret R_T can be bounded as

$$R_T \leq \widehat{\mathcal{R}}_T^{\mathcal{B}} + \sum_{t=1}^{T} \varepsilon_t(S_t) \leq \widehat{\mathcal{R}}_T^{\mathcal{B}} + 7\sqrt{ST \ln \frac{4ST}{\delta}}$$

with probability at least $1 - \delta/2$, where the last inequality follows by Lemma D.5. From what we argued in Appendix D.2.3, we can upper bound $\widehat{\mathcal{R}}_T^{\mathcal{B}}$ using the high-probability regret bound of `DAda-Exp3`. Notice that the delays incurred by `DAda-Exp3` via `MetaBIO` are those given when providing the estimates $\big(\widetilde{\theta}_t\big)_{t \in [T]}$. We denote these delays by $\widetilde{d}_1, \ldots, \widetilde{d}_T$, and the total delay perceived by `DAda-Exp3` is thus $\widetilde{\mathcal{D}}_T = \sum_{t=1}^{T} \widetilde{d}_t$. Hence, from the improved bound for `DAda-Exp3` in Equation (D.2), we have that

$$\widehat{\mathcal{R}}_T^{\mathcal{B}} \leq 2\sqrt{2KT\left(3\ln(K) + \ln\left(4/\delta\right)\right)} + 2\sqrt{\widetilde{\mathcal{D}}_T\left(3\ln(K) + \ln\left(4/\delta\right)\right)} + \frac{\sigma_{\max} + 2}{2} \ln(4/\delta)$$

holds with probability at least $1 - \delta/2$. The combination of the above two inequalities, together with Lemma 6.1, concludes the proof. □

D.2.5 Regret of `AdaMetaBIO`

Theorem 6.3. *Let $\delta \in (0,1)$. If we run* `AdaMetaBIO` *with* `DAda-Exp3`*, then the regret of* `AdaMetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$R_T \leq 3\min\left\{7\sqrt{ST \ln \frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 6\sqrt{KTC_{K,2\delta}} + 2\sqrt{\mathcal{D}_\Phi C_{K,2\delta}} + (\sigma_{\max} + 2)\ln \frac{8}{\delta} \tag{6.9}$$

with probability at least $1 - \delta$.

Proof of Theorem 6.3. Let $t^* \in [T]$ be the last round before `AdaMetaBIO` switches from `DAda-Exp3` to `MetaBIO`, i.e., the last round that satisfies $\mathfrak{D}_{t^*} C_{K,4\delta} \leq 49ST \ln \frac{8ST}{\delta}$. Then, define

$$a^* \in \arg\min_a \sum_{t=1}^{T} \theta(s_t(a)) \,.$$

We may decompose regret as

$$R_T = \sum_{t=1}^{t^*} \Big(\theta(S_t) - \theta(s_t(a^*))\Big) + \sum_{t=t^*+1}^{T} \Big(\theta(S_t) - \theta(s_t(a^*))\Big)$$

$$\leq \underbrace{\sum_{t=1}^{t^*}\theta(S_t) - \min_{a\in\mathcal{A}}\sum_{t=1}^{t^*}\theta(s_t(a))}_{R_{t^*}} + \underbrace{\sum_{t=t^*+1}^{T}\theta(S_t) - \min_{a\in\mathcal{A}}\sum_{t=t^*+1}^{T}\theta(s_t(a))}_{R_{t^*:T}} \ .$$

The incurred delay until time t^* is $\mathfrak{D}_{t^*}$. Thus, from Equation (D.4), we get that the following bound

$$R_{t^*} \leq 2\sqrt{2Kt^*C_{K,2\delta}} + \sqrt{2t^*\ln\frac{12}{\delta}} + 2\sqrt{\mathfrak{D}_{t^*}C_{K,2\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{6}{\delta} \tag{D.11}$$

holds with probability at least $1-\delta/2$, where we recall that $C_{K,\delta} = 3\ln K + \ln(12/\delta)$. If our algorithm never switches, then $t^* = T$ and we get the bound in (D.11) for R_T. Note that this is no greater than the upper bound in the statement as $\sqrt{\mathfrak{D}_T C_{K,2\delta}} \leq 7\sqrt{ST\ln(8ST/\delta)}$ by definition of t^* in this case.

Otherwise, we use the switching condition $\sqrt{\mathfrak{D}_{t^*}C_{K,2\delta}} \leq 7\sqrt{ST\ln(8ST/\delta)}$ along with the fact that $\sqrt{t^*\ln(12/\delta)} \leq \sqrt{Kt^*C_{K,2\delta}}$ to get

$$R_{t^*} \leq 3\sqrt{2Kt^*C_{K,2\delta}} + 14\sqrt{ST\ln\frac{8ST}{\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{6}{\delta} \ . \tag{D.12}$$

Furthermore, Theorem 6.2 directly gives us an upper bound for $R_{t^*:T}$ since `AdaMetaBIO` runs `MetaBIO` for $t > t^*$ with the confidence parameter set to $\delta/2$. We just need to bound the total incurred delays of these rounds, namely $\widetilde{\mathcal{D}}_{t^*:T}$. Let σ'_t be the outstanding observations for any round $t > t^*$ as perceived by the execution of `MetaBIO` starting after round t^*, that is, when considering only delays $(d_t)_{t>t^*}$. It is immediate to observe that $\sigma'_t \leq \sigma_t$ and thus $\max_{t>t^*}\sigma'_t \leq \max_{t>t^*}\sigma_t$. Moreover, from Lemma 6.1 we have

$$\widetilde{\mathcal{D}}_{t^*:T} \leq \mathcal{D}_{\Phi'} \ ,$$

where Φ' denotes a set of $\min\{T-t^*, 2S\sigma'_{\max}\}$ rounds with the largest delays among $(d_t)_{t>t^*}$, with $\sigma'_{\max} := \max_{t>t^*}\sigma'_t$. So we have

$$\mathcal{D}_{\Phi'} \leq \mathcal{D}_{\Phi}$$

due to the fact that $|\Phi'| = \min\{T-t^*, 2S\sigma'_{\max}\} \leq \min\{T, 2S\sigma_{\max}\} = |\Phi|$. Therefore, from Theorem 6.2 we obtain

$$R_{t^*:T} \leq 2\sqrt{2K(T-t^*)C_{K,3\delta}} + 7\sqrt{ST\ln\frac{8ST}{\delta}} + 2\sqrt{\mathcal{D}_{\Phi}C_{K,3\delta}} + \frac{\sigma_{\max}+2}{2}\ln\frac{8}{\delta} \tag{D.13}$$

with probability at least $1-\delta/2$. We conclude the proof by combining Equations (D.12) and (D.13) along with the fact that $\sqrt{t^*} + \sqrt{T-t^*} \leq \sqrt{2T}$ to get that the bound

$$R_T \leq 6\sqrt{KTC_{K,2\delta}} + 3\min\left\{7\sqrt{ST\ln\frac{8ST}{\delta}}, \sqrt{\mathcal{D}_T C_{K,2\delta}}\right\} + 2\sqrt{\mathcal{D}_{\Phi}C_{K,2\delta}} + (\sigma_{\max}+2)\ln\frac{8}{\delta}$$

holds with probability at least $1-\delta$. □

D.2.6 Expected Regret Analysis of `AdaMetaBIO` with `Tsallis-INF`

Proposition 6.1. *If we execute* `AdaMetaBIO` *with* `Tsallis-INF` *(Zimmert and Seldin, 2020), and use the switching condition* $\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST\ln(2ST)}$ *at each round* $t \in [T]$*, where* $\mathfrak{D}_t = \sum_{j=1}^{t} \sigma_j$*, then the regret of* `AdaMetaBIO` *in the BIO setting with adversarial action-state mappings and stochastic losses satisfies*

$$\mathbb{E}\left[R_T\right] \le 4\sqrt{2KT} + 2\sqrt{2\mathcal{D}_\Phi \ln K} + 4\min\left\{3\sqrt{ST\ln(2ST)}, \sqrt{2\mathcal{D}_T \ln K}\right\} .$$

Proof of Proposition 6.1. We begin by studying of expected regret of `MetaBIO` and we then give a regret analysis of `AdaMetaBIO`. When running `MetaBIO`, we use the unbiased empirical mean estimators $(\widehat{\theta}_t)_{t\in[T]}$ as the mean loss estimates, rather than the lower confidence bounds $(\widetilde{\theta}_t)_{t\in[T]}$. The expected regret is defined as

$$\mathbb{E}[R_T] = \sum_{t=1}^{T} \mathbb{E}\left[\theta(S_t)\right] - \sum_{t=1}^{T} \theta(s_t(a^*)) ,$$

where we fix any $a^* \in \arg\min_{a\in\mathcal{A}} \sum_{t=1}^{T} \theta(s_t(a))$. Here we use a version of `Tsallis-INF` that is tailored for the delayed bandits problem (Zimmert and Seldin, 2020), which guarantees a bound in expectation on the regret

$$\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a) \coloneqq \sum_{t=1}^{T} \widehat{\theta}_t(S_t) - \sum_{t=1}^{T} \widehat{\theta}_t(s_t(a))$$

against any fixed action $a \in \mathcal{A}$, using the loss estimates $\big(\widehat{\theta}_t\big)_{t\in[T]}$. Observe that this regret is defined in terms of our estimates, as required in our case. By Zimmert and Seldin (2020, Theorem 1), `Tsallis-INF` guarantees that its expected regret is

$$\mathbb{E}\left[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \widehat{\theta}_t(S_t) - \sum_{t=1}^{T} \widehat{\theta}_t(s_t(a^*))\right] \le 4\sqrt{KT} + \sqrt{8\widetilde{\mathcal{D}}_T \ln K} \le 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} ,$$

where the last inequality uses Lemma 6.1. Then, we can focus on our notion of regret and use the above regret bound to obtain that

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[R_T - \widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\right] + \mathbb{E}\left[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\theta(S_t) - \widehat{\theta}_t(S_t)\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\right)\right] + \mathbb{E}\left[\widehat{\mathcal{R}}_T^{\text{Tsallis}}(a^*)\right] \\
&\le \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(\theta(S_t) - \widehat{\theta}_t(S_t)\right)\right]}_{\Delta} + \mathbb{E}\left[\sum_{t=1}^{T}\left(\widehat{\theta}_t(s_t(a^*)) - \theta(s_t(a^*))\right)\right] + 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} .
\end{aligned} \tag{D.14}$$

We know that our mean estimator is unbiased. Therefore, we have that $\mathbb{E}\left[\widehat{\theta}_t(s_t(a^*))\right] = \theta(s_t(a^*))$ for any $t \in [T]$, meaning that the second term in the right-hand side of (D.14) is equal to zero.

On the other hand, we can apply Lemma D.2 to get the following bound for Δ that holds with probability at least $1 - \delta/2$ for any $\delta \in (0,1)$:

$$\Delta \le \min\left\{\frac{1}{2}\sum_{t=1}^{T} \varepsilon_t(S_t), T\right\} , \tag{D.15}$$

where we recall that $\varepsilon_t(s) = \sqrt{\frac{2}{N_t'(s)} \ln \frac{4ST}{\delta}}$. In particular, the inequality $\Delta \le T$ is true in general. By Lemma D.5, we can bound the right-hand side of (D.15) as

$$\frac{1}{2}\sum_{t=1}^{T} \varepsilon_t(S_t) \le \frac{7}{2}\sqrt{ST \ln \frac{4ST}{\delta}}$$

when conditioning on the event as in the statement of Lemma D.2. If we denote such an event as $\mathcal{E}$, we have that $\mathbb{P}\left(\overline{\mathcal{E}}\right) \le \delta/2$ and that $\mathbb{E}\left[\Delta \mid \mathcal{E}\right] \le \frac{7}{2}\sqrt{ST \ln\left(4ST/\delta\right)}$. As a consequence, we notice that

$$\mathbb{E}\left[\Delta\right] = \mathbb{E}\left[\Delta \mid \mathcal{E}\right]\mathbb{P}\left(\mathcal{E}\right) + \mathbb{E}\left[\Delta \mid \overline{\mathcal{E}}\right]\mathbb{P}\left(\overline{\mathcal{E}}\right) \le \frac{7}{2}\sqrt{ST \ln \frac{4ST}{\delta}} + \frac{\delta}{2}T \le 5\sqrt{ST \ln\left(2ST\right)} + 1$$

where in the last inequality we set $\delta = 2/T$. Since we assume that $S \ge 2$, we can easily observe that $\mathbb{E}\left[\Delta\right] \le 6\sqrt{ST \ln\left(2ST\right)}$. Plugging this into Equation (D.14) gives us

$$\mathbb{E}\left[R_T\right] \le 4\sqrt{KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST \ln(2ST)} . \tag{D.16}$$

At this point, we can proceed to the proof of the overall bound on the expected regret of `AdaMetaBIO`. The behaviour of `AdaMetaBIO` follows the same principle as before, but the switching condition is different:

$$\sqrt{8\mathfrak{D}_t \ln K} > 6\sqrt{ST \ln(2ST)} .$$

Similar to the analysis of `AdaMetaBIO` in Appendix D.2.5, we decompose the regret into

$$\mathbb{E}[R_T] \le \underbrace{\sum_{t=1}^{t^*} \mathbb{E}\left[\theta(S_t)\right] - \min_{a\in\mathcal{A}} \sum_{t=1}^{t^*} \theta(s_t(a)) +}_{R_{t^*}} \underbrace{\sum_{t=t^*+1}^{T} \mathbb{E}\left[\theta(S_t)\right] - \min_{a\in\mathcal{A}} \sum_{t=t^*+1}^{T} \theta(s_t(a))}_{R_{t^*:T}} ,$$

where t^* is the last round satisfying $\sqrt{8\mathfrak{D}_{t^*}} \le 6\sqrt{ST \ln\left(2ST\right)}$. Then, we have

$$\mathbb{E}[R_{t^*}] \le 4\sqrt{Kt^*} + \sqrt{8\mathfrak{D}_{t^*} \ln K} . \tag{D.17}$$

If $t^* = T$ then $R_{t^*} = R_T$ and we get the bound in Equation (D.17), where we note that $\sqrt{8\mathfrak{D}_T \ln K} \le 6\sqrt{ST \ln\left(2ST\right)}$ by definition of t^* in this case, and we can replace $\mathfrak{D}_T$ by $\mathcal{D}_T$. Otherwise, $t^* < T$ and we can apply the bound for `MetaBIO` from Equation (D.16), along with the fact that the total incurred delay after round t^* is upper bounded by $\mathcal{D}_\Phi$, in order to derive an upper bound for $\mathbb{E}[R_{t^*:T}]$ that is

$$\mathbb{E}[R_{t^*:T}] \le 4\sqrt{K(T-t^*)} + \sqrt{8\mathcal{D}_\Phi \ln K} + 6\sqrt{ST \ln(2ST)} . \tag{D.18}$$

Finally, if we use the fact that $\sqrt{8\mathfrak{D}_{t^*}} \le 6\sqrt{ST \ln(2ST)}$ (by definition of t^*) in Equation (D.17), and combine it with Equation (D.18), we conclude that

$$\mathbb{E}[R_T] \le 4\sqrt{2KT} + \sqrt{8\mathcal{D}_\Phi \ln K} + 2\min\Big\{6\sqrt{ST\ln(2ST)}, \sqrt{8\mathcal{D}_T \ln K}\Big\}\ ,$$

where we also used the fact that $\sqrt{t^*} + \sqrt{T - t^*} \le \sqrt{2T}$. □

D.3 Proofs for the Lower Bounds

Theorem 6.5. *Suppose that the action-state mapping is adversarial and the losses are stochastic and that $d_t = d$ for all $t \in [T]$. If $T \ge \min\{S, d\}$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret $\mathbb{E}\left[R_T\right] = \Omega\big(\sqrt{\min\{S, d\}T}\big)$.*

Proof of Theorem 6.5. Assume without loss of generality that $K = 2$ and let $\mathcal{S} \coloneqq \{h_1, \dots, h_S\}$ be the finite set of possible states. Let $S' \coloneqq \lfloor \min\{S/2, d\} \rfloor$ and let $I_1, \dots, I_T$ be the actions chosen by the considered algorithm. Split the T time steps into $m \coloneqq \lfloor T/S' \rfloor$ blocks $B_1, \dots, B_m$ of equal size S', eventually leaving $\le S' - 1$ extra time steps. We assume with no loss of generality that the last step corresponds to the end of the m-th block. The feedback formed by the losses of the actions chosen by the algorithm in a certain block is received only after the last time step of the same block since $S \le 2d$. Define $b_i \coloneqq (i-1)S' + 1$ for all $i \in [m]$. We assume that the learner receives *all* the realized losses $\ell_t(s_t(A))$ for all $t \in B_i$ and all $A \in \{1, 2\}$ at the end of each block, which means that we are in a full information setting, as this only helps the algorithm.

Now, we define a specific sequence of assignments from actions to states, and construct losses so that the expected regret becomes sufficiently large. Let $s_t(A) \coloneqq h_{2(t-b_i)+A}$ for all $t \in B_i$, all $i \in [m]$ and all $A \in \{1, 2\}$; this means that, for the first time step of any block, actions 1 and 2 will be assigned to states h_1 and h_2 respectively, then to h_3 and h_4 respectively in the next time step of the same block, and so on. Let $\varepsilon \coloneqq \frac{1}{4}\sqrt{\frac{S'}{2T\ln(4/3)}} \in \left[0, \frac{1}{4}\right]$ and let $\theta^{(A)} \in \mathbb{R}^2$ be a vector of mean losses such that $\theta^{(A)}_i \coloneqq \frac{1}{2} - \mathbb{I}\{i = A\}\,\varepsilon$, for each $A \in \{1, 2\}$. We simplify the notation with $\mathbb{E}_A\left[\cdot\right] \coloneqq \mathbb{E}\left[\cdot \,\middle|\, \theta^{(A)}\right]$ and $\mathbb{P}_A\left(\cdot\right) \coloneqq \mathbb{P}\left(\cdot \,\middle|\, \theta^{(A)}\right)$, where the conditioning on $\theta^{(A)}$ means that we sample losses for each state assigned to $i \in \{1, 2\}$ such that they are Bernoulli random variables with mean $\theta^{(A)}_i$. In particular, conditioning on $\theta^{(A)}$, we sample independent Bernoulli random variables $X^i_1, \dots, X^i_m$ with mean $\theta^{(A)}_i$, one for each block, for $i \in \{1, 2\}$. Then, the losses are defined as $\ell_t(s_t(i)) \coloneqq X^i_j$ for each $t \in B_j$ and each $j \in [m]$.

We can now proceed to show a lower bound for the expected pseudo-regret. Let T_i be the number of times the learner chooses action i over all T time steps. The expected pseudo-regret over the two instances determined by $\theta^{(k)}$ for $k \in \{1, 2\}$ adds up to

$$\mathbb{E}_1\left[R_T\right] + \mathbb{E}_2\left[R_T\right] = \varepsilon\left(2T - \mathbb{E}_1\left[T_1\right] - \mathbb{E}_2\left[T_2\right]\right)\ .$$

Following the standard analysis, we show that the difference $\mathbb{E}_2\left[T_2\right] - \mathbb{E}_1\left[T_2\right]$ is such that

$$\mathbb{E}_2\left[T_2\right] - \mathbb{E}_1\left[T_2\right] \le T \cdot d_{\mathrm{TV}}(\mathbb{P}_2, \mathbb{P}_1) \le T\sqrt{\frac{1}{2}D_{\mathrm{KL}}\left(\mathbb{P}_1 \,\|\, \mathbb{P}_2\right)}\ ,$$

where the last step follows by Pinsker's inequality.

Let $\lambda_i \coloneqq \{(I_t, \ell_t(S_t(1)), \ell_t(S_t(2))) \mid t \in B_i\}$ be the feedback set known to the learner by the end of block B_i, and let $\lambda^i \coloneqq (\lambda_1, \dots, \lambda_i)$ be the tuple of all feedback sets up to the end of block B_i. Denote by $\mathbb{P}_{k,i}(\cdot)$ the probability measure of feedback tuples λ^i conditioned on $\theta^{(A)}$. By the chain rule for the relative entropy, we can observe that

$$\begin{aligned}
D_{\mathrm{KL}}(\mathbb{P}_1 \,\|\, \mathbb{P}_2) &= \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1\left(\lambda^{i-1}\right) D_{\mathrm{KL}}\left(\mathbb{P}_{1,i}\left(\cdot \mid \lambda^{i-1}\right) \,\|\, \mathbb{P}_{2,i}\left(\cdot \mid \lambda^{i-1}\right)\right) \\
&\le \sum_{i=1}^{m} \sum_{\lambda^{i-1}} \mathbb{P}_1\left(\lambda^{i-1}\right) 16\varepsilon^2 \ln(4/3) \\
&= 16m\varepsilon^2 \ln(4/3) \ ,
\end{aligned}$$

where we used the fact that each relative entropy $D_{\mathrm{KL}}\left(\mathbb{P}_{1,i}\left(\cdot \mid \lambda^{i-1}\right) \,\|\, \mathbb{P}_{2,i}\left(\cdot \mid \lambda^{i-1}\right)\right)$ corresponds to the sum of the relative entropy between two Bernoulli distributions with means $1/2$ and $1/2-\varepsilon$ and that between Bernoulli distributions with means $1/2-\varepsilon$ and $1/2$, respectively, which is upper bounded by $16\varepsilon^2 \ln(4/3)$ for $\varepsilon \in [0, 1/4]$. This follows by an application of the chain rule for the relative entropy, as well as from the fact that the distribution of I_t is the same under both $\mathbb{P}_{1,i}\left(\cdot \mid \lambda^{i-1}\right)$ and $\mathbb{P}_{2,i}\left(\cdot \mid \lambda^{i-1}\right)$, for all $t \in B_i$ and any λ^{i-1}. Therefore, we have that

$$\mathbb{E}_2[T_2] - \mathbb{E}_1[T_2] \le 2\varepsilon T \sqrt{2m \ln(4/3)}$$

which also implies that

$$\mathbb{E}_1[R_T] + \mathbb{E}_2[R_T] \ge \varepsilon T \left(1 - 2\varepsilon \sqrt{2\frac{T}{S'} \ln(4/3)}\right) = \frac{\varepsilon T}{2} \ge \frac{1}{8}\sqrt{\frac{\lfloor S/2 \rfloor T}{2\ln(4/3)}} \ge \frac{1}{8}\sqrt{\frac{ST}{6\ln(4/3)}} \ ,$$

where we used the facts that $m \le T/S'$ and that $\lfloor S/2 \rfloor \ge S/3$ for any integer $S \ge 2$. This means that the expected pseudo-regret of the learner has to be $\frac{1}{16}\sqrt{\frac{ST}{6\ln(4/3)}}$ at least in one of the two instances. Now, for $S > 2d$ we use the same construction, but now we only use $2d$ states, which leads to the promised $\Omega(\sqrt{\min\{S,d\}T})$ lower bound. □

Theorem 6.6. *Suppose that the action-state mapping is adversarial, the losses are stochastic, and that $d_t = d$ for all $t \in [T]$. If $T \ge d+1$ then there exists a distribution of losses and a sequence of action-state mappings such that any (possibly randomized) algorithm suffers regret*

$$\mathbb{E}[R_T] = \Omega\Big(\min\Big\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\Big\}\Big) \ .$$

Proof of Theorem 6.6. Let $S' \coloneqq \min\left\{\lfloor \frac{S}{2} \rfloor, \lfloor \frac{T}{d+1} \rfloor\right\} \ge 1$. We consider the first $(d+1)S'$ rounds of the game and divide them into S' blocks $B_1, \dots, B_{S'}$ of same length $d+1$. In this way, we ensure that the feedback for any time step in some block is revealed to the learner only after its final round.

Without loss of generality, we can assume that the learner observes all the losses of one block immediately after its last time step; this only helps the learner since they would observe only the incurred losses at possibly later rounds otherwise. We can further simplify the problem by assuming that losses are deterministic functions of the states, i.e., $\ell_t \equiv \theta$ for every round t. This also means that the problem turns into an easier, full-information version of our problem with deterministic losses. Now, let the adversary choose the action-state mappings such that for each block index i and each action $a \in \mathcal{A}$, $S_t(a) = S_{t'}(a) \in \{s_{2i-1}, s_{2i}\}$ for all $t, t' \in B_i$. Furthermore, we assume that the losses are chosen such that $\theta(s_{2i-1}) \in \{0, 1\}$ and $\theta(s_{2i}) = 1 - \theta(s_{2i-1})$ for all $i \in [S']$. In this construction, the learner cannot obtain any useful information from the states of a block because of the delays. Moreover, the states observed in one block are not observed again in the other blocks.

It thus suffices to prove a lower bound for a standard full information game with S' rounds and loss range $[0, d+1]$. Hence, we can conclude that the expected regret of any algorithm has to be

$$\mathbb{E}\left[R_T\right] = \Omega\left((d+1)\sqrt{S'}\right) = \Omega\left(\min\left\{(d+1)\sqrt{S}, \sqrt{(d+1)T}\right\}\right) . \qquad \square$$

D.4 Action-State Mappings and Loss Means Used in the Experiments

Table D.1 and Table D.2 describe the instances used to generate the data for the experiments of Section 6.6.

Mean loss	$s=1$	$s=2$	$s=3$
$\theta(s)$	0.2	0.4	0.8

Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	0.8	0.1	0.1
$a=2$	0.4	0.5	0.1
$a=3$	0.3	0.7	0.0
$a=4$	0.5	0.3	0.2

Table D.1: Mean losses and stochastic action-state mapping for Experiment 1 in Section 6.6.

Mean loss	$s=1$	$s=2$	$s=3$
$\theta(s)$	0	1	1

Environment 1

Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	0.06	0.47	0.47
$a=2$	0	0.50	0.50
$a=3$	0	0.50	0.50
$a=4$	0	0.50	0.50

Environment 2

Mapping	$P(1\|a)$	$P(2\|a)$	$P(3\|a)$
$a=1$	1	0	0
$a=2$	0.94	0.03	0.03
$a=3$	0.94	0.03	0.03
$a=4$	0.94	0.03	0.03

Table D.2: Mean losses and stochastic action-state mappings for Experiment 2 in Section 6.6.

Appendix E

Proof Details for Chapter 9

E.1 Boosting Decision Trees with Bounded Depth

Theorem 9.2. *Let $\mathcal{X}$ be any domain. For any concept c and any hypothesis class $\mathcal{H}$ over $\mathcal{X}$, if there exist $\gamma \in (0, \frac{1}{2})$ and $d \in \mathbb{N}$ such that $\operatorname{depth}^c_{\mathcal{H}}\big(\frac{1}{2} - \gamma \mid P\big) \leq d$ for all $P \in \mathcal{P}(\mathcal{X})$, then $\operatorname{depth}^c_{\mathcal{H}}(\varepsilon \mid P) \leq \frac{d}{2\gamma^2} \log \frac{1}{2\varepsilon}$ for all $P \in \mathcal{P}(\mathcal{X})$ and all $\varepsilon > 0$.*

We use a surrogate loss $G(q) \coloneqq \sqrt{q(1-q)}$, where $0 \leq q \leq 1$. Since $\min\{q, 1-q\} \leq G(q)$, the surrogate loss bounds from above the classification error of the majority vote. For a distribution $P \in \mathcal{P}(\mathcal{X})$, let $G_P(c) \coloneqq G\big(P(c=1)\big)$. Let the conditional surrogate loss of $f \colon \mathcal{X} \to \{0,1\}$ be

$$G_P(c \mid f) \coloneqq P(f=0)G\big(P(c=1 \mid f=0)\big) + P(f=1)G\big(P(c=1 \mid f=1)\big) \ . \tag{E.1}$$

Finally, given a decision tree T with leaves $\mathcal{L}(T)$, define the conditional surrogate loss of T as

$$H_P(c \mid T) \coloneqq \sum_{z \in \mathcal{L}(T)} P(z) G(p_{c|z}) \ , \tag{E.2}$$

where $P(z)$ is the probability that $x \sim P$ is mapped to leaf z in the tree T and $p_{c|z} \coloneqq P(c = 1 \mid z)$. Our goal is to construct an $\mathcal{H}$-based decision tree T such that $H_P(c \mid T) \leq \varepsilon$, implying that $L_P(T, c) \leq \varepsilon$ because $G\big(p_{c|z}\big)$ bounds from above the probability that $T(x) \neq c(x)$ conditioned on x being mapped to z in T.

Our variant of `TopDown`, called `TopDownLBL` (TopDown Level-By-Level), starts from a single-leaf tree T with a majority-vote label and works in phases. In each phase, we replace each leaf $z \in \mathcal{L}(T)$ of the current tree T with a suitably chosen $\mathcal{H}$-based d-depth tree T_z using the same criterion as `TopDown`. The main difference is that the weak learners adopted by `TopDownLBL` consist of $\mathcal{H}$-based trees of depth bounded by d, which generalize from the individual decision stumps of $\mathcal{H}$ as in `TopDown` (corresponding to the case $d = 1$). Hence, at the end of each phase, the depth of T increases by at most d. The algorithm stops if and when $H_P(c \mid T) \leq \varepsilon$.

We use the two following lemmas.

Lemma E.1 (Takimoto and Maruoka (2003, Lemma A.1))**.** *Let P be a balanced distribution, i.e., $P(c = 1) = P(c = 0) = \frac{1}{2}$. Let $f \colon \mathcal{X} \to \{0,1\}$ be such that $L_P(f, c) \leq \frac{1}{2} - \gamma$ for some $\gamma \in (0, \frac{1}{2})$. Then, $G_P(c \mid f) \leq (1 - 2\gamma^2)G_P(c)$.*

Lemma E.2 (Takimoto and Maruoka (2003, Proposition 5)). *Let P be a distribution and P' its balanced version. If $G_{P'}(c \mid h) \leq (1-\beta) G_{P'}(c)$ for some $\beta > 0$ then $G_P(c \mid h) \leq (1-\beta) G_P(c)$.*

Proof of Theorem 9.2. Our algorithm `TopDownLBL` can be equivalently viewed as building a $\mathcal{H}'$-based tree T', where $\mathcal{H}'$ is the class of $\mathcal{H}$-based d-depth trees. Any $\mathcal{H}'$-based tree T' can be transformed into a $\mathcal{H}$-based tree T in a top-down fashion simply by listing the nodes at each level of T' starting from the root, and iteratively replacing every decision stump $h' \in \mathcal{H}'$ with the corresponding $\mathcal{H}$-based tree $T_{h'}$. Then, each leaf $z \in \mathcal{L}(T_{h'})$ of $T_{h'}$ is replaced by copies of the left or right subtree of the decision stump h' in T' based on the values (0 or 1) of the label λ_z of z. Clearly, the depth of T is at most d times the depth of T'.

We now bound the drop in $H_P(c \mid T')$ when a leaf z in the $\mathcal{H}'$-based tree T' is replaced by a decision stump in $\mathcal{H}'$. Let P the distribution over $\mathcal{X}$ conditioned on x being mapped to z and let P' its "balanced" version satisfying $P'(c=1) = P'(c=0) = \frac{1}{2}$. Because of our weak learning assumption, we know there exists $h'_z \in \mathcal{H}'$ with error at most $1/2 - \gamma$ on P'. By Lemma E.1, $G_{P'}(c \mid h'_z) \leq (1-2\gamma^2) G_{P'}(p'_{c|z})$, where $p'_{c|z} = \frac{1}{2}$ because of the balanced property of P'. Hence, by Lemma E.2,

$$G_P(c \mid h'_z) \leq (1-2\gamma^2) G_P(p_{c|z}) \,. \tag{E.3}$$

Let T'_z be the tree T' in which we replaced a leaf $z \in \mathcal{L}(T')$ with the decision stump $h'_z \in \mathcal{H}'$. Using Equation (E.3),

$$H_P(c \mid T') - H_P(c \mid T'_z) = \big(G_P(p_{c|z}) - G_P(c \mid h'_z)\big) P(z) \geq 2\gamma^2 G_P(p_{c|z}) P(z) \,. \tag{E.4}$$

Now let T'_i be the tree after the algorithm has run for i phases. Using the above inequality for each $z \in \mathcal{L}(T'_i)$, we obtain

$$H_P(c \mid T'_i) - H_P(c \mid T'_{i+1}) \geq \sum_{z \in \mathcal{L}(T'_i)} 2\gamma^2 G_P(p_{c|z}) P(z) = 2\gamma^2 H_P(c \mid T'_i) \,. \tag{E.5}$$

Hence, after m phases,

$$L_P(T'_m, c) \leq H_P(c \mid T'_m) \leq \big(1-2\gamma^2\big)^m H_P(c \mid T'_0) \leq \frac{1}{2} e^{-2m\gamma^2} \,, \tag{E.6}$$

where T'_0 is the initial tree consisting of a single leaf z and, in the last inequality, we used the fact that $H_P(c \mid T'_0) = G_P(p_{c|z}) \leq \frac{1}{2}$ and the inequality $1 - x \leq e^{-x}$. The proof is concluded by noting that $\frac{1}{2} e^{-2m\gamma^2} \leq \varepsilon$ for $m \geq \frac{1}{2\gamma^2} \log \frac{1}{2\varepsilon}$. □

We remark that we recover the standard setting of boosting decision trees when $d = 1$. In this special case, our result matches the depth lower bound mentioned by Kearns and Mansour (1999), while guaranteeing a $\mathcal{O}\big(2^{\mathrm{depth}^{\mathrm{c}}_{\mathcal{H}}(\varepsilon \mid P)}\big) = \mathcal{O}\big((1/\varepsilon)^{1/(2\gamma^2)}\big)$ tree-size upper bound that is analogous to the ones by Kearns and Mansour (1999) and Takimoto and Maruoka (2003).

E.2 Further Proofs for the Algebraic Characterization

E.2.1 Algebraic Characterization

Lemma 9.1. *Let $\mathcal{X}$ be any domain and $\mathcal{H} \subseteq 2^{\mathcal{X}}$. Then, $\mathrm{clos}(\sigma(\mathcal{H})) = \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$.*

Proof of Lemma 9.1. We begin with the proof of the first identity in the statement. The inclusion $\mathrm{clos}(\mathrm{Alg}(\mathcal{H})) \subseteq \mathrm{clos}(\sigma(\mathcal{H}))$ immediately follows by definition of closure. We now show that the converse is also true. Let $T \in \mathrm{clos}(\sigma(\mathcal{H}))$. Fix a distribution $P \in \mathcal{P}(\mathcal{X})$ and $\varepsilon > 0$. By definition of closure, there exists a sequence $A_1, A_2, \ldots \in \sigma(\mathcal{H})$ such that $\lim_{i\to\infty} P(A_i \triangle T) = 0$. Consequently, for every $\varepsilon > 0$ there exists some $i \in \mathbb{N}^+$ such that $P(A_i \triangle T) \leq \varepsilon$. Thus, we can assume without loss of generality that the sequence $(A_i)_{i\in\mathbb{N}^+}$ satisfies $P(A_i \triangle T) \leq \varepsilon_i$ for the choice $\varepsilon_i := 2^{-i}$, for each $i \in \mathbb{N}^+$ (as we can select such a subsequence). Denote the restriction of P to $\sigma(\mathcal{H})$ as $P\big|_{\sigma(\mathcal{H})}$, that is $P\big|_{\sigma(\mathcal{H})} \colon \sigma(\mathcal{H}) \to \mathbb{R}_{\geq 0}$ and $P\big|_{\sigma(\mathcal{H})}(A) := P(A)$ for all $A \in \sigma(\mathcal{H})$. It is well known that, for each i, we can select an element $B_i \in \mathrm{Alg}(\mathcal{H})$ with $P\big|_{\sigma(\mathcal{H})}(B_i \triangle A_i) \leq \varepsilon_i$ (see, e.g., Halmos (2013, Theorem D, Section 13)); hence, $P(B_i \triangle A_i) \leq \varepsilon_i$. By the triangle inequality $P(T \triangle B_i) \leq 2\varepsilon_i = 2^{-i+1}$ for any i, which also implies that $\lim_{i\to\infty} P(T \triangle B_i) = 0$ for the sequence $(B_j)_{j\in\mathbb{N}^+}$ in $\mathrm{Alg}(\mathcal{H})$. Therefore, $T \in \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$. □

E.2.2 Algebraic Characterization for Countable Domains

Theorem 9.5. (Characterization for VC classes and countable domains) *Let $\mathcal{X}$ be any countable domain, let c be any concept, and let $\mathcal{H}$ be a VC hypothesis class over $\mathcal{X}$. Then:*

1. *c is approximable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H})$.*
2. *c is approximable but not interpretable by $\mathcal{H}$ if and only if $c \in \sigma(\mathcal{H}) \setminus \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*
3. *c is uniformly interpretable by $\mathcal{H}$ if and only if $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$.*

Proof of Theorem 9.5. Item 1 follows from Lemma E.3 and Theorems 9.3 and 9.4, whereas items 2 and 3 follow from Lemma E.5. □

Lemma E.3. *Let $\mathcal{X}$ be any countable domain and $\mathcal{H}$ be any hypothesis class over $\mathcal{X}$. Then,* $\mathrm{clos}(\mathrm{Alg}(\mathcal{H})) = \sigma(\mathcal{H})$*.*

Proof. Clearly $\sigma(\mathcal{H}) \subseteq \mathrm{clos}(\sigma(\mathcal{H})) = \mathrm{clos}(\mathrm{Alg}(\mathcal{H}))$ by Lemma 9.1. Now we prove the converse. Let $A \in \mathrm{clos}(\sigma(\mathcal{H}))$ and let $P \in \mathcal{P}(\mathcal{X})$ such that $P(x) > 0$ for all $x \in \mathcal{X}$; note that P exists as $\mathcal{X}$ is countable and it also means that $\mathrm{supp}(P) = \mathcal{X}$. By definition of $\mathrm{clos}(\sigma(\mathcal{H}))$, there exists a sequence $(A_i)_{i\in\mathbb{N}^+}$ in $\sigma(\mathcal{H})$ such that

$$\lim_{i\to\infty} P(A \triangle A_i) = 0 \ . \tag{E.7}$$

By selecting an appropriate subsequence, we can assume $P(A \triangle A_i) \leq 2^{-i}$ for all $i \in \mathbb{N}^+$ without loss of generality. Define

$$B_i := \bigcap_{j\geq i} A_j \qquad \forall i \in \mathbb{N}^+ \tag{E.8}$$

and observe that $B_i \in \sigma(\mathcal{H})$ for each $i \in \mathbb{N}^+$. Note that

$$P(A \triangle B_i) = P\Bigg(A \triangle \bigcap_{j\geq i} A_j\Bigg) \leq \sum_{j\geq i} P(A \triangle A_j) \leq 2^{-i+1} \ . \tag{E.9}$$

Note also that $B_i \subseteq A$ for all $i \in \mathbb{N}^+$. Suppose indeed this was not the case, then $B_i \setminus A \neq \emptyset$. Hence, by definition of B_i, there exists some $x \in A_j \setminus A$ for all $j \geq i$. Since $P(x) > 0$ by the choice of P, we have the contradiction

$$\lim_{i\to\infty} P(A\triangle A_i) \geq P(x) > 0\ . \tag{E.10}$$

Now consider the set

$$B \coloneqq \bigcup_{i\in\mathbb{N}^+} B_i = \lim_{i\to\infty} B_i\ . \tag{E.11}$$

Note that by construction $B \in \sigma(\mathcal{H})$.[*] Moreover, since $B_i \subseteq B_{i+1}$ and $B_i \subseteq A$ for all $i \in \mathbb{N}^+$, we have that the sequence $(A\triangle B_i)_{i\in\mathbb{N}^+}$ is downward monotone and thus

$$P(A\triangle B) = P\Bigg(\bigcap_{i\in\mathbb{N}^+} A\triangle B_i\Bigg) = \lim_{i\to\infty} P(A\triangle B_i) = 0\ . \tag{E.12}$$

Given that P has full support, this implies $A = B$. □

Definition E.1. *Let $\mathcal{X}$ be any set. A sequence $(h_i)_{i\in\mathbb{N}}$ in $2^{\mathcal{X}}$ is* pointwise convergent *to $h \in 2^{\mathcal{X}}$ if*

$$\forall x \in \mathcal{X} \quad \exists i_x \in \mathbb{N} : \ \forall i \geq i_x \qquad x \in h_i \iff x \in h\ . \tag{E.13}$$

Proposition E.1. *If $\mathcal{X}$ is countable then every infinite sequence $(h_i)_{i\in\mathbb{N}}$ in $2^{\mathcal{X}}$ contains an infinite subsequence that is pointwise convergent.*

Let $\mathcal{H} \subseteq 2^{\mathcal{X}}$ and let $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H})$ be the family of all subsets of $\mathcal{X}$ that are the pointwise limit of some sequence in $\mathcal{H}$. Clearly $\mathcal{H} \subseteq \mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \subseteq 2^{\mathcal{X}}$.

Lemma E.4. *If $\mathcal{X}$ is countable then $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) = \mathrm{clos}(\mathcal{H})$.*

Proof. To see that $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \subseteq \mathrm{clos}(\mathcal{H})$, recall the definition of pointwise convergence, and note how it implies that if a sequence $(h_i)_{i\in\mathbb{N}}$ converges pointwise to h then $\lim_{i\to\infty} P(h_i\triangle h) = 0$ for every $P \in \mathcal{P}(\mathcal{X})$. To see that $\mathrm{clos}_{\mathrm{pw}}(\mathcal{H}) \supseteq \mathrm{clos}(\mathcal{H})$, choose any sequence $(h_i)_{i\in\mathbb{N}}$ that converges to some $h \in \mathrm{clos}(\mathcal{H})$ under an appropriate distribution $P \in \mathcal{P}(\mathcal{X})$ such that $\mathrm{supp}(P) = \mathcal{X}$ (which exists as $\mathcal{X}$ is countable); observe that this implies the pointwise convergence of $(h_i)_{i\in\mathbb{N}}$ to h. □

Lemma E.5. *If $\mathcal{X}$ is countable and $\mathcal{H}$ is a VC class over $\mathcal{X}$, then $\mathrm{clos}(\mathrm{Alg}_d(\mathcal{H})) \subseteq \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$ for every $d \in \mathbb{N}$.*

Proof. Let $d \in \mathbb{N}$ and $c \in \mathrm{clos}(\mathrm{Alg}_d(\mathcal{H}))$. By Lemma E.4, $c \in \mathrm{clos}_{\mathrm{pw}}(\mathrm{Alg}_d(\mathcal{H}))$, so there exists an infinite sequence of trees $(T_i)_{i\in\mathbb{N}}$ in $\mathrm{Alg}_d(\mathcal{H})$ that converge pointwise to c. Without loss of generality, we may assume that every T_i is a complete tree of depth d.[†] Now consider the sequence $(h_i^1)_{i\in\mathbb{N}}$ of decision rules used by the first node (say, the root) of those trees. By Proposition E.1 there is an infinite subsequence $(h_{i_j}^1)_{j\in\mathbb{N}}$ that is pointwise convergent to some $h^1 \in \mathcal{H}$. Now consider the infinite sequence of trees $(T_{i_j})_{j\in\mathbb{N}}$, and repeat the argument for the second node (say, a child of the root corresponding to a specific output of the decision stump at the root). By repeating the argument $2^d - 1$ times (one for every internal node of the trees) we obtain an infinite sequence $(T_i^*)_{i\in\mathbb{N}}$ of trees

[*] In particular, $B = \liminf_{i\to\infty} A_i$.

[†] One can always complete T_i using internal nodes that hold, e.g., the decision rule of the root.

in $\mathrm{Alg}_d(\mathcal{H})$ that converge pointwise to c and such that at every node v the decision rules converge pointwise to some h^v. Now let T^* be the decision tree obtained by using h^v as decision rule at v. We observe that $T^* = c$. Let $x \in \mathcal{X}$. By Definition E.1, for each node v there exists i_x^v such that $x \in h_i^v$ iff $x \in h^v$ for every $i \geq i_x^v$, where h_i^v is the stump used at v by T_i^*. By letting $i_x \coloneqq \max_v i_x^v$ it follows that $x \in h_i^v$ iff $x \in h^v$ for every $i \geq i_x$ and all nodes v simultaneously. Therefore all trees T_i^* with $i \geq i_x$ send x to the same leaf, and moreover that leaf remains the same if we use h^v at v. Note also that, since $(T_i^*)_{i\in\mathbb{N}}$ is infinite, then we can assume that every leaf predicts the same label in all T_i^* (since there is certainly an infinite subsequence that satisfies such a constraint). It follows that $(T_i^*)_{i\in\mathbb{N}}$ converges pointwise the tree T^* that uses the limit stump h^v at v. But the labeling of $(T_i^*)_{i\in\mathbb{N}}$ converges pointwise to c, too. We conclude that $T = T^*$. Finally, note that by construction $h^v \in \mathrm{clos}_{\mathrm{pw}}(\mathcal{H})$, and thus $h^v \in \mathrm{clos}(\mathcal{H})$ by Lemma E.4, for all v; hence, $T^* \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$. It follows that $c \in \mathrm{Alg}(\mathrm{clos}(\mathcal{H}))$. □

E.3 Remarks on the Graded Complexity Measure

In Section 9.7 we demonstrated more general guarantees for any graded complexity measure Γ, given any domain $\mathcal{X}$ and any hypothesis class $\mathcal{H}$ over $\mathcal{X}$. Observe that, when $\mathcal{H}$ is a non-VC class, item $(3')$ of Theorem 9.6 states an upper bound on the Γ-complexity rate of order $\mathcal{O}\big(\frac{1}{\varepsilon^d}\big)$ for a constant $d \in \mathbb{N}$. This bound is indeed larger compared to the previous guarantee of $\mathcal{O}(\log(1/\varepsilon))$ on the depth of $\mathcal{H}$-based decision trees (Theorem 9.1) and it has to do with the generality of the definition of graded complexity measure.

Keeping this in mind, we remark that it is possible to recover the $\mathcal{O}(\log(1/\varepsilon))$ Γ-complexity rate bound under a stronger assumption on the graded complexity measure Γ. In particular, it is sufficient for Γ to satisfy

$$\Gamma(f_1 \cup f_2) \leq 1 + \max\{\Gamma(f_1), \Gamma(f_2)\} \qquad \forall f_1, f_2 \in \mathrm{Alg}(\mathcal{H}) \; . \tag{E.14}$$

Note that this condition is satisfied when Γ corresponds to the depth of $\mathcal{H}$-based decision trees. For example, consider a similar representation of trees as in Equation (9.15) using directly $\mathcal{H}$ for the decision rules of the internal nodes.

Thus, we can follow the same steps as in the proof of Theorem 9.6 with a particular focus on the construction of A from the decision tree T in Equation (9.15). It immediately follows that $\Gamma(A_v^T) \leq 3 + \Gamma(A_v) + \max\big\{\Gamma(A_u^T), \Gamma(A_w^T)\big\}$ for any internal node $v \notin \mathcal{L}(T)$, where u and w are, respectively, the left and right child of v. Now, let $\rho(z) \subseteq \mathcal{V}(T)$ be the nodes along the path from the root of T to the leaf $z \in \mathcal{L}(T)$. We can thus show that

$$\Gamma(A) = \mathcal{O}\left(\max_{z\in\mathcal{L}(T)} \sum_{v\in\rho(z)} (\Gamma(A_v) + 1)\right) = \mathcal{O}\big((k+1)\cdot \mathrm{depth}(T)\big) = \mathcal{O}\Big(\log\frac{1}{\varepsilon}\Big) \; , \tag{E.15}$$

where we used the fact that T has $\mathrm{depth}(T) \leq \frac{1}{2\gamma^2}\log\frac{1}{2\varepsilon}$ and that $\Gamma(A_v) \leq k$ for any internal node v of T.