**Title**

Learning from an unknown environment



**Author**

Muthukumar, Vidya

**Publication Date**

2020

Peer reviewed|Thesis/dissertation

Learning from an unknown environment

by

Vidya K Muthukumar

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anant Sahai, Chair
Professor Peter L. Bartlett
Professor Jean C. Walrand
Professor Shachar Kariv

Fall 2020

Learning from an unknown environment

# Abstract

This dissertation initiates fundamental lines of inquiry into better understanding learning from data provided by other agents who have a possible strategic incentive. In many use cases of modern machine learning (ML), these agents will not be acting in isolation, and it is critical for them to directly interact with other strategic agents. For example, several pioneering works in cognitive radio designed multi-agent mechanisms whose equilibrium outcome was efficient spectrum sharing among multiple cognitive radio agents. However, the attainment of these equilibria requires *co-design* among the agents: they have to know each other's utility functions and basic strategic nature, which could be stochastic, adversarial, competitive, or cooperative. In the conceptualized utopia of spectrum sharing, both of these will be unknown a-priori and will need to be *learned* through repeated interaction. Nor does this scenario arise only in cognitive radio: applications of swarm robotics, reinforcement learning and e-commerce all involve the intersection of principles of ML with multi-agent systems that are not co-designed.

The ensuing twin questions of *how agents should learn from strategically generated data*, as well as *how such strategic behavior will manifest*, are well-posed and non-trivial even under the simplest possible instance cases of ML, such as predicting a binary sequence generated by an unknown environment. We consider the first three categories of strategic nature: stochastic, adversarial, and competitive. The first part of this thesis designs algorithms for "optimal" learning in an unknown environment, where the notion of optimality is defined as being able to adapt to the nature of the environment *on-the-fly* without knowing this nature beforehand. The on-the-fly nature of this goal is formalized in the classical framework of online learning. While the traditional goal of online learning is regret minimization *with respect to a given model class*, I motivate that adapting the model class, itself, in an online fashion is essential to transition from guarantees on regret to guarantees on reward. Accordingly, we design robust approaches, inspired by seminal approaches to purely stochastic model selection, to work in both stochastic and adversarial environments for online learning with full-information and limited-information feedback.

The second part of this thesis considers a strategic, but non-adversarial agent generating the data that is being used for learning. Such an agent is typically selfish and rational, rather than being simply malicious — thus, their behavior manifests in a more complex manner than an adversarial agent's. I introduce a new *frequentist* framework to approximately express such an agent's incentives and trade-offs involved in reaching the Stackelberg equilibrium of the ensuing non-zero-sum game. This is in agreement with the classical Bayesian-repeated-game asymptotic theory, now with constructive strategies for both players. Interestingly, through this framework we can show that the agent is incentivized to reveal, not obfuscate, her information to the learner. This thesis concludes by showing a surprising *divergent* outcome in day-to-day behavior that is fundamental to the property of no-regret online learning when deployed in multi-player, game-theoretic environments. This suggests a possible re-examination of learning dynamics, inspired by behavioral game theory, in future work.

To Amma, Appa, Ashwin, Gayathri, and how could I forget the Paatis.

# Contents

# List of Tables

# Acknowledgments

First and foremost, I would like to thank my advisor, Professor Anant Sahai, who is one of the most creative researchers whom I have had the fortune to work with. Thank you for helping me unlock my own inner creativity that is essential to doing exciting and enjoyable research. I especially cherish the wide-ranging freedom that you gave me in deciding where my research passion lay, at a critical juncture in my PhD. Not only that, you unhesitatingly joined me on the ride — your readiness to always try new things and stay curious taught me what doing research is all about. Working with you has been challenging, unexpected, and fun in turn, and always a pleasure — I can't thank you enough. I also wish to thank you for always lending an ear on the many occasions I've brought career-related questions to your desk instead of theorems; it means a lot. You have been a truly fantastic mentor.

Special thanks goes to Professors Peter Bartlett and Jean Walrand, the EECS members of my thesis committee who have greatly influenced the content in this dissertation. Collaborations with Peter and his research group have played an out-sized role in enriching my research. Peter, I will always appreciate your readiness to hear about new ideas and encourage them while also giving me access to the wealth of information needed to truly unlock their potential. Since my third year, I also found a second research home in your wonderful research group, and will always remember the group meetings and group outings extremely fondly. Jean, your group's seminars on "network economics" and your own groundbreaking work in this area were a big influence on my research interests early in my PhD. You also have a remarkable ability to use deceptively simple-looking models to get at the heart of a research problem, which has been a direct inspiration for my own research ethos.

Deciding to work on the intersection of learning and games initially felt like a bold step, and Professor Shachar Kariv's input on my research directions early on gave me immeasurable encouragement to pursue my interests. Other faculty I've especially enjoyed my interaction with are Martin Wainwright and Kannan Ramchandran. Martin, your sharpness and clarity of thought is reflected in your awesome foundational book on high-dimensional statistics; without this resource, I'm sure I would have been lost in the wilderness for a much longer time. The many barbs and witticisms traded with Ashwin have provided me with levity when it's often needed. Kannan, you have always had a smile and a hello ready for graduate students in the lab, and it seems like a small thing, but it means a lot.

I wish to thank all my teachers — Venkat Anantharam, Thomas Courtade, Michael Jordan, Ali Niknejad, Christos Papadimitriou, Anant Sahai, Martin Wainwright, and Jean Walrand — for showing me why Berkeley EECS is such an awesome place to be at. I especially would like to thank Anant and Ali, who instructed me as a TA in EE16A and set me on the right path to learn the challenging art of teaching myself, and Christos, whose stellar treatment of algorithmic game theory played a large role in shaping my interests.

Graduate research is hardly a solo effort, and I am grateful for all of the awesome research collaborations, both past and present, that have not only been fruitful and engaging, but always remind me how supported I really am. Thanks to (last-name alphabetical order) Peter Bartlett, Mikhail (Misha) Belkin, Niladri Chatterji, Thomas Courtade, Angel Daruna,

Kate Harrison, Daniel Hsu, Vijay Kamble, Brian Kingsbury, Abhishek Kumar, Cheng Mao, Saska Mojsilovic, Adhyyan Narang, Ashwin Pananjady, Tejaswini Pedapati, Soham Phade, Nalini Ratha, Mitas Ray, Anant Sahai, Prasanna Sattigeri, Vignesh Subramanian, Samuel Thomas, Kush Varshney, Kailas Vodrahalli, Martin Wainwright, and Chai-Wah Wu. Special thanks goes to: Niladri and Soham for being my partners-in-crime in online learning and games; Misha and Daniel for all the fun conversations about “modern machine learning"; Kate for being the best graduate student mentor I could have asked for and a continual inspiration; and Kush Varshney and Saska Mojsilovic at IBM Research for a truly holistic, genre-bending research internship at the Science for Social Good program. Shout-outs to my firecracker mentees, Mitas, Kailas and Adhyyan — I am fortunate to have worked with them and seen them grow. I must specially thank Andrew Thangaraj at IIT-Madras for introducing me to the beauty of information theory early in my undergraduate years (as well as supporting my academic visit to Munich together with Professor Gerhard Kramer — thank you both!), and Babak Hassibi for generously supporting me for a wonderful summer at Caltech. These experiences really inspired me to actually pursue a PhD in EECS.

Berkeley has been special in facilitating memorable collaborations outside of research. My partners-in-crime in course-developing EE16A were not only super fun to work with, but also inspired much of my teaching philosophy. Women in Computer Science and Engineering (WICSE) has played a tremendous role in supporting me and other graduate women during my time here. Special thanks to my bad-ass co-president Orianna DeMasi and WICSE’s grand mom, Sheila Humphreys — you both are amazing. Shirley Salanio has been a wonderful human being and a pillar of support during all six years at Berkeley. Shirley, I marvel at how you respond to as many emails a day with as many smiley faces as you do. :) Thanks also goes to Kim Kail for patiently handling our numerous reimbursement requests, always promptly and with a smile.

A big reason I’ve considered myself fortunate to be at Berkeley is the presence of the glorious Simons Institute next door. The fantastic workshops, and vibrant research culture on display, really showed me how much of a team effort research is. I am grateful to Nika Haghtalab, Vijay Kamble, and Thodoris Lykouris for encouraging me to pursue a career in academia. Thanks also to Gireeja Ranade for meticulously-but-encouragingly critiquing at least two important research talks!

I’ve been lucky to have a “big tent" lab as my working environment, and it took me six whole years to become tired of my office. For this, thanks goes to all BLISSers past and present. Special thanks to: Kabir for being the height of cool; Kate for excellent research and life hacks; Vijay for showing us that grad school could actually be fun; Orhan for being a perpetually entertaining desk-mate and an even better friend; Vasuki for starting required conversations on social justice; Rashmi and Nihar for all the words of wisdom and frisbee; Fanny for all the concerts you dragged us to, and all the triathlons you failed to drag us to but inspired us nonetheless, and most of all, for the picture of Ashwin’s proposal to me that we will never forget; and Yuting for being the sometimes-needed yin to Fanny’s Yang. Thanks also to Peter Bartlett’s awesome group — especially Kush, Niladri, Xiang, and Aldo — for all the good times and jokes, usually at the expense of each other (Kush). Thanks

to Po-Ling Loh and Aaditya Ramdas, whose fearlessness and can-do attitude inspired me in both long-distance cycling and research pursuits; and Mayur, Siva, Nadia and Despina, who have all stayed buddies long after the AIDS/LifeCycle ride finished. Old-and-new friends have continually reminded me that life exists outside of graduate school — Poorna Kumar, Karuna Agarwal, Neha Nathan, Ramya Anand, Karishma Sureka, Vishwanath Saragadam, Amod Mital, Vaishnavi Surendra, Sanjay Guruprasad, thanks for all the food, drink and memories. Thanks to my music gurus, Asha aunty and Savitha aunty, for helping me keep up my passion in music, always a pleasant distraction from research.

As we tackle the many peaks and valleys of graduate school, we realize just how much would be impossible without the unconditional love and support of family. I have been indescribably lucky to have my parents, Vijaya and Muthu, close by in the Bay Area. Thank you Appa for introducing me to K'Nex, for spending many hours doing math puzzles with me and for always believing in me; and thank you Amma for instilling in me the value of kindness, for the endless supply of home cooked meals which helped me survive grad school, and for always putting my needs before yours. You both have also been an integral part of my musical journey, from spotting my early passion for classical piano and Carnatic vocal music to nurturing it with love and effort through all these years. During my time at graduate school, Glorious Gayath graduated from one east coast behemoth and went to another to start her own grad school journey. I often joke that she walked through undergrad as a grad student — and she has always lent an ear when I wanted it. Little sister, thanks for all the lemon drizzle cakes, and being my best girlfriend over the last six years.

I have been fortunate to have had the grace and grit of my two wise grandmothers, Bombay Paati and Madras Paati, in my life — thank you for everything. Thanks to my late Periappa and Thatha who showered me with more love and affection than I thought was possible — I love you, and I miss you, and it breaks my heart that you cannot be here at the end of my PhD journey.

Special thanks to my parents-in-laws, Indira Aunty and Swathi Uncle, for all their encouragement and support over the years from near and afar — the affectionately delivered *kodubale*, Sunday crosswords, and *Maathrubhootham* have given me a window into your vibrant household. Thanks to sister-in-law Kasturi for the malamute videos, cooking projects, and lending her experienced writer's eye to my academic statements. We've enjoyed your many visits to the Best Coast.

I married my best friend, Ashwin Pananjady, midway through graduate school. We met way back in undergrad, and I could write a whole thesis on his sparkling persona and how it's impacted every aspect of my life — from "producing" theorems, to cycling across California, to your incredible proposal on Mt Rainier, to our grand Indian wedding, to going on the faculty job market together, it has been an incredible partnership through life, wheels, and math. You have shown me, by your own resilient example, how to meet adversity with laughter instead of tears: your unfailing ability to do this will always inspire me. Words cannot express how thankful I am to have you in my life, and to have undertaken this journey with you. Here is to a whole lifetime of adventures together, both inside and outside of the academia.

# Chapter 1

# Introduction

The classical paradigm of statistical inference [1] and its contemporary manifestation, machine learning (ML) [2], are seeing a historic resurgence and tremendous empirical success in tasks of supervised prediction and unsupervised pattern recognition. This success is largely owed to the combination of highly expressive models [3], Big Data, and GPU compute [4], and has been replicated in a number of application domains. These domains includes computer vision [5] and natural language processing [6], in which deep neural networks today achieve "state-of-the-art" empirical performance. Some of this success has also been replicated in the more difficult task of reinforcement learning [7], which consists of learning optimal control policies in a stochastic environment with unknown dynamics.

This progress has been largely driven through single-agent perspectives. Underlying single-agent learning is the classical statistical assumption [1, 2] that the learner has access to a batch of data, that has been generated by a *natural process*. Stated alternatively, the learner is assumed to interact with a fixed stochastic environment, the parameters of which are being learned about.

However, ML is increasingly being discussed in the context of settings where this assumption is not viable. For example, the principles of ML are currently being applied to mechanism design in online marketplaces. In these applications, the data that is being learned from is at least partially impacted by the actions of a selfish agent who, herself, possesses unknown incentives. In the AI milieu, agents can no longer be assumed to act in isolation—they will increasingly be forced to interact *with each other* in addition to the fixed environment. As we motivate in Sections 1.1 and 1.2, the success of both current and future engineering applications critically involves, by their very nature, engaging with core questions in this kind of *multi-agent learning*.

On the other hand, interactions between multiple agents of a *known strategic nature* have been modeled through the classical framework of game theory. The applications of game theory and mechanism design are diverse both in engineering (e.g. path routing through networks, power grids, and dynamic spectrum sharing) and e-commerce (e.g. auctions, matching markets, ride-sharing, hiring platforms). However, in the traditional, best understood framework of game theory, the strategic nature of all agents involved is *known* to be selfish and

rational with all utility functions as common knowledge; therefore, none of the agents are engaging in learning. Of course, the assumption of rationality has seen significant push-back from the behavioral economics community, and a wide range of models for human strategic behavior have been theoretically postulated as well as empirically evaluated on human data [8–11]. However, the time-scale of this theoretical postulation and empirical evaluation is on the matter of months or even years. As we motivate in Sections 1.1 and 1.2, automated multi-agent interaction allows for the same wide range of spectrum of strategic behavior, usually unknown a-priori; moreover, decisions now need to be made *real-time*. To optimally make decisions in this unknown environment, the necessary process of learning of strategic nature is ideally integrated with real-time interactions.

The above perspectives necessitate a fundamental intersection of our understanding of single-agent learning with our understanding of multi-agent game theory. Sections 1.3 reviews our existing understanding of these classical areas of research and sets up the mathematical framework for this dissertation. At the same time, this section highlights the inherent non-trivialities involved in intersecting learning and games. As we describe in Section 1.4, the core problems can be split into two broad categories:

1. Can we understand how to optimally learn the nature of an unknown environment, and perform decision-making that is almost as optimal as though we had known the nature of this environment beforehand?

2. Can we understand (approximations) of the optimal behavior of a strategic agent in the presence of a learner?

Before describing the formal framework for this thesis, and setting up the above core problems, we ground the reader in two real-world case studies in which questions at the intersection of learning and strategic behavior are increasingly paramount. We start by considering the design of automated, online marketplaces.

## 1.1 Case study: Learning in online marketplaces

As paraphrased by Paul Milgrom in his classic text "Putting Auction Theory To Work" [12], game theorists in the $20^{th}$ century "plied their trade" on two important application domains: large-scale auctions in the public sector, and matching markets. Quote from Milgrom [12, page 2, Chapter 1]:

*"An article in 1995 in the New York Times hailed one of the first US spectrum auctions as 'The Greatest Auction Ever'. The British spectrum auction of 2000, which raised about* \$34 *billion, earned one of its academic designers a commendation from the Queen and the title 'Commander of the British Empire'...The National Resident Matching Program, by which* $20,000$ *US physicians are matched annually to hospital residency programs, implemented a*

*new design in 1998 with the help of the economist-game theorist Alvin Roth.* ***By the mid-nineties, thirty-five years of theoretical economic research about fine details of market design was suddenly bearing very practical fruit.***"

These fruits of this success involve human agents and the design of simple (e.g. the celebrated Vickrey-Clark-Groves mechanism [13–15]), robust (e.g. the property of obvious-strategy-proof-ness [16]) and computationally efficient (e.g. SAT solvers in the spectrum incentive auction [17, 18]) mechanisms that are adjusted as per numerous discussions between academics and government policy experts (for a representative report on some of these discussions, see [19]). Today, we are witness to a new avatar of *automated-agent* participation in mechanism design, together with dynamic mechanisms that are designed real-time. Prominent examples of this include Google's Ad-Words auctions and online matching implemented in applications like hiring and ride-sharing platforms. The design of these mechanisms will now be done with tremendous amounts of incomplete information, both from the perspective of the mechanism designer and the participating agents. In what follows, we highlight the numerous sources of incomplete information that arise, and the ensuing non-trivialities that ensue in optimal market design. We focus on auction design as a representative example for the sake of brevity, while noting that similar questions arise in other online platforms such as matching markets [20, 21].

## Representative example: Modern, private-sector auction design

The classic problem of (approximately) optimal auction design directly engages with strategic agents in a very obvious manner. Large-scale auction design has traditionally been restricted to the public sector, where the goal is to maximize *social welfare* of all the agents. However, one of the most prominent recent examples of large-scale auction design, advertising auctions, is carried out in the private sector. The mechanism designers here are companies like Google, Microsoft and Yahoo!; consequently, the aim is frequently to maximize *seller revenue*. Hence, we focus on the objective of revenue maximization.

We consider one of the simplest and oldest models for auctions, the *single-item auction*[1] with symmetric bidders [23]. Here, as depicted in the block diagram in Figure 1.1, the (three) bidders have valuations $(v_1, v_2, v_3)$ for the single item drawn from a common probability distribution $F(\cdot)$. A seller's chosen mechanism maps submitted bids $(b_1, b_2, b_3)$ to an allocation of the item (to one of the bidders), and corresponding payment rules that all of the bidders are bound to follow. Importantly, the bids are being submitted by strategic bidders, and so the bid of agent $i$, i.e. $b_i$, need not be equal to her valuation, i.e. $v_i$. Thus, the goal of seller mechanism design is to maximize her obtained revenue, and indirectly, to *elicit* the bidders' valuation information. One way of doing this is by designing a mechanism that incentivizes the bidders to report their true values in the ensuing Bayes-Nash equilibrium; such mechanisms are called *truthful*, or *incentive-compatible* mechanisms.

[1] Multiple items can also be sold, leading to the challenging problem of combinatorial auction design. For a representative discussion on the computational difficulties involved, see [22, Chapter 11].

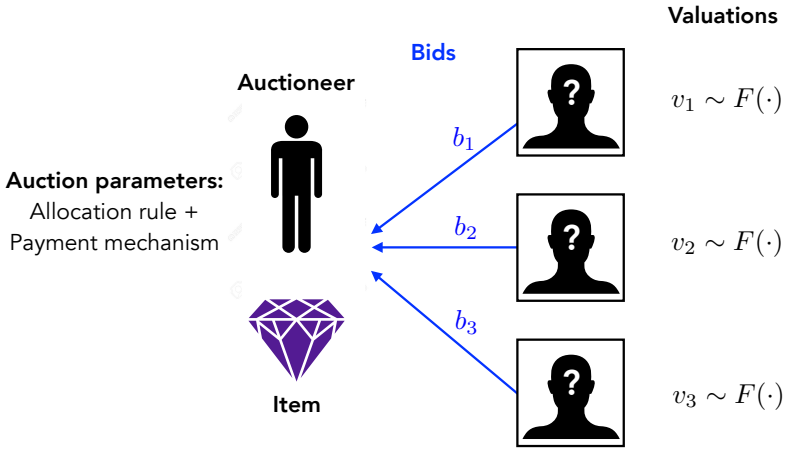



Figure 1.1: Block diagram of a simple single-item auction with 3 bidders whose valuations $(v_1, v_2, v_3)$ are drawn from the same probability distribution $F(\cdot)$. In general, the bids $(b_1, b_2, b_3)$ need not be equal to the valuations $(v_1, v_2, v_3)$. The auction design is parameterized by an allocation rule, that decides who receives the item, and a payment rule, that decides how much bidders pay. These rules are functions only of $(b_1, b_2, b_3)$ and possibly use knowledge of the valuation distribution $F(\cdot)$, but never the valuations themselves. This figure was made using Keynote.

Myerson's seminal theory [23] for this "one-shot", single-item auction design postulates that a particular type of reserve price implemented together with a Vickrey (second-price) auction is revenue-optimal among all mechanisms, truthful or otherwise. Here, the reserve price is simply a price for the item posted by the seller: if none of the bidders meet this price, the seller will not sell the item at all. Importantly, the optimal reserve price *intricately depends on knowledge of the valuation distribution* $F(\cdot)$. This is an extremely idealized assumption, as in practice the distributions of preferences among bidders can widely vary and are unknown to the seller. Below, we motivate that practically effective auction design requires a non-trivial component of *learning* about the bidders in a number of ways[2].

1. In modern applications of auction design like online advertising, the bidders' valuation distributions for item(s) are typically unknown. Regardless of the statistical model for the unknown valuation distributions, it is clear that for effective Myersonian auction design, the optimal reserve prices need to be estimated from past data.

[2]Judging from academic talks given by researchers in industry [24], the issues that arise in practical private-sector auction design are even more numerous, and several are as yet unformulated mathematically. Thus, the list provided above is far from comprehensive, but provides a good starting point for motivation.

2. It has been observed in practical deployments of auctions that bidders consider first-price, or "winner-pays-bid" auctions as maximally trustworthy. Such auctions are no longer truth-telling, and their analysis deviates from the classical theory (for a survey of this recent research, see Jason Hartline's recent tutorial [25]). Nevertheless, these non-truthful mechanisms are widely used in practice and can also enjoy improved guarantees on welfare in *prior-free mechanism design*, i.e. design with robust guarantees for any valuation distribution [26]. Such guarantees intricately involve optimizing the reserve prices used in the auction design from past samples of data. Extensive field experiments on advertising auctions [27] have shown that these optimizations on reserve prices, while not publicly available, have a material effect on increasing revenue.

Invariably, while we do not have access to the details of these auctions, we know that at a high level the auction parameters are optimized as a function of data obtained from previously run auctions. Moreover, the recent paradigm of dynamic[3] mechanism design (e.g. [39, 42, 43, 47]), in which these optimizations are done real-time and online, is increasingly realistic to model modern market design with automated agents. Thus, learning, in some form or the other, plays a significant role in the design, as does modeling the strategic behavior of bidders who interact with the mechanism.

Efforts to effectively integrate learning into modern market design have been recent, promising, and numerous. A complete survey of all of the literature in private-sector auction design is beyond the scope of this thesis; however, several key questions remain to be answered that connect directly to the fundamental issues discussed here. In general, the problem of learning optimal auctions from provided bidder data is highly non-trivial as bidders are heterogeneous in their preferences — moreover, they could possess sizable incentives for complex dynamic strategic behavior as a consequence of long-term interaction with the mechanism. Nor has this problem been ignored by the research community in economics and computation, as detailed below:

1. Even the problem of learning optimal auctions from heterogeneous, but myopic bidder data involves important modeling questions. The best case is homogeneity, in which case a statistical learning theory perspective can be used to provably learn approximations to the Myerson-optimal auction from samples of valuations [31–38]. This theory is valid if samples of valuations can be reliably extracted. This might happen if, for e.g., bidders are myopic and have little incentive to perform strategic manipulation. On the other hand, the worst case of heterogeneity in bidder preferences is when valuation distributions can vary arbitrarily from round-to-round; in this scenario, ideas from worst-case online learning can be applied to obtain guarantees on revenue, as in [39].

[3]We here use the word "dynamic" to indicate the presence of *online learning*; dynamic mechanism design is also commonly used to describe online algorithmic mechanism design which does not have a learning component, but is aimed to maximize revenue well over the single-shot setting when the valuation distribution is known, but realizations are not [28–30]. We do not discuss this work here as it does not directly engage with issues of learning.

2. The possibility of dynamic strategic behavior poses an even more challenging problem: in the above single-item auction setting, we need to learn the valuation distribution $F(\cdot)$ from samples $(b_1, \ldots, b_n)$ for any hope of optimal auction design. Is it possible to do this when the samples are themselves being generated by strategic bidders? It turns out that when the seller and bidder are equally patient, the worst-case kind of dynamic strategic behavior could cancel out any potential benefit of learning [40, 41]. However, when bidders are less patient, learning-based schemes, primarily based on preserving truthfulness, can do almost as well as the Myerson auction in hindsight [42, 43]. Moreover, even when they are incentivized to, bidders may not necessarily engage in dynamic strategic manipulations of their bids as the optimal manipulation could be extremely complicated, and not analytically evaluate-able or computationally tractable.

3. Most practical auctions are *non-truthful*. Learning such auctions will likely pose even deeper challenges at the intersection of learning and strategic behavior. This is because even in the purely statistical setting, samples are of bids which are strategic manipulations on true values. Nevertheless, under certain settings, inference of these auctions' revenue and welfare guarantees is possible using systematic properties of the agents' bids as a function of their values in Bayes-Nash equilibrium [44–46].

The above discussion shows that modeling the strategic behavior of bidders is an extremely complicated task, as is designing learning algorithms for auction design that respect all possible kinds of strategic behavior. We first note that the goal of a "all-purpose" learning algorithm that adapts to the behavioral model that best describes bidder behavior, while sometimes utopian, is a worthy methodological goal that is directly connected to this dissertation's goal of learning from an unknown environment. In fact, the goal of adaptivity in auctions has already seen some research attention: in particular, repeated auctions can be shown to be robust to several such "types" of strategic behavior in a setting where the valuation distributions are known [47] — here, the only component of the bidders that is being learned is their (discrete) strategic type. The situation where both strategic behavior and valuation distributions are unknown is more challenging and engages with the heart of issues of learning intersected with strategic behavior. The methodology for adaptivity in Chapters 2 and 3, while developed in simple learning-theoretic models, has particular promise for auction design given the recent advances in statistical as well as worst-case learning of optimal auctions. These kinds of interactions also demonstrate the need to better model strategic behavior of bidders in the presence of learning. Simple rules that approximate dynamic strategic behavior, if and when they exist, are therefore of sizable interest to realistic market design. From the above discussion, we conclude that any treatment of dynamic mechanism design with a learning component needs to engage with the twin issues of: a) methodology for online learning in dynamic, potentially strategic environments; b) intellectual tools to understand manifestations of bidder strategic responses to seller mechanisms designed with a learning component.

As a final comment, we remark that these issues are not completely foreign to traditional deployments of auction theory with social-welfare maximization and/or human bidders. Human behavior *is* unpredictable, and auctions need to be both simply designed and robust to deviations from the expected/modeled behavior. However, the scope of the modeling problem is significantly more acute in dynamic mechanism design with automated agents: the uncertainty in information is greater, and errors in modeling lead to consequences that are both real-time and accumulate over time.

## 1.2 Case study: Game theory in spectrum regulation

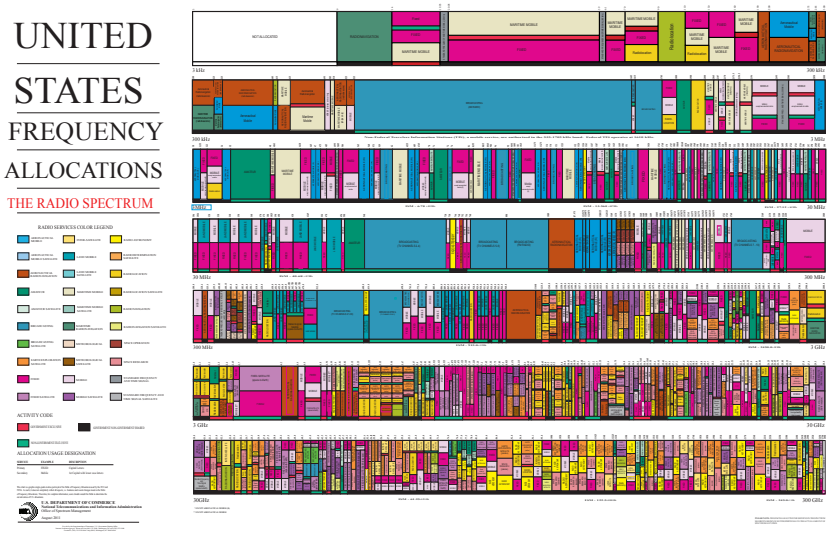



Figure 1.2: Depiction of spectrum allocation according to the command-and-control paradigm in the United States. Figure credit to U.S. Department of Commerce, National Telecommunications and Information Administration, Office of Spectrum Management.

Our second case study comprises the landscape of intelligent cognitive radio technology [48] and its role in enabling dynamic spectrum sharing [49]. Wireless spectrum is a valuable resource with tremendous economic opportunity [50]; however, its supply in recent times has not kept up with demand. Rather than a lack of availability of the raw resource, the reason for this is primarily inefficiencies in allocation of the resource across both space and time. Figure 1.2 illustrates the traditional *command-and-control* approach to spectrum regulation: a single user gets a license for dedicated access to the spectrum band, and (typically) does not need to share this band. Historically, command-and-control cleanly ensured spectrum usage rights for license holders, and did not preclude or restrict entrants when spectrum was plentiful. However, this ability breaks down today in the face of high demand and spectrum scarcity — it does not allow for robust and flexible sharing of the spectrum

band across either space or time. This inefficiency, while more recent, was predicted by the prescient analysis of the economist Ronald Coase more than 50 years ago [51].

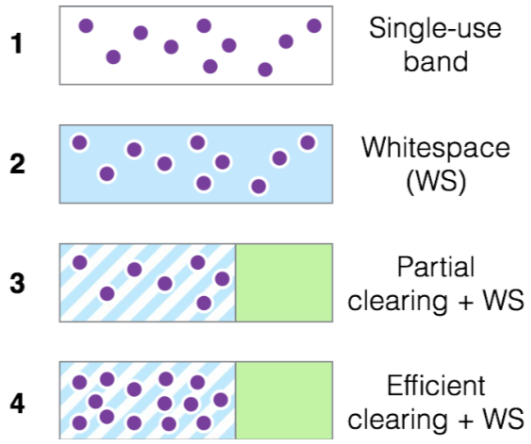



Figure 1.3: An illustration of the various options for spectrum re-purposing. Incumbents are shown as purple dots, while white-spaces are blue and cleared spectrum is green. White represents unused spectrum, and the white and blue pattern represents spectrum that could, but need not, support white-space rules. Figure from [52].

Over the past few decades, ambitious market-driven approaches to re-purpose and reallocate wireless spectrum, primarily through auctions [19, 53, 54], have substantially improved *spatial* efficiency of spectrum usage. For example, several of the higher television broadcasting channels are seldom used by most locations in the United States, resulting in a large amount of "TV white-space" that is widely utilized by secondary users of spectrum (for a summary of opportunities afforded by the presence of TV white-spaces, see Kate Harrison's pioneering thesis [55]). In fact, the FCC recently conducted an incentive auction that successfully completely clears these TV bands for long-term-evolution (LTE) usage while efficiently "repacking" the few TV stations, that were originally present in these bands, into lower bands. See Figure 1.3 for a schematic of the eventual outcome of this "efficient clearing" method. While market-driven efforts to re-purpose spectrum have already paid dividends in efficiency of spectrum usage, they can only go so far. This is because they do not address the problem of *dynamic, temporal spectrum sharing*. For example, spectrum is often still owned by a primary user who uses it only 5% of the time — the rest of the 95% of the time, this valuable resource is wasted!

The central question whether we can ensure the protection of primary users' rights, while also allowing the unused spectrum to be better utilized. Being able to do this necessitates modeling dynamic spectrum sharing through real-time[4], game-theoretic interaction between

[4]Another significant option for temporal spectrum sharing that we do not discuss in detail here is the framework of pricing competition games [54, 56] between multiple primary users that "lease out" spectrum

multiple rational, intelligent users of the spectrum. The justification for considering secondary users of spectrum as rational and intelligent comes from the remarkable advent of cognitive radio technology in the last 20 years [48]. This technology allows for the design of radios[5] to intelligently sense channels for spectrum availability, as well as adapt to the environment in an agile manner [57–62]. In the absence of any external mechanism, such intelligent radios that are symmetric (in the sense that they cause equal amounts of interference harm to each other) can self-enforce[6] against one another and share the spectrum in an equitable manner; but in scenarios of unequal harm, the less vulnerable device dominates [63].

In particular, if the nature of harm is unidirectional (e.g. if packets from two users collide, the packet is dropped for only one of the users), we need to create external incentives in the form of explicit enforcement for equitable use of the resources. Kristen Ann Woyach, in her PhD thesis [64], laid out a pioneering framework of *light-handed regulation*, implementable in real-time [65], and inspired by economic analyses of criminal law [66–68] to model desired interaction between asymmetric users of the spectrum. In this model, the users are asymmetric in two senses: their demands of the resource are different, but so are their modes of access and rights of usage. On one hand, our goal from the point of view of a *primary user* (who has typically paid for the spectrum) is to ensure as little secondary interference to her overall activity as possible. On the other hand, our goal from the point of view of a *secondary user* (who is typically sensing for the spectrum for free) is to ensure a "best-efforts" guarantee to a compliant and technically desirable secondary. We will now briefly describe the mechanism by which these goals are theoretically achievable.

## A mechanism for light-handed regulation

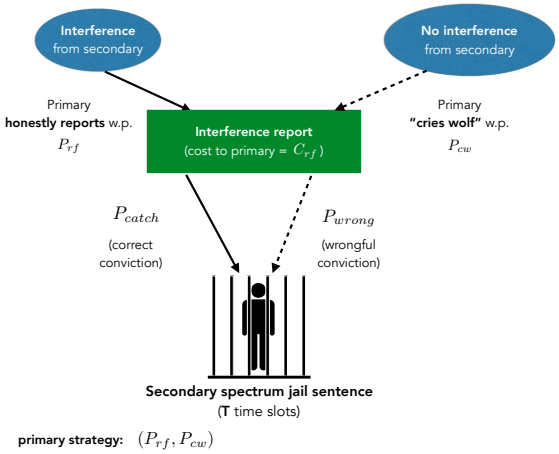
The important paradigm shift advocated by Woyach in her thesis constitutes spectrum regulation through light-handed, *ex-post* enforcement. At a high level, such enforcement allows secondary users to use spectrum in an unrestricted manner, but subjects them to "punishment" if they are "caught" in the act of interference to a primary user. More concretely, this light-handed mode of enforcement works through a third-party enforcer in conjunction with the participation of the primary user herself:

1. The mode of "punishment" for a secondary user, if caught, is a *spectrum jail sentence for radios*: the secondary is forced to turn off his transmissions for a stipulated period of time denoted by $T$. (This stipulated period of time is known-in-advance to the secondary, but only enforced if the secondary is actually caught interfering.)

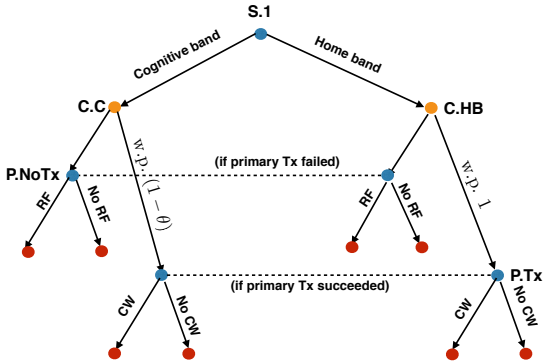
when it is not being used to potential secondary users. While these models are interesting and have been partially put into practice, they are only successful for spectrum sharing on the time scales of weeks or months, owing to the monetary nature of the transactions. Here, we are most interested in the question of *real-time* spectrum sharing amongst many users.

[5]In fact, in a remarkable leap from theory to practice, much of this technology is now software-implementable.

[6]This self-enforcement also underlies the practical success of technologies like WiFi!




(a) Flowchart showing the effect of spectrum jail enforcement on a secondary user.




(b) Extensive form of the ensuing primary-secondary game. The primary strategy is to report or not report interference at every step, and the secondary strategy is to transmit in the spectrum band or another (home) band.

Figure 1.4: The spectrum jail enforcement system as created by Woyach [64] and the ensuing primary-secondary game. Both figures were made using Keynote.

2. When a secondary user causes interference, the primary has the option (at a cost $C_{\mathsf{rf}}$) of reporting this interference to the third-party enforcer. With a positive probability, the interference is correctly attributed to the secondary user and the secondary user is jailed.

Figure 1.4a contains a schematic of this mode of enforcement, and Figure 1.4b provides a summary of the ensuing primary-secondary game in normal form. While we refer the interested reader to [69] for the details of the analysis, we provide here a brief summary of the guarantees afforded by this enforcement mechanism. Intuitively, it is clear that a sufficiently large jail sentence $T$ would deter a strategic secondary from over-transmitting, thus guaranteeing a primary user a minimal level of protection. Muthukumar and Sahai [69] provide concrete lower bounds on this jail sentence $T \geq T_0(\gamma)$, that also rely on estimates of minimal primary demand and the fidelity of the enforcement mechanism ($P_{\mathsf{catch}}$), to guarantee that the primary is protected a fraction of $\gamma$ of the time. ($\gamma$ can approach arbitrarily close to 1, and the sentence $T_0(\gamma)$ will increase accordingly.) Notably, the calculation of the minimal jail sentence depends critically on the nature of strategic behavior by secondaries: *selfish and rational secondaries* tend to require smaller jail sentences to be *deterred*, while *malicious/adversarial secondaries* need larger jail sentences to be *shielded from*. Alternatively stated, the same choice of enforcement mechanism would afford greater levels of protection against selfish and rational secondaries than malicious secondaries!

It is intuitively clear that this light-handed enforcement mechanism can ensure adequate levels of primary protection — but we want to go further. We want to show that technically desirable secondaries are incentivized to actually utilize spectrum holes. The original work

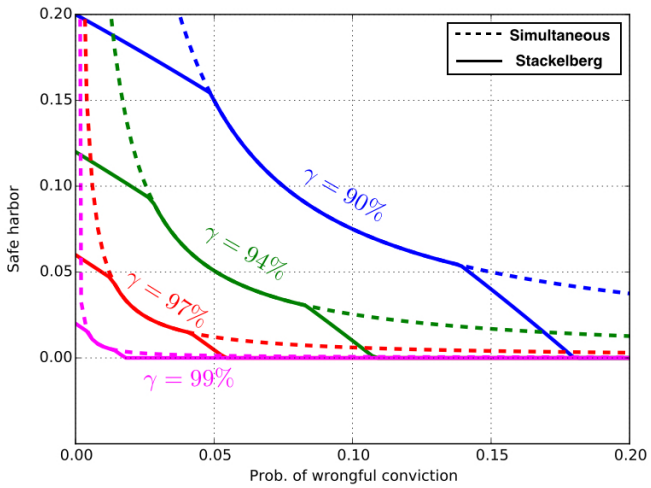



Figure 1.5: Figure showing how the size of the "safe harbor" of technically desirable secondaries changes for various primary performance requirements, given by $\gamma$. Figure from [69], and details of how parameters are set there-in.

of Woyach [64], in accordance with the criminal law literature, framed such a "desirable" secondary as one who would honestly use spectrum, i.e. cease to transmit if a primary transmission was detected. However, even honest secondaries can cause substantial harm to the primary if they frequently incorrectly sense that spectrum is available, even when it is not. Muthukumar and Sahai [69] critically re-framed the notion of a desirable secondary as one that will cause sufficiently little interference harm on average, whether from cheating or honest usage of the spectrum in conjunction with the primary. Using this re-framing, their main result shows that the (Stackelberg as well as simultaneous) equilibrium of the primary-secondary game can result in a sufficiently desirable secondary co-existing in the same spectrum band with the primary, *while causing minimal interference harm to the primary.*

For the class of such desirable secondaries, designated as a "safe harbor", efficient *and* equitable spectrum sharing is possible with the primary. The size of this safe harbor as a function of wrongful conviction rates, for various levels of stipulated primary protection, is depicted in Figure 1.5.

## The need for engagement with learning

From the above, the benefits of a light-handed mechanism are demonstrated to be two-fold: not only is equitable *and* efficient spectrum sharing possible in equilibrium, but also this

mechanism incentivizes the development of desirable secondary technology. While these equilibrium guarantees are a significant and promising step towards equitable and efficient spectrum sharing, they are under-lied by several assumptions embedded in the idealized framework of "one-shot" game-theoretic interaction between primary and secondary user(s); particularly that the primary and secondary user know important parameters about each other. In the absence of this critical assumption, the primary and secondary would conceivably reach this desirable equilibrium through some form of *learning* each other's information—but how they might do this is quite unclear. We give a sense of the scope of these informational assumptions below.

1. The primary acts with knowledge of important secondary parameters, such as the probability that he can cause her interference when they use the spectrum band at the same time.

2. The secondary acts with knowledge of important primary parameters, such as her inherent demand for spectrum.

3. The primary and secondary both act with knowledge of parameters of the enforcement mechanism, such as the cost of reporting interference, and parameters of the environment.

4. The primary assumes knowledge of the strategic nature of the secondary — whether malicious or selfish/rational.

In a dynamic spectrum sharing environment, it is not clear that we can make any of the above assumptions a-priori. After all, the primary and secondary user are neither co-designed nor cooperative; therefore, they will not know anything about each other a-priori, nor can they necessarily rely[7] on a trusted exchange of information such as their operating parameters. With this in mind, communication technology is increasingly designed with a ML component, see e.g. [70]. The objective of such technology is to successfully learn about the environment (as a conglomerate of nature and the aggregate effect of primary/secondary users), and transmit optimally. A more subtle point is that the primary has to engage in learning of her own — about the secondary's strategic nature. This is important because, as we saw, a primary should interact differently with malicious and rational secondary users, and the presence of the latter ensures relatively favorable spectrum utility. Other categories of secondaries that were not considered above are *agnostic secondaries*, that transmit with a fixed probability invariant to primary actions and enforcement mechanism; and *cooperative secondaries*, which include primary throughput and overall spectral efficiency as a part of their own utility function.

The scenarios of agnostic and cooperative users are expected to be far more benign for a primary user, and if successfully detected, can lead to greatly improved spectral efficiency.

[7] In particular, designing private protocols for doing this could involve inference about the agents' strategic types as well.

In fact, DARPA's second version of the Spectrum Collaboration Challenge[8] aimed to bring separately designed teams of cognitive radios together to "learn to cooperate" to maximally utilize spectrum in both a fair and efficient manner. The elephant in the room is that "learning to cooperate" in the absence of co-design requires the ability to detect an intent of cooperation, and maintain robust/optimal performance in the absence of this intent. Moreover, the practical importance of adaptive regulation in the presence of artificially intelligent, learning agents beyond the domain of spectrum cannot be over-emphasized. The nascent field of algorithmic fairness in ML begins to address a future ecosystem in which stakeholders and regulators, by default, interact with learning agents real-time. A foundational understanding of game-theoretic considerations in environments with learning, whether competitive or cooperative, will be essential to effective regulatory design in the modern AI age.

## 1.3 Review: Learning in *known* environments

The two case studies we considered have amply demonstrated that a fundamental understanding of learning in game-theoretic environments is required for further intellectual and engineering progress. This thesis explores this fundamental question using the simplest possible models that retain non-triviality. We will now review the important mathematical tools and framework that we will use for this dissertation.

The semantics, mathematics and automation of "learning" as an abstract concept have extremely rich history in statistics [1, 2], information theory [71], control theory [72] and artificial intelligence [7, 73]. In this thesis, "learning" is used as an umbrella term to describe the merging of statistical inference from past data with optimal decision-making in a variety of *known* environments. We broadly categorize these environments into *statistical* and *strategic* (in the non-cooperative sense) below[9].

We will see that the conceptual philosophy and technical tools used are both quite different for statistical and strategic environments. Throughout the discussion, we highlight the history of research that attempts to bridge these environments, and evidence for the depth of the problems formulated in this thesis.

### Statistical learning

We denote a batch of data by $\{Z_i \in \mathcal{Z}\}_{i=1}^n$ and a sequence of data by $\{Z_t \in \mathcal{Z}\}_{t\geq 1}$. The statistical learning paradigm assumes that underlying the composition of this data is an unknown component that needs to be inferred (this could be a parameter or a function from a non-parametric class with special geometric structure [74]), together with stochastic noise that is independent of the unknown component[10]. This model is notably broad; it can include non-stationarity in the data generation process, and dependencies across the data

[8] For details, see `https://www.darpa.mil/program/spectrum-collaboration-challenge`

[9] We discuss cooperative considerations as future work in Chapter 6.

[10] Typically, the presence of stochastic noise is what makes the learning problem non-trivial.

points. Correspondingly, the nature of the goals of statistical learning is also broad, ranging from prediction to data re-generation to optimal decision-making. We now describe the setting of supervised prediction from batch data as a common sub-paradigm of statistical learning. Chapters 2 and 3 will also engage with the online variant of this problem (in statistical as well as adversarial environments).

**Example 1.** *In supervised statistical learning from batch data, the data is given by* $\{Z_i = (X_i, Y_i)\}_{i=1}^n$ *where the features are given by* $X_i$ *iid* $\sim P$*, and the output is an unknown function of the features, i.e.*

$$Y_i = f^*(X_i) + W_i \tag{1.1}$$

*where* $f^* \in \mathcal{F}$*, and* $W_i$ *is iid stochastic noise. Typically, the function class* $\mathcal{F}$ *is either parametric, meaning that* $\mathcal{F} := \Theta$ *and* $f^*(\cdot) := f_{\theta^*}(\cdot)$*, or non-parametric; for example,* $\mathcal{F}$ *can represent the space of convex functions or smooth functions or a reproducing kernel Hilbert space [75, Chapter 12].*

*The primary goal in supervised statistical learning is to provide an* ***estimate*** *of the unknown function, denoted by* $\widehat{f}$*, and use this estimate to obtain accurate predictions on a fresh sample of data. In other words, the goal is to ensure that for a new sample* $(X, Y)$ *generated in the same way as above, we can get*

$$\ell(Y, \widehat{f}(X)) \tag{1.2}$$

*to be as small as possible, where* $\ell(\cdot, \cdot)$ *is an appropriately defined non-negative-valued loss function. Commonly studied loss functions include the squared loss function and the* $0-1$ *loss function.*

The statistical learning paradigm is remarkably broad, and remains a very active area of research today. What is essential to this paradigm is the existence of a *natural process*, i.e. a distribution $P \in \mathbb{P}(\cdot)$ that underlies the generation of data *that does not respond, or change in a time-varying manner, to the learner/decision maker's actions*. Also note that we have de-facto assumed the frequentist formulation in the above informal discussion, as in [1, Chapter 3] and forthcoming formulations. All of the setups in statistical learning are also well-formulated and studied under Bayesian models; indeed, Bayesian is also heavily used in economic decision (expected-utility) theory as well as game theory with incomplete information [76–78]. For most of this thesis, we will use the frequentist statistical paradigm by default. In other words, the unknown function $f^* \in \mathcal{F}$ can be arbitrary, and the goal of frequentist statistical estimation is to design estimators $\widehat{f}_n := \phi(\{Z_i\}_{i=1}^n)$ with one, or both of the following properties:

1. Asymptotic *consistency*, i.e.

$$\widehat{f}_n(\cdot) \to f^*(\cdot) \text{ almost surely as } n \to \infty \tag{1.3}$$

2. Non-asymptotic *minimax optimality*, i.e. for any value of $n \geq n_0$, we have

$$\sup_{f^* \in \mathcal{F}} \mathbb{E}\left[\ell(\widehat{f}_n(X), Y)\right] \sim \inf_{\phi(\cdot)} \sup_{f^* \in \mathcal{F}} \mathbb{E}\left[\ell(\phi(\{Z_i\}_{i=1}^n)(X), Y)\right], \tag{1.4}$$

where the $\sim$ indicates *order*-optimality in terms of dependencies on $n$ as well as relevant complexity measures of the function space $\mathcal{F}$.

The above definitions are the gold standard for optimality in frequentist theory of statistical estimation[11]. On the other hand, the Bayesian theory [1, Chapter 7] defines a prior distribution $\mathbb{P}(\mathcal{F})$ on the function $F$, i.e. the generative model for data is $Y = F(X) + W$. Then, *with knowledge of this prior distribution*, the maximum-a-posteriori (MAP) estimate under the above framework would be:

$$\widehat{f}_{n,\mathsf{Bayes.}} := \max_{f \in \mathcal{F}} \mathbb{P}\left(F = f \middle| (Z_i)_{i=1}^n\right), \tag{1.5}$$

and this is easily shown to be the quantity that minimizes the Bayes risk, i.e. $\mathbb{E}\left[\ell(f(X), Y)\right]$.

Applied statistical methodology, which is not our focus here, uses both frequentist and Bayesian principles. From a more conceptual standpoint, we can see from the above definitions that notions of optimality are harder to formalize in a frequentist setup than a Bayesian setup; on the other hand, the MAP estimate could be complicated in its analytical expression, and difficult to interpret. The frequentist paradigm is considered to be more robust than the Bayesian paradigm, which is most effective when the prior distribution is well-specified; and allows for much broader classes of estimators that are often under-lied by simple principles that highlight fundamental trade-offs involved in statistical estimation or decision-making. In game-theoretic settings with incomplete information, most statistical modeling has been Bayesian owing to the simplicity in defining equilibrium concepts [76, Chapters 4 and 7]; however, in Chapter 4 of this thesis, we make a case for introducing frequentism to the realm of game theory.

Regardless of the modeling choice of frequentism or Bayesian-ness, the fact remains that learning algorithms designed with such a statistical assumption in mind, while optimal for statistical environments, are dramatically brittle if the statistical assumption is *violated*. Putting this in the context of both of our case studies, this has important practical ramifications:

1. An auction mechanism that is designed for learning from statistically generated bidder data will see a significant drop in revenue in the presence of long-term strategic bidders who aim to game the mechanism for their own benefit.

2. A user of spectrum that senses the environment as though it is static will see a dramatic drop in throughput from the presence of malicious or competitive users of the spectrum.

[11] Distributional notions of convergence, like asymptotic normality, are also important when considering statistical inference, but our focus here is on estimation error guarantees.

## Sequential learning in non-cooperative, game-theoretic environments

The discussion above shows that we need to consider the possibility of the data $\{Z_t\}_{t\geq 1}$ being generated with a strategic incentive in mind. The way we will do this is by modeling the interaction as a repeated game between learner, whom we designate as "Alice", and the generator of data, whom we designate as "Bob". At round $t$, Alice's strategy is precisely the action that she chooses to take, $A_t$; and Bob's strategy is the sample of data $Z_t$. As in the online decision-making setup, Alice will choose a (possibly randomized) strategy $\{A_t\}_{t=1}^T$ with the aim of (approximately) maximizing her reward function,

$$R := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(Z_t; A_t)\right].$$

The principal difference is in how Bob will generate his data. Instead of $Z_t$ being independent and identically distributed across $t$ (or even evolving according to a stationary stochastic process), Bob now has a reward function of his own, denoted by $g$, and will be generating samples $\{Z_t\}_{t=1}^T$ with the aim of (approximately) maximizing this reward function,

$$G := \mathbb{E}\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} g(Z_t; A_t)\right]. \tag{1.6}$$

Furthermore, we will assume what is commonly called the *full-information* feedback structure[12]: Alice can observe not only her reward $r(Z_t; A_t)$ at every round, but also the data $Z_t$. Similarly, Bob can observe Alice's action $A_t$ at every round.

We can now try and ask the same question as before, i.e. what is Alice's (approximately) optimal strategy against Bob in this repeated-game. At a high level, it is easy to see that the nature of this strategy intricately depends on the strategic nature of Bob, i.e. the nature of his utility function. Ironically, the case that is easiest to study involves Bob being adversarial to Alice. In other words, under the adversarial model, Bob observes Alice's action $A_t$ at round $t$ and generates data $\{Z_t\}_{t=1}^T$ to minimize Alice's reward, i.e. $g(\cdot,\cdot) = -r(\cdot,\cdot)$. In other words, this constitutes a *repeated, zero-sum game*.

Interestingly, most known guarantees for optimality for Alice are framed in the form of a quantity called *worst-case regret* [80, 81], with respect to the best single action that Alice could have used in hindsight. The idea of regret goes back to Jim Hannan [80] and has roots in the seminal work by David Blackwell on approachability of convex sets over time [82]. While it is most applicable to adversarial environments, it has also been posited

[12]Even less is known about the limited-information feedback in realistic game-theoretic settings: it is a nascent but active area of research today. While the adversarial bandits [79] paradigm is popular in the online learning literature, it does not realistically model even zero-sum game theoretic interaction, as the adversary in the adversarial bandits model essentially has full-information access to the observed rewards.

for use in *any* strategic environment by noting that the worst-case kind of unknown strategic behavior is, in fact, adversarial. Constructive algorithms that satisfy this no-regret property have seen a resurgence in the online learning literature in both the computer science [83, 84] and economics [85] literature. These algorithms are largely interesting in the context of zero-sum games because when played against each other, the time averaged payoff of both players converges to the unique Nash equilibrium payoff, or value, of the zero-sum game. Thus, no-regret algorithms are viewed as a way to approach Nash equilibrium in zero-sum games. Even with all of this promise, it is important here to note that the very metric of regret does not, actually, tell the full story: Chapters 2 and 3 demonstrate the importance of selecting an offline benchmark in online learning and demonstrate that minimizing regret is not always synonymous with maximizing reward. Moreover, Chapter 5 shows that no-regret algorithms, when played against one another, can exhibit surprising day-to-day behavior even in the simplest $2 \times 2$ zero-sum games.

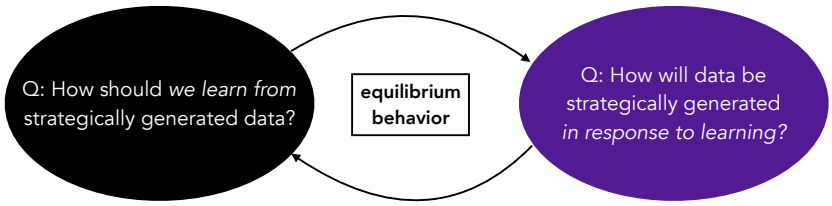



Figure 1.6: Illustration of infinite-loop reasoning that unavoidably arises in repeated game-theoretic interaction with learning. Instead of formalizing approximate optimality for agents in isolation, we need to formalize approximate game-theoretic equilibria. This figure was made using Keynote.

Even if we do assume the full validity of the metric of regret for the moment, we also note that the above *frequentist* perspectives are rooted in optimality from the perspective of adversarial regret; and in fact the process of minimizing regret against an adversary is really a strategy of defense, rather than a learning algorithm. In other words, Alice already knows that Bob will want to hurt her as much as possible, and is playing to avoid such hurt in the long run. Our understanding of how Alice should play against a *competitive* Bob, especially when his utility function is unknown to Alice, is much less mature, and the questions become far more nuanced. It is worth noting that the game between Alice and Bob is now *non-zero-sum*, and here the most basic solution concept of game theory, the Nash equilibria (NE), remains elusive in several basic ways. For example, in the worst case, a celebrated result in algorithmic game theory shows that computing a NE is computationally hard, even when players know each other's utility functions [86]. When players are both learning from one another, we also know that no notion of "uncoupled" dynamics will succeed in approaching NE for certain non-zero-sum games, even asymptotically [87].

At a higher level, it stands to reason that Alice should find it easier to interact with a non-cooperative but non-adversarial Bob, as we saw in the spectrum enforcement games. The catch is that Bob's utility function is still unknown, and Alice may still have to worry about the possibility of Bob being totally malicious[13]. What could happen in such non-cooperative interaction is explored at length in Chapter 4 of this thesis. As demonstrated in Figure 1.6, studying how Alice should optimally learn from Bob's data becomes unavoidably intertwined with the question of how Bob should optimally shape his data in response to Alice's learning. Rather than formalizing optimality for Alice and Bob in isolation, we need to appeal to repeated game equilibria concepts. Chapter 4 postulates explicit strategies that Alice and Bob could follow in an approximate notion of sub-game-perfect equilibrium, and formalizes the beginnings of a new frequentist paradigm for repeated game theory. While the classical study of repeated game equilibria is Bayesian [76, 78], which allows the concept to be exactly formalized, we really want to understand *simple and explicit algorithms that Alice and Bob will use to interact*, as opposed to just their summary properties (such as eventually garnered utility). On the flip-side, formalizing frequentist guarantees in a game-theoretic environment remains highly challenging, and our understanding as of now is preliminary.

The above insights are especially applicable in a *one-sided learning* setup, where only Alice is learning from Bob. This greatly simplifies the problem. When multiple agents are learning from one another in a non-zero-sum setting, many questions are open. Frequentist perspectives in the non-zero-sum milieu are more on the heuristic side, although decentralized (internal) no-regret dynamics do have the intriguing property of converging to the broader concept of correlated equilibrium. For a survey on this literature, see [22, Chapter 4].

## 1.4 Intersecting learning and strategic behavior

This dissertation is fundamentally about a first-principles effort to intersect learning and strategic behavior, and is split into two parts from the philosophical point of view of "Alice", the learner, and "Bob", the unknown strategic agent. From the point of view of Alice, the primary question is *constructive*: how should she learn from data provided by an "unknown" Bob, who could be stochastic, competitive, adversarial or cooperative? Chapters 2 and 3 examine candidate algorithms for Alice that can adapt between two of the intellectually easier categories: stochastic and adversarial. To engage with the more nuanced (and realistic) categories of competitively and cooperatively generated data, we need to take Bob's point of view. The primary question then becomes *intellectual*: what are the core underlying principles that approximate optimal, or equilibrium, strategic behavior for Bob in the presence of a learning agent (Alice)? Chapters 4 and 5 formulate and solve initial lines of inquiry into this question under various informational frameworks.

[13]This is, in fact, used as informal justification for studying no-adversarial-regret algorithms even in non-zero-sum games, although we note that this justification is significantly weaker, and accordingly the economics literature designates these algorithms as "adaptive heuristics".

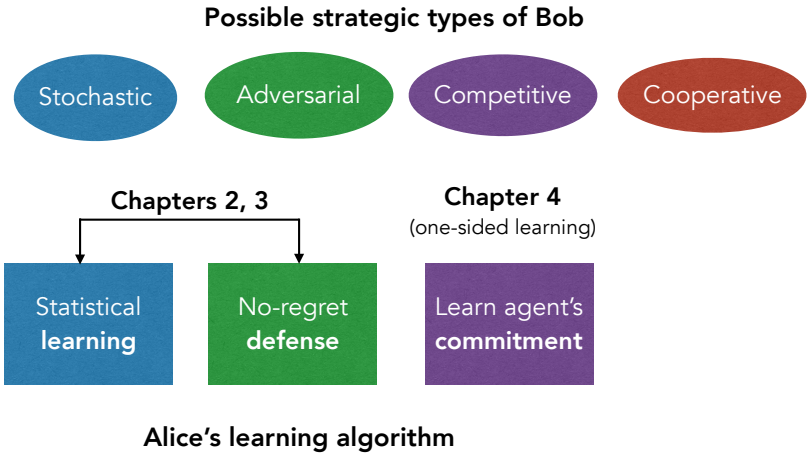



Figure 1.7: Composite strategic types of Bob and approximately optimal learning strategies used by Alice corresponding to each type. This figure was made using Keynote.

## Part I: Adaptivity In Alice's Learning

Part I of the thesis focuses on the methodological design of algorithms for Alice that adapt between statistical and strategic environments. If Alice knew Bob's type before-hand, Section 1.3 shows the way for designing her optimal learning algorithm. An idealistic goal for Alice would be to *adapt* to Bob's type, and obtain performance almost as good as though she had known the type beforehand. This goal of adaptivity between at least two types, stochastic and adversarial, frames Chapters 2 and 3 of this thesis. As depicted in Figure 1.7 hypothesis-testing her observations of Bob's behavior for one of the multiple possible types in a round-robin manner. Once Alice "detects" one of these types, she would then proceed with the corresponding optimal learning algorithm. This approach does not easily work against strategic environments, which can manifest in subtle ways: for example, adversarial agents can masquerade as stochastic for a while and fool naive "round-robin" approaches. In Chapter 2, we formalize the goal of adaptivity through the lens of full-information online learning, and build on sophisticated adaptive algorithms in the online learning literature that were designed with the goal of regret-minimization. We expand this perspective by showing that regret minimization is not always synonymous with reward maximization, and use this to motivate adaptive offline benchmark, or model selection, as an important component in adaptive learning algorithms. Chapter 2 designs *successful adaptive model selection between both stochastic and adversarial environments* in the full-information feedback environment. Work in this chapter is joint with Peter L. Bartlett, Mitas Ray, and Anant Sahai, and is contained in the paper [88]. Chapter 3 provides similar guarantees in the more challenging model of limited-information feedback. Work in this chapter is joint with Peter L. Bartlett and Niladri S. Chatterji, and is contained in the paper [89].

Even between stochastic and adversarial responses, several open problems remain in

the broad goal of adaptivity, particularly when limited-information and dynamic feedback is present, and with modern ML models. We briefly discuss some of these problems in Chapter 6.

## Part II: Bob's Incentives In The Presence Of Learning

As depicted in Figure 1.6, understanding Alice's "optimal" strategy against a competitive Bob with unknown utility function resists a clean decision-theoretic formulation, precisely because the interaction is one of a repeated non-zero-sum game with incomplete information. In other words, Alice's optimal learning strategy has to be analyzed in conjunction with Bob's optimal strategy in the presence of learning. Part II of this thesis focuses on candidate rules for Bob's optimal strategy in the presence of Alice, who is learning from her past interactions.

Chapter 4 examines this interaction under a major simplifying assumption of *one-sided learning*: in other words, Bob knows Alice's utility function, but not vice-versa, so Alice is the only player engaging in learning. This assumption of one-sided learning, also called information asymmetry, allows us to make a natural connection to the economics literature on reputation building under Bayesian formulations of incomplete information. Our aim is to understand approximate sub-game-perfect equilibrium strategies for both Alice and Bob in the frequentist milieu. While this is a very challenging task, and we do not entirely succeed in this goal, Chapter 4 designates strategies for Alice and Bob that satisfy multiple criteria for credibility and are thus strong candidates for an approximate repeated-game equilibrium. Our finding is surprising on the surface, but is entirely consistent with the indirect conclusion of classical Bayesian formulations: *Bob implicitly shares his information with Alice in these strategies.* This constitutes a very interesting incentive alignment that happens in a setting of non-cooperative interaction. Work in this chapter is joint with Anant Sahai, and is contained in the paper [90] (see [91] for a shorter version that appeared at ACM Economics and Computation 2019).

With *two-sided learning*, the questions become even more preliminary, and a first principles study of the eventual outcome of commonly used dynamics is essential as a starting point. Chapter 5 outlines such a study in the simplest case of zero-sum games, and shows some surprising properties of the ubiquitous no-regret dynamics. Work in this chapter is joint with Soham Rajesh-Phade and Anant Sahai, and is contained in the forthcoming preprint [92].

Chapter 6 concludes this dissertation with a brief discussion of two topics crucial to this research agenda that were not directly explored in this thesis; namely: a) engaging with modern ML models for agents, and b) understanding cooperative behavior.

# Part I

# Adaptivity In Learning

# Chapter 2

# Adaptivity in online prediction

We begin by describing the problem of adapting between a class of stochastic environments and an adversarial environment that is generating the data. Our goal is to maximize *reward* almost as well as if we had known the unique environment generating the data beforehand.

Note that this goal of adaptivity, from a methodological perspective, is immediately important in several practical applications. A notable example is mechanisms in online marketplaces, e.g. auctions, that were discussed in Section 1.1 of this thesis. An auctioneer repeatedly interacting with bidders aims to maximize her revenue well over the single-shot auction setting, by attempting to predict the valuations of their future customers from historical data provided by previous customers. Naturally, the auctioneer needs to "robustify" her mechanism to guard against strategic bidding. However, if the bidders *turn out* to behave predictably, she would like to leverage this to greatly increase her revenue. An ideal mechanism for the auctioneer would be one that is robust to strategic manipulation, but also exploits predictable bids.

The broadly stated goal of adaptivity in online learning has a rich history [93–99], and continues to be an extremely active area of research [100–105]. Much (but not all) of this work centers around adaptive *regret* guarantees with respect to a fixed offline benchmark. However, we demonstrate in this chapter (in Section 2.2) that the goal of minimizing regret is not always synonymous with maximizing reward, and offline benchmark selection has to be a central component of adaptivity in online learning, even in the most basic sequential prediction paradigm. The main contribution of this chapter is to design online learning algorithms that are able to adapt in two important ways:

1. Between stochastic and adversarial environments of a given model complexity.

2. Between different model classes within a stochastic (or adversarial) environment.

The problem of offline benchmark selection is synonymous with model selection, and thus our main contribution in this chapter is to develop methodology for *online model selection*. Data-driven model selection also has a rich history in purely stochastic environments dating back to ideas of model complexity penalization (Akaike's information criterion and Bayesian

information criterion heavily used in applied statistical methodology, structural risk minimization that underlies statistical learning theory [106–108]), and data-driven validation of model classes, the latter of which is ubiquitous in the practice of modern machine learning. Our online model selection algorithms use ideas of complexity penalization and validation respectively, and are thus directly inspired by this history.

## 2.1 Basic setup

In the spirit of supervised machine learning, we will consider a contextual prediction setting over $T > 0$ rounds, in which we receive context-output pairs $(X_t, Y_t)_{t=1}^T$. We consider $X_t \in \mathcal{X}^D, Y_t \in \mathcal{X}$, where $\mathcal{X} = \{0, 1\}$ is the binary alphabet[1]. It will also be natural to consider the truncated version of $X_t$ that only represents the last $d$ coordinates – we denote this by $X_t(d)$, with the convention that $X_t := X_t(D)$. Note that this includes the traditional *contextual tree prediction* paradigm [109], as well as the *universal sequence prediction* paradigm [110] in which the context $X_t(d) = \{Y_s\}_{s=t-d}^{t-1}$ comprises of previous observation values itself.

We follow the online supervised learning paradigm: before round $t$, we are given access to $X_t$, but not $Y_t$. Let $\mathcal{F}_D$ denote the set of all tree experts, expressed as Boolean functions from $\mathcal{X}^D$ to $\mathcal{X}$. We will also be considering tree experts that map from the sub-contexts $\{X_t(h)\}$ to outputs $Y_t$, denoted by $\mathbf{f}_h \in \mathcal{F}_h$ for all values of $h$ in $\{0, 1, \ldots, D\}$. (In universal prediction, these can be thought of as *finite-memory predictors*.) We use the shorthand notation $\mathbf{f} := \mathbf{f}_D \in \mathcal{F}_D$. We define the *order* of a tree expert, denoted by $\mathrm{order}(\mathbf{f}_h)$, as the minimum value of $d \leq h$ for which its functionality can be expressed equivalently in terms of a function from $\mathcal{X}^d$ to $\mathcal{X}$. That is,

$$\mathrm{order}(\mathbf{f}_h) := \min\{d \leq h : \text{ there exists } \mathbf{f}'_d \in \mathcal{F}_d \text{ s.t. } \mathbf{f}_h(x(h)) = \mathbf{f}'_d(x(d)) \text{ for all } x(h) \in \mathcal{X}^h\}. \tag{2.1}$$

We define our randomized online algorithm for *prediction using tree experts* in terms of a sequence of probability distributions $\{\mathbf{w}_t^{(\mathsf{tree})}\}_{t=1}^T$ over the set $\mathcal{F}_D$ of all tree experts. Note that $\mathbf{w}_t^{(\mathsf{tree})}$ cannot depend on $\{(X_s, Y_s)\}_{s \geq t+1}$ or $Y_t$. We denote the realization of the prediction at time $t$ by $\widehat{Y}_t \in \mathcal{X}$, and the distribution on $\widehat{Y}_t$ by $\mathbf{w}_t$ (clearly induced by $\mathbf{w}_t^{(\mathsf{tree})}$). After prediction, the actual value $Y_t$ is revealed, and the expected loss is modeled as $0-1$ loss depending on whether we get the prediction right. Formally, we have $\mathbf{l}_t = \big[\mathbb{I}[Y_t \neq 0] \quad \mathbb{I}[Y_t \neq 1]\big]$, and the expected loss of the algorithm in round $t$ is given by $\langle \mathbf{w}_t, \mathbf{l}_t \rangle = w_{t,1-Y_t}$. We also call this the expected *prediction error* of the algorithm. We denote as

[1] As a general note, all our analysis can easily be extended to the $m$-ary case. We present the binary case for simplicity.

shorthand

$$L_{t,\mathbf{f}} := \sum_{s=1}^{t} \mathbb{I}[Y_s \neq \mathbf{f}(X_t(h))] \text{ for all } \mathbf{f} \in \mathcal{F}_h, h \leq D$$
$$L_{X,t,y} := \sum_{s=1}^{t} \mathbb{I}[X_s = X; Y_s \neq y] \text{ for all } X \in \mathcal{X}^h, h \leq D, y \in \mathcal{X}$$
$$\mathbf{L}_{X,t} := \begin{bmatrix} L_{X,t,0} & L_{X,t,1} \end{bmatrix} \text{ for all } X \in \mathcal{X}^h, h \leq D.$$

The traditional quantity of *regret* measures the loss of an algorithm with respect to the loss of the algorithm that possessed oracle knowledge of the best single "action" to take in hindsight, after seeing the entire sequence offline. In the context of online supervised learning, this "action" represents the best $d^{th}$-order Boolean function $\widehat{F}_d(T) \in \mathcal{F}_d$. The expected regret with respect to the best $d^{th}$-order tree expert is defined as $R_{T,d} := \sum_{t=1}^{T} \langle \mathbf{w}_t, \mathbf{l}_t \rangle - L_{T,\widehat{F}_d(T)}$.

## Stochastic-vs-adversarial assumptions on data

In general, we make no assumptions on our data and we would like to minimize the *adversarial* regret $R_{T,d}$ for every value of $d$. The optimal scaling for this regret, with matching lower bound, is known (e.g. [84]) to be

$$R_{T,d} = \mathcal{O}(\sqrt{T \cdot 2^d}). \tag{2.2}$$

However, as we have mentioned informally, we would like to get greatly improved regret rates for data generated in a certain way (without a-priori knowledge of such generation). To start, we work with the following general *stochastic, stationary, predictable condition* on the responses given the covariates.

**Definition 2.1.1** (Stationary $d^{th}$-order stochastic condition on responses)**.** *We say that our data $(X_t, Y_t)_{t \geq 1}$ satisfies the stationary stochastic condition on responses if, at every round $t$, we have*

$$Y_t \Big| \{X_t(h), (X_{t-1}, Y_{t-1}), \ldots, (X_1, Y_1)\} \sim P^*(\cdot | X_t(h)) \tag{2.3}$$

*for all $X_t(h) \in \mathcal{X}^h$ and $h \in \{0, 1, \ldots, D\}$. More-over, the condition is said to be $d^{th}$-order realizable if we have $Y_t | \{X_t, (X_s, Y_s)_{s=1}^{t-1}\} \sim P^*(\cdot | X_t(d))$ for all $t$ and all $X_t \in \mathcal{X}^D$. Note that the realizability condition implies that $Y_t$ is independent of all previous observations given $X_t(d)$.*

For this setting, it is natural to define the best "external predictor" for any $h \leq d$:

$$f^*(x(h)) :\in \arg\max_{y \in \mathcal{X}} P^*(y | x(h)) \text{ for all } x(h) \in \mathcal{X}^h. \tag{2.4}$$

We further assume that the best $d^{th}$-order predictor is unique[2], i.e.

$$P^*(f^*(x(d))|x(d)) > P^*(y|x(d)) \text{ for all } y \neq f^*(x(d)) \text{ and for all } x(d) \in \mathcal{X}^d.$$

and denote the parameter[3]

$$\beta(x(d)) = P^*(f^*(x(d))|x(d)) \tag{2.5}$$

$$\beta_d^* := \min_{x(d)\in\mathcal{X}^d} \beta(x(d)). \tag{2.6}$$

In tandem with the stationary stochastic condition on responses conditioned on contexts, we need to specify the generative model for the contexts $\{X_t\}_{t\geq 1}$. We specify three generative models that we will use to prove our results.

**Definition 2.1.2** (Independent, identically distributed contexts.)**.** *This case considers $X_t$ i.i.d $\sim Q_D^*(\cdot)$, where $Q_D^*$ constitutes a distribution on $\mathcal{X}^D$ that is supported on the whole of $\mathcal{X}^D$. We also denote the marginal distributions on $X_t(h)$ by $Q_h^*(\cdot)$ for all $h = 0, 1, \ldots, D$.*

**Definition 2.1.3** ($d^{th}$-order Markovian model.)**.** *This case considers contexts to be previous realizations of the responses themselves, i.e. $X_t = (Y_{t-D}, \ldots, Y_{t-1})$, thus, the process $\{Y_t\}_{t\geq 1}$ constitutes a $d^{th}$-order Markov process. We assume that this Markov process has a stationary distribution; thus, overloading notation, we define $Q_h^*(\cdot)$ as the stationary distribution on $(Y_{t-h}, \ldots, Y_{t-1})$ for all $t \geq h+1$, and all $h = 0, \ldots, D$. We again assume that $Q_h^*(\cdot)$ is supported on the whole of $\mathcal{X}^h$ for all $h = 0, \ldots, D$.*

**Definition 2.1.4** (Periodic contexts.)**.** *This case considers* ***deterministic, periodic contexts***

$$X_t := (X_{t-1} + 1) \mod 2^D, \tag{2.7}$$

*and[4] $X_1 \sim$ Unif$(\mathcal{X}^D)$. Note that this automatically implies that $X_t(h) := (X_{t-1}(h) + 1) \mod 2^h$ and, marginally, $X_1(h) \sim$ Unif$(\mathcal{X}^h)$ for any $h \in \{0, 1, \ldots, D\}$.*

The last definition of periodicity in contexts is quite a strong one; we make it primarily to illustrate the heart of our analysis of online model selection through validation. We use the more general assumptions of stochastic contexts (iid and Markov) for our first online model selection procedure, which uses principles of structural risk minimization. Since our proofs are essentially identical for both the iid and Markov cases, we will treat these cases together in our theorem statements as well as proofs.

For all of the cases above, we can define the important notions of asymptotic *unpredictability* for all model orders $h \in \{0, 1, \ldots, D\}$. The definitions and notation are directly inspired by information-theoretic limits on sequence compression and prediction [110].

[2]This is the fundamental *Tsybakov margin condition* [111] that is essential for eventual learn-ability of the best predictor.

[3]Note that the uniqueness of best-predictor assumption directly implies that $\beta^* > 1/2$, since we are working with a binary alphabet.

[4]We add the randomness into the first context for convenience in defining the expected unpredictability $\pi_h^*$ to be identical for all rounds.

**Definition 2.1.5** ( [110]). *For data $(X_t, Y_t)_{t\geq 1}$ satisfying the stationary stochastic condition, and the first two cases of iid contexts (Definition 2.1.2) and Markovian responses (Definition 2.1.3), we define its asymptotic unpredictability under the $h^{th}$-order predictive model by*

$$\pi_h^* := \sum_{x(h)\in\mathcal{X}^h} Q_h^*(x(h)) \left[1 - \max_{y\in\mathcal{X}}\{P^*(y|x(h))\}\right]. \tag{2.8}$$

*For the case of periodic contexts (Definition 2.1.4), we define asymptotic unpredictability under the $h^{th}$-order predictive model by*

$$\pi_h^* := \frac{1}{2^h} \sum_{x(h)\in\mathcal{X}^h} \left[1 - \max_{y\in\mathcal{X}}\{P^*(y|x(h))\}\right]. \tag{2.9}$$

*In both cases, the sequence $\{\pi_h^*\}_{h=0}^D$ can easily be verified to decrease in $h$.*

In all the contextual models, it is easy to show that the optimal *Follow-the-Leader* algorithm, tailored to model order $d$, can give us regret

$$R_{T,d} = \mathcal{O}\left(\frac{2^d}{(2\beta_d^* - 1)^2}\right). \tag{2.10}$$

As we will see in the next subsection, the regret rates in Equations (2.2) and (2.10) can be simultaneously achieved *with a-priori knowledge of the model order $d$* using existing adaptive algorithms.

## Adaptive algorithms for a known model order

In this section, we present a simple generalization of the AdaHedge algorithm [94, 95, 98] as our choice for the "base" adaptive algorithm corresponding to each model order $d$. Before specifying the actual choice of learning rate, we start with the definition of a base algorithm corresponding to model order $d \in \{0, 1, \ldots, D\}$.

**Definition 2.1.6.** *The base algorithm $\mathcal{A}_d$ corresponding to function class $\mathcal{F}_d$ comprises of an exponential weights update over functions in $\mathcal{F}_d$:*

$$(w_{t,\mathbf{f}}^{(\mathsf{tree})})^{(d)}(\eta_t^{(d)}) \propto \begin{cases} e^{-\eta_t^{(d)}\cdot L_{t-1,\mathbf{f}}} \text{ if } order(\mathbf{f}) \geq d \\ 0 \text{ otherwise .} \end{cases}$$

*The overall weight vector is denoted by $(\mathbf{w}_t^{(\mathsf{tree})})^{(d)}(\eta_t^{(d)})$.*

A good data-adaptive choice of $\{\eta_t\}_{t\geq 1}$ has been an intriguing question of significant recent interest. The idea (borrowing language from [112]) is that we want to "learn the correct learning rate" for the problem. We consider a particularly elegant choice based on the algorithm AdaHedge, that was defined for the simpler experts setting. We denote $\eta_{s_1}^{s_2} = \{\eta_s\}_{s=s_1}^{s=s_2}$ for shorthand.

**Definition 2.1.7** ([98]). *The* AdaHedge *learning rate process* $\{\eta_t^{(h)}\}_{t\geq 1}$*, corresponding to every base algorithm* $\mathcal{A}_h$ *for every* $h \in \{0, 1, \ldots, D\}$*, is described as*

$$\eta_t^{(h)} = \frac{2^d \ln 2}{\Delta_{t-1}^{(d)}((\eta^{(d)})_1^{t-1})}, \tag{2.11}$$

*where the cumulative and instantaneous mixability gaps are defined as below:*

$$\Delta_t^{(d)}((\eta^{(d)})_1^t) := \sum_{s=1}^{t} \delta_s^{(d)}(\eta_s^{(d)}), \text{ where} \tag{2.12}$$

$$\delta_s^{(d)}(\eta_s^{(d)}) := \langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, \mathbf{l}_s^{(\mathsf{tree})} \rangle + \frac{1}{\eta_s} \ln \langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, e^{-\eta_s \mathbf{l}_s^{(\mathsf{tree})}} \rangle \tag{2.13}$$

$$\tag{2.14}$$

For every choice of $h$, the base algorithm $\mathcal{A}_h = \text{AdaHedge}(h)$ can easily be shown to obtain the adversarial regret guarantee $R_{T,h} = \mathcal{O}(\sqrt{T \cdot 2^h})$. Proposition 2.1.8 (restated in Section 2.6 as Proposition 2.6.18) illustrates the stochastic regret guarantee obtainable by $\mathcal{A}_h$.

**Proposition 2.1.8.** *Let* $h \geq d$*. For any sequence* $\{X_t, Y_t\}_{t=1}^T$*, we have*

$$R_{T,d}(\mathcal{A}_h) = \mathcal{O}\Big(\frac{2^h}{(2\beta^* - 1)} \sqrt{\Big(h + \ln\Big(\frac{(D-d)}{\epsilon(2\beta^* - 1)^2}\Big)\Big)}\Big)$$

*simultaneously for all* $h \in \{d, \ldots, D\}$ *with probability at least* $(1-\epsilon)$ *on a sequence* $(X_t, Y_t)_{t\geq 1}$ *that satisfies the* $d^{th}$*-order stochastic condition with parameter* $\beta^*$*.*

The proof of Proposition 2.1.8 is simply a standard application of the proof of AdaHedge to contextual prediction, and is also contained in Section 2.6. Observe that the dependencies of the regret guarantee on the number of rounds, $T$, is optimal in both adversarial and stochastic environments—however, the dependence on the true model order $d$ critically depends on the choice of $h$. Observe that if we just picked $h = d$, the above proposition provides the optimal stochastic regret guarantee; however, if $h > d$, the dependence on $h$ is exponential and thus highly sub-optimal. Also, observe that any guarantee on $R_{T,d}$ vanishes completely if we instead chose $h < d$, as we will demonstrate next, this regret will in general be linear in $T$.

We can already see from the above discussion that there is a sizable cost to both over-specifying and under-specifying the model order in regret minimization. This provides a hint that the very goal of regret minimization does not tell the full story. We now make the limitations of regret clear through a simple example.

## 2.2 The need for adaptive model selection/The limitations of regret

The simplest use case of online learning constitutes the example of binary sequence prediction, where there are no contexts $\{X_t\}_{t\geq 1}$, but there are responses $\{Y_t \in \{0,1\}\}_{t\geq 1}$ that we wish to predict from past data. Here, regret is minimized with respect to the performance obtained by predicting the best fixed choice of letter — $y = 0$ or $y = 1$ — on all rounds. This corresponds to the quantity $R_{T,0}$ in our formulation of contextual prediction.

Binary sequence prediction by itself has seen sizable attention dating back to the work of Thomas Cover, and more recently in the context of adaptivity [110, 113–115]. This is because the binary sequence prediction problem, while appearing simple on the surface, captures much of the central non-trivialities in online learning[5]. We now consider the composition of unpredictable, or "adversarial" binary sequences relative to a given choice of regret-minimizing algorithm. For example, deRooij et al [98] considered the following choice of binary sequence as "unpredictable" with respect to several algorithms, including AdaHedge, that adapt between stochastic and adversarial environments:

$$Y_t = \begin{cases} 1 \text{ if } t \text{ even} \\ 0 \text{ if } t \text{ odd.} \end{cases} \tag{2.15}$$

This example was taken slightly further by Gravin, Peres, and Sivan [116], who studied the optimal choice of adversarial sequence against exponential weights families.

**Definition 2.2.1.** *The sequence that is **adversarial** to exponential weights families of algorithms is defined for a fixed number of rounds $T$ as follows: Fix $T_0 := T - T^p$ for some $0 < p < 1$. Then, we pick*

$$Y_t = \begin{cases} 1 \text{ if } t \leq T_0, \ t \text{ even} . \\ 0 \text{ if } t \leq T_0, \ t \text{ odd} . \\ 1 \text{ otherwise.} \end{cases} \tag{2.16}$$

Notably, the adversaries in Equations (2.15) and (2.16) can both be shown to incur $\omega(\sqrt{T})$ on exponential-weights families, including sophisticated methods that learn the learning rate like AdaHedge. The reason that this property continues to hold for AdaHedge is, roughly, as follows: since the number of 1's never becomes significantly greater than the number of 0's; thus, for most of the duration of the game, AdaHedge treats these sequences as unpredictable. This is clearly undesirable behavior: the "adversary", as described above, is easily predictable to the human eye—simply predicting a 1 after seeing a 0, and vice-versa, would lead to very few prediction errors on both examples.

[5]Indeed, the example that demonstrates the $\omega(\sqrt{T})$ lower bound on regret is precisely a binary sequence prediction example [81, Chapter 3].

We can define a "FOLLOW-THE-LEADER(1)" algorithm (henceforth abbreviated to FTL) with this spirit, as below:

$$\widehat{Y}_t := \begin{cases} 1 \text{ if } \sum_{s:Y_{s-1}=Y_{t-1}} Y_s \geq \sum_{s:Y_{s-1}=Y_{t-1}} (1 - Y_s) \\ 0 \text{ otherwise .} \end{cases} \tag{2.17}$$

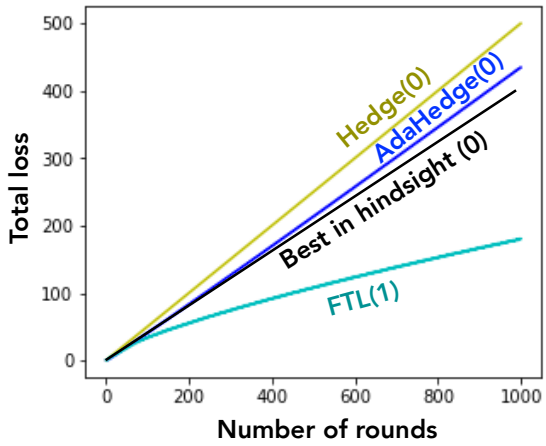



Figure 2.1: Sub-optimality of $0^{th}$-order regret minimizing algorithms on "adversarial" sequence due to under-fitting. Figure from [88].

Equation (2.17) simply uses the maximum likelihood estimation principle for Bayes-optimal classification of the sequence into $\{0, 1\}$ at every round. Thus, it corresponds to the optimal algorithm we would use if we *knew* that $\{Y_t\}_{t\geq 1}$ evolves according to a 1-memory Markov process. It can be verified that this approach would clearly incur only constant, as opposed to $\sqrt{T}$, regret on the adversarial sequence — however, the regret metric $R_{T,0}$ does not capture the full extent of the benefit. Figure 2.1 compares the *overall loss* (i.e. number of prediction errors) of three algorithms: HEDGE(0), ADAHEDGE(0), and FTL(1). It also provides, for reference, the overall loss of the best letter in hindsight (which, here, corresponds to predicting 1 all the time). It is clear that FTL(1) is doing even better (and by a significant amount) than the offline benchmark, and thus its improvement over the other regret-minimizing algorithms is actually *linear* in $T$.

Thus, adversaries designed for online learning algorithms using the $0^{th}$-order offline benchmark tend to be highly predictable under a more complex benchmark (here, corresponding to 1-memory Markov predictors). This suggests that we may want to increase the complexity of our offline benchmark so as to "catch" all such complex, but highly predictable patterns. In the above reasoning, we do not want to stop at predictors using a memory of length 1: of course, a sequence with memory of dependency 2 can be designed to be an "adversary" to all such algorithms using an offline benchmark. In general, we may want to define some maximal memory length $D$ for our benchmark, and use a regret-minimizing algorithm with

respect to this benchmark. Such an algorithm would capture all statistical patterns in the data, including the most complex possible, if they existed. The equivalent algorithms for this case, as discussed in Section 2.1, are FTL($D$), HEDGE($D$), and ADAHEDGE($D$). Notably, the classical *universal prediction paradigm* [110, 113] takes this reasoning even further and uses prediction schemes inspired by Lempel-Ziv compression to *asymptotically* capture stochastic patterns with an infinite memory, i.e. $D \to \infty$.

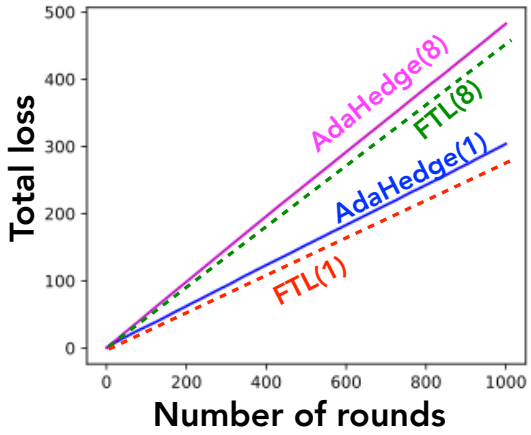



Figure 2.2: Sub-optimality of regret-minimizing algorithms with respect to a higher-than-needed model order, e.g. ADAHEDGE(8), on 1-memory sequences due to over-fitting. Figure from [88].

However, de-facto using the most complex model order is *not* optimal in a non-asymptotic sense. Figure 2.2 shows the stark sub-optimality in AdaHedge($D$)'s overall loss with respect to AdaHedge(1). In regret-minimizing language, while AdaHedge($D$) does incur a *constant* regret that does not grow with $T$, the value of this constant is too large. Instead of the best possible regret guarantee of $R_{T,1} = \widetilde{\mathcal{O}}(2^1)$, ADAHEDGE($D$) incurs a much larger regret scaling as $R_{T,1} = \widetilde{\Theta}\left(2^D\right)$. This larger constant is what is showing up in Figure 2.2 as a high sub-optimality in performance.

Thus, it is clear that we need to adaptively select the offline benchmark as a function of the properties of the data: there is a significant price paid in overall loss (or reward) by under-specifying *or* over-specifying the model order. In general, the true model order $d$ (in this example, $d = 1$) is unknown to the algorithm. This implies that in addition to adapting between stochastic and adversarial environments, data-adaptive model selection is essential to successful adaptive online learning methodology. We now discuss the prior art on model selection in stochastic and adversarial environments.

## 2.3 Related work in stochastic and adversarial model selection

Data-driven model selection in *offline*, purely stochastic environments is central to applied statistics and machine learning methodology; thus, it has a rich history. Traditional statistical learning theory concepts such as the Vapnik-Chervonenkis dimension and Rademacher complexity [117, 118] demonstrate a general principle by which the estimation error (which is synonymous with regret) increases as we increase the complexity of the model class (which is synonymous with the complexity of the offline benchmark). Therefore, classical approaches to model selection explicitly penalize overly complex models in a *model selection criterion*. This criterion is known as Akaike's information criterion or Bayes' information criterion [119] in classical statistics, and structural risk minimization (SRM) [106] in statistical learning theory. More recently, SRM has been shown to perform data-adaptive model selection successfully with high probability through estimation error bounds that are obtained via empirical process theory. Theoretical work on SRM is broad in scope and we refer the reader to [108] for a review. More recently, an alternative method of model selection constitutes evaluating a trained procedure tailored to a particular model class on a separate "hold-out" set, and using the measured error as a proxy for the population test error that using this model class would incur. The most sophisticated variants on this validation-style approach include *cross-validation*, which interleaves a single data-set into multiple training and "hold-out" categories [120, 121].

The SRM approach can be directly plugged into full-information online learning algorithms *when the existence of stochasticity is known*; while the validation approach is non-trivial to analyze even when stochasticity is known. Whether SRM or validation-based approaches are used, the equivalent problem of purely stochastic model selection in limited-information feedback settings is far more challenging. We discuss partial solutions to model selection under bandit feedback in Chapter 3, but several open problems remain.

On the other side, the paradigm for *adversarial* regret minimization was laid out in the discrete "experts" setting in seminal work (for a review, see [84]), and subsequently lifted up to the more general *online convex optimization* framework (for a review, see [122]). The next natural goal was adaptivity to several types of "easier" instances while preserving the worst-case guarantees. Most pertinent to our work are the easier *stochastic losses* [98], under which the greedy Follow-the-Leader algorithm achieves regret $\mathcal{O}(1)$. In the experts setting, multiple algorithms have been proposed [94, 95, 98, 100, 101, 123] that adaptively achieve $\mathcal{O}(1)$ regret. Some of these guarantees have been extended to online optimization [103]. As we will see, naively extending these analyses to the contextual prediction problem gives a pessimistic $\mathcal{O}(2^D)$ regret bound. In our work, we show that we can get the best of *many worlds* and greatly improve the exponent to $\widetilde{\mathcal{O}}(2^{2d})$, reducing the dependence on the maximum model complexity $D$ from exponential to linear.

Recent guarantees on adapting to a simpler model class, *but not to stochasticity*, have also been developed [93, 97, 99–102, 104]. Many of these approaches [97, 99, 102, 104] do not

improve the $\widetilde{\mathcal{O}}(\sqrt{T})$ rate for stochastic data, and can thus be thought of as purely adversarial algorithms. However, some of them are notably far broader in scope than the binary sequence paradigm, and deal with online optimization [99] and online supervised learning [104] with "multi-scale" predictors.

The problem of simultaneous stochastic and adversarial model selection is also not entirely new. We address in particular two recent algorithms, AdaNormalHedge [101] and Squint [100], both of which obtain second-order quantile regret bounds in terms of a "variance" term and the correct model complexity. The *analysis* of both AdaHedge and AdaNormalHedge *in the stochastic regime* avoids the model selection issue, and yields a pessimistic $\mathcal{O}(2^{D+d})$ regret bound. Squint cleverly applies the Bernstein condition [123] and obtains the optimal stochastic rate of $\mathcal{O}(2^d)$ for the special "realizability" case. However, the elegance of these approaches comes at some cost to broader applicability, as discussed below:

1. The computational complexity of Squint necessarily scales linearly with the number of experts, which in the case of the contextual prediction problem is a prohibitive *double-exponential-in-D* complexity.

2. The analysis of both AdaNormalHedge and Squint in the stochastic regime requires data to be realizable under the $d^{th}$-order stochastic condition, and does not explicitly connect to statistical model selection frameworks.

3. AdaNormalHedge and Squint are designed for explicit complexity hierarchies in model classes; thus, will not yield meaningful model selection guarantees even empirically in modern environments where highly complex (indeed *over-parameterized* with respect to data) models provide the most successful performance [124, 125].

For online model selection to be broadly applicable, we need all of these ingredients—i.e. computational efficiency, guarantees under model mis-specification, *and* flexible methodology in a variety of statistical environments beyond traditional complexity hierarchies. Regardless of the choice of algorithm, it appears that satisfying the first two conditions requires directly reasoning about model selection guarantees in a probabilistic sense. Both of the algorithms described in this chapter do precisely this by explicitly connecting to the methodology of SRM and validation. While our main results are still restricted to the realizable case of a $d^{th}$-order model generating the data, we are optimistic about being able to apply our techniques to analyze online prediction error in mis-specified environments in the future (for a detailed discussion on this point, see Section 2.7).

Finally, regarding the third condition, a theoretical analysis of online validation methodology is of especial interest, as it is the only known *empirically successful* methodology for model selection that does not explicitly encode model complexity information, and is therefore heavily used in modern machine learning. Our result constitutes one of the first such theoretical analyses in an online prediction environment, even if for traditional complexity hierarchies. As we discuss in Chapter 6, theoretical analysis of online validation in modern ML environments is a future topic of sizable interest.

## 2.4 Online model selection through complexity penalization

Our first algorithm leverages the framework for *offline* structural risk minimization (SRM) in purely stochastic environments. The central idea in SRM is to use a model selection criterion that, in addition to minimizing training error on data, penalizes overly complex models that have the potential to over-fit. In other words, the model selection criterion trades off *estimation error*, i.e. the complexity of the model class, and *approximation error*, i.e. the approximability of the best model-in-class to the actual data.

Since online prediction in adversarial environments needs to be randomized, our first algorithm essentially implements SRM in a Bayesian manner. Essentially, we use an exponential-weights update on tree experts equipped with a *time-varying, data-dependent learning rate* and a suitable prior distribution on tree experts, where the prior acts precisely as the model complexity regularizer. We start by describing the structure of the prior distribution.

**Definition 2.4.1.** *For any non-negative-valued function $g : \{0, 1, \ldots, D\} \to \mathbb{R}_+ \cup \{0\}$, we define the prior distribution on all tree experts in $\mathcal{F}_D$, $\mathbf{w}_{1,\mathbf{f}}^{(\mathsf{tree})}(g) = \frac{\sum_{h=order(\mathbf{f})}^{D} g(h)}{Z(g)}$, where $Z(g)$ is the normalizing factor.*

We select a function $g(\cdot)$ and use the prior defined above to effectively down-weight more complex experts. We will see that the choice of prior is crucial to recovering stochastic model selection. We now describe our algorithm.

**Definition 2.4.2.** *The algorithm* SRMOverAdaHedge($D$) *whose prior is derived from the function $g(\cdot)$ updates its probability distribution on tree expert as follows:*

$$w_{t,\mathbf{f}}^{(\mathsf{tree})}(\eta_t; g) = \frac{\left(\sum_{h=order(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}{\sum_{\mathbf{f}' \in \mathcal{F}_D} \left(\sum_{h=order(\mathbf{f}')}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}'}}}. \tag{2.18}$$

*and learning rate update $\{\eta_t\}_{t \geq 1}$ made as below:*

$$\eta_t = \frac{\ln 2}{\Delta_{t-1}(\eta_1^{t-1})}, \tag{2.19}$$

*where $\Delta_t(\eta_1^{t-1})$ is called the "cumulative mixability gap" at time $t$ and is given by*

$$\Delta_t(\eta_1^{t-1}) := \sum_{s=1}^{t} \delta_s(\eta_s) \text{ where} \tag{2.20}$$

$$\delta_s(\eta_s) := \langle \mathbf{w}_s(\eta_s),\, \mathbf{l}_s \rangle + \frac{1}{\eta_s} \ln \langle \mathbf{w}_s(\eta_s),\, e^{-\eta_s \mathbf{l}_s} \rangle. \tag{2.21}$$

The algorithm SRMOVERADAHEDGE($D$) appears to have a prohibitive computational complexity of $\mathcal{O}(|\mathcal{F}_D|) = \mathcal{O}(2^{2^D})$. However, the distributive law enables a clever reduction in computational complexity to $\mathcal{O}(2^D)$. The main idea is that instead of keeping track of cumulative losses of all the $2^{2^D}$ functions in $\mathcal{F}_D$, represented by $\{L_{t,\mathbf{f}}\}_{\mathbf{f}\in\mathcal{F}_D}$, we only need to keep track of the cumulative losses of making certain predictions as a function of certain contexts, represented by $\{\{L_{x,t,y}\}_{y\in\mathcal{X}}\}_{x\in\mathcal{X}^D}$. This reduction was first considered for tree expert prediction in the worst-case [109], with a fixed learning rate $\eta > 0$, and can easily be extended to the broader class of exponential-weights updates. The idea is that the update on probability distribution on *tree experts*, described in Equation (2.18) – can be equivalently written as a computationally faster update on probability distribution on *predictors*:

$$w_{t,y}(\eta_t; g) = \frac{\sum_{h=0}^{D} g'(h;\eta_t) e^{-\eta_t L_{X_t(h),t,y}}}{\sum_{h=0}^{D} g'(h;\eta_t) \left(\sum_{y\in\mathcal{X}} e^{-\eta_t L_{X_t(h),t,y}}\right)} \text{ where} \tag{2.22a}$$

$$g'(h;\eta_t) = g(h) \prod_{x(h)\neq X_t(h)} \left(\sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}}\right) \tag{2.22b}$$

The equivalence is in the sense that the expected loss incurred by updates (2.18) and (2.22a) is the same. The computational complexity of the latter update is $\mathcal{O}(2^D)$ per iteration, as shown in Proposition 2.7.8.

We now consider the realizable case in which the data is actually coming from model order $d$. We study the algorithm SRMOVERADAHEDGE($D$) with the following choice of model-order-proportional prior function.

$$g_{\mathsf{prop}}(h) = 2^{-2^{h+1}} \tag{2.23}$$

We will later see that this choice of prior effectively recovers the SRM framework in the online setting. Our first result shows that the algorithm with this choice of prior helps us effectively learn the model order while staying worst-case robust.

**Theorem 2.4.3.** *1. For any sequence $\{X_t, Y_t\}_{t=1}^T$, the algorithm* SRMOVERADAHEDGE($D$) *with prior defined according to function $g_{\mathsf{prop}}(\cdot)$ gives us regret rate*

$$R_{T,d} = \mathcal{O}\left(\sqrt{T}2^d\right) \tag{2.24}$$

*with respect to the best $d^{th}$-order tree expert in hindsight for every $d \in \{0, 1, \ldots, D\}$.*

*2. Consider any $\delta \in (0, 1]$. Let the stationary stochastic sequence $(X_t, Y_t)_{t\geq 1}$ satisfy the $d^{th}$-order stationary stochastic condition on responses given contexts with parameter $\beta^*$, and either of Definitions 2.1.2 or 2.1.3. Denote $\alpha_{d-1,d} := \frac{\pi^*_{d-1} - \pi^*_d}{2}$. Then,*

SRMOVERADAHEDGE($D$) *with prior function* $g_{\mathsf{prop}}(\cdot)$ *incurs regret with probability greater than or equal to* $(1-\delta)$*:*

$$R_{T,d} = \mathcal{O}\left(2^{2d}\left(\frac{d^2}{\alpha_{d-1,d}^2}\ln\left(\frac{d}{\alpha_{d-1,d}^2\epsilon}\right) + \frac{D\cdot d}{(\alpha^*)^2}\ln\left(\frac{D}{\alpha^*\epsilon}\right)\right)\right) \tag{2.25}$$

*where* $\alpha^* = \min\{\alpha_{d-1,d}, (2\beta^* - 1)\}$.

Note that this regret bound is non-trivial for model selection, but sub-optimal owing to the extraneous factor of $2d$ in the exponent.

## 2.5 Online model selection through validation

Theorem 2.4.3 above provides a positive, but non-trivial, model selection result. The main difficulty with adapting SRM to the online setting is meshing the idea of non-uniform priors, which adapts model complexity, with the data-dependent learning rate, which adapts between stochasticity and adversity. The sub-optimalities in our bounds are reminiscent of past discussions on the difficulty of adapting exponential weights-style algorithms to work with non-uniform prior weighting, and indeed the algorithms ADANORMALHEDGE and SQUINT were created to resolve this difficulty. However, as discussed in Section 2.3, the elegance of these approaches comes at the cost of their broader applicability in computational efficiency, robustness to model mis-specification and model selection beyond complexity hierarchies.

We aim for a resolution of this sub-optimality while aiming to preserve broad applicability. We now consider a second natural way of doing online model selection, also motivated by an ubiquitous approach to purely statistical model selection: *data-driven validation.* To see the usefulness of data-driven validation in online prediction, recall that Figures 2.1 and 2.2 displayed the sub-optimality of algorithms using under-specified *or* over-specified models in overall loss.

*The hope with an online validation approach is that we can fruitfully use the information provided by this algorithmic sub-optimality, in an online fashion, to perform optimal model selection.*

Remarkably, the roots of an online validation approach have been explored in purely adversarial environments as well. The seminal work of Hutter and Poland [93] use a "meta-expert" layer for randomized algorithm selection, and show that optimal purely adversarial model selection is possible. We define this meta-algorithmic framework below.

**Definition 2.5.1.** *For every base algorithm* $\mathcal{A}_h$*, we denote* $h_t(h) := \langle (\mathbf{w}_t^{(\mathsf{tree})})^{(h)}, \mathbf{l}_t^{(\mathsf{tree})}\rangle$ *and* $H_t(h) := \sum_{s=1}^{t-1} h_t(h)$*. Then, our meta-algorithm* METAALG($\{\mathcal{A}_h(\{\eta_t^{(h)}\}_{t=1}^T)\}_{h=0}^D; \{\eta_t\}_{t\geq 1}$) *chooses weighted prediction*

$$\mathbf{w}_t^{(\mathsf{tree})} := \sum_{h=0}^{D} q_t(h;\eta_t)(\mathbf{w}_t^{(\mathsf{tree})})^{(h)} \tag{2.26}$$

*where we have*

$$q_t(h;\eta_t) \propto e^{-\eta_t H_{t-1}(h)} \text{ for all } h \in \{0,1,\ldots,D\},$$

*and the learning rate $\eta_t$ is chosen as in Equation (2.19). We define the algorithm* ValidationOverAdaHedge($D$) *as* MetaAlg *when the base algorithms are chosen as* $\mathcal{A}_h = \text{AdaHedge}(h)$.

This algorithmic structure is inspired by the hierarchy-of-experts in adaptive FTPL (Theorem 9 of Hutter and Poland [93]) and can be implemented with any *meta-learning rate* schedule. While Hutter and Poland consider meta-learning rate schedules that lead to small-loss and quantile bounds, we use the AdaHedge meta-learning rate schedule to give the strongest possible stochastic regret bounds[6]. Further, we use the base algorithms $\{\mathcal{A}_d = \text{AdaHedge}(d)\}_{d=0}^{D}$. We call the resulting algorithm ValidationOverAdaHedge($D$). We show in Proposition 2.7.9 that the computational complexity-per-iteration[7] of ValidationOverAdaHedge($D$) is $\mathcal{O}(D)$; this is even better than the $\mathcal{O}(2^D)$ complexity for SRMOverAdaHedge($D$).

We will see that the resulting model selection analysis in stochastic environments manifests as a natural *online* form of leave-one-out cross-validation [120, 121], which cannot be reduced to the traditional analyses of AdaHedge in well-behaved stochastic environments [95]. Our model selection guarantees for ValidationOverAdaHedge($D$) are stated below.

**Theorem 2.5.2.** *1. For any sequence $\{X_t, Y_t\}_{t=1}^{T}$,* ValidationOverAdaHedge($D$) *with prior defined according to function $g_{\mathsf{prop}}(\cdot)$ gives us regret rate*

$$R_{T,d} = \mathcal{O}\left(\sqrt{T \cdot 2^d \ln 2} \cdot \ln D + \ln D + 2^d \ln 2\right) \tag{2.27}$$

*with respect to the best $d^{th}$-order tree expert in hindsight for every $d \in \{0,1,\ldots,D\}$.*

*2. Consider any $\delta \in (0,1]$. Let the stationary stochastic sequence $(X_t, Y_t)_{t\geq 1}$ satisfy the $d^{th}$-order stationary stochastic condition on responses given contexts with parameter $\beta^*$, and Definition 2.1.4 on the contexts. Then,* ValidationOverAdaHedge($D$)

[6]Note that, relative to SRMAdaHedge($D$), we have given the algorithm additional flexibility by using base algorithms that each use a different learning rate. (This idea also shows up in other adaptive online learning approaches, notably, to meet the goal of *multi-scale adaptivity*.) This additional flexibility will underlie improved dependencies in the exponent for model selection, with the caveat that the analysis is significantly more involved in the stochastic regime.

[7]We are considering time complexity here; it is easy to verify that the spatial complexity of the algorithm will still be $\mathcal{O}(2^D)$ as information pertaining to all $2^D$ of the contexts needs to be stored.

*incurs regret*

$$R_{T,d} = \mathcal{O}\Big(d^2 \cdot 2^d + d \ln\left(\frac{d}{\delta}\right) + d^2 \left(\ln\left(\frac{D}{\delta}\right)\right)^2 \cdot 2^d + D^3 \ln\left(\frac{D^2}{\delta} \ln\left(\frac{D}{\delta}\right)\right) + D^{3/2} \cdot d \cdot 2^d \left(\ln\left(\frac{D}{\delta}\right)^{3/2}\right)\Big) \tag{2.28}$$

$$\tag{2.29}$$

*with probability at least* $(1-\delta)$*. Ignoring the* $\ln(1/\delta)$ *dependencies, this gives us a regret bound* $R_{T,d} = \widetilde{\mathcal{O}}\left(d^2 \cdot 2^d + D^3 + D^{3/2} \cdot d \cdot 2^d\right)$*, which is optimal up-to polynomial factors in* $(d, D)$*.*

The proofs of Theorem 2.5.2 follows from a careful combination of adversarial-stochastic interpolation and structural risk minimization, and is deferred to the appendix. Of particular technical involvement is the proof of ruling out higher-order models in model selection (Appendix 2.6). This is difficult because it involves obtaining confidence intervals on averages of test errors across different rounds, which are highly dependent quantities. These dependencies also form the central non-triviality in obtaining confidence intervals on estimates of the test error via leave-$k$-out cross-validation [126–129]. However, here, we are able to deal with the dependencies through martingale arguments to study the *online* evolution of the validation error. The periodic-contexts assumption that we have made (Definition 2.1.4) is for relative simplicity in this proof structure.

Theorem 2.5.2 shows that the efficient algorithm ValidationOverAdaHedge($D$) obtains optimal (up-to polynomial factors in $(d, D)$) regret rates as would be achieved by an algorithm that had oracle knowledge about the presence of stochasticity *and* the model order. This is the strongest possible side information that an algorithm could conceivably possess keeping the online learning problem non-trivial.

Finally, it is worth noting that these positive model selection results for validation are still in a standard complexity-hierarchy setup, and in fact we crucially rely on the evolution of over-fitting error to rule out over-fitting models. It is a fascinating question to ask whether online validation ideas will also be effective in modern ML setups where prediction error can *decrease* as model complexity increases: both in over-parameterized neural networks and the "double descent" models on random features [124, 125]. If the answer turned out to be positive, adaptive online validation would be an extremely attractive "one-size-fits-all" solution to online model selection in a variety of statistical learning environments.

Figure 2.3 displays the benefits of online model selection, whether through SRM or validation, on our 1-memory Markov chain example in the context of binary sequence prediction.

## 2.6 Proofs

In this section, we collect the proofs of Theorems 2.4.3 and 2.5.2. The proofs, while diverging significantly in model selection methodology, follow a common high-level sequence of steps.

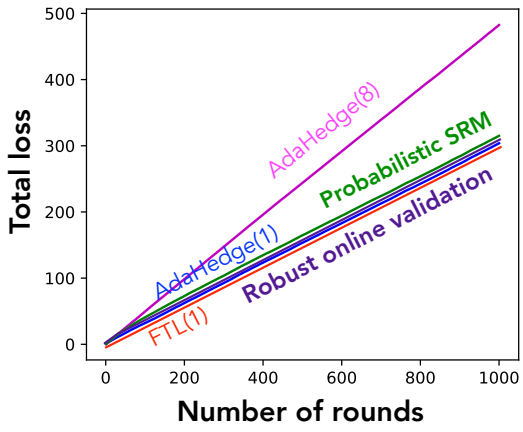



Figure 2.3: Approximate optimality of model selection algorithms using SRM and validation on 1-memory Markov chain. Figure from [88].

| Notation | Meaning |
|---|---|
| $\mathcal{F}_h$ | Set of all $h^{th}$-order tree experts $\mathbf{f} : \mathcal{X}^h \to \mathcal{X}$ |
| $l_{t,\mathbf{f}}^{(\text{tree})}$ | Instantaneous loss suffered by tree expert $\mathbf{f}$ |
| $L_{t,\mathbf{f}}^{(\text{tree})}$ | Cumulative loss suffered by tree expert $\mathbf{f}$ |
| $\mathbf{l}_{t,\mathbf{f}}^{(\text{tree})} = [l_{t,\mathbf{f}}^{(\text{tree})}]_{\mathbf{f}\in\mathcal{F}_D}$ | Instantaneous losses suffered by tree experts in $\mathcal{F}_D$ |
| $\mathbf{L}_{t,\mathbf{f}}^{(\text{tree})} = [L_{t,\mathbf{f}}^{(\text{tree})}]_{\mathbf{f}\in\mathcal{F}_D}$ | Cumulative losses suffered by tree experts in $\mathcal{F}_D$ |
| $l_{t,y}$ | Instantaneous loss from predicting $y \in \mathcal{X}$ |
| $L_{x,t,y}$ | Cumulative loss from predicting $y \in \mathcal{X}$ after seeing $x \in \mathcal{X}^h$ |
| $\widehat{F}_h(t) := \arg\min_{\mathbf{f}\in\mathcal{F}_h} L_{t,\mathbf{f}}$ | Best $h^{th}$-order tree expert at time $t$ |
| $\widehat{L}_{t,h} := L_{t,\widehat{F}_h(t)}$ | Cumulative loss suffered by tree expert $\widehat{F}_h(t)$ |
| $R_{T,h}$ | Regret with respect to best $h^{th}$-order tree expert |

Table 2.1: Basic notation for regret minimization under contextual experts framework.

We begin by collecting basic notation that is common to the framework, and then provide the sequence of steps for both proofs in parallel.

## Basic notation

Tables 2.1 and 2.2 contain basic notation for the regret minimization framework and algorithm-specific notation respectively for both algorithms, SRMOVERADAHEDGE($D$) and VALIDATIONOVERADAHEDGE($D$).

Tables 2.1 and 2.2 recap the basic notation for regret minimization and important algorithmic notation, and are useful to look at while reading the proof of the second-order bound.

| Notation | Meaning |
|---|---|
| $\eta_1^T = \{\eta_t\}_{t=1}^T$ | Sequence of learning-rates |
| $g : \{0, 1, \ldots, D\} \to \mathbb{R}_+$ | Prior function on order of tree expert |
| $\mathbf{w}_1(g) \Big\vert \mathbf{w}_1^{(\mathsf{tree})}(g)$ | Initial distribution on prediction \| choice of tree expert |
| $\mathbf{w}_t(\eta_t; g)$ | Distribution at round $t$ on prediction |
| $\mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g)$ | Distribution at round $t$ on choice of tree expert |
| $Z(g)$ | Normalizing factor for initial distribution on tree experts |
| $h_t(\eta_t; g)$ | Instantaneous expected loss incurred by algorithm at time $t$ |
| $H_t(\eta_1^t; g)$ | Cumulative expected loss incurred by algorithm at time $t$ |
| $\delta_t(\eta_t; g)$ | Instantaneous mixability gap of algorithm at time $t$ |
| $\Delta_t(\eta_1^t; g)$ | Cumulative mixability gap of algorithm at time $t$ |
| $v_t(\eta_t; g)$ | Instantaneous variance of loss incurred by algorithm at time $t$ |
| $V_t(\eta_1^t; g)$ | Cumulative variance of loss incurred by algorithm at time $t$ |

Table 2.2: Notation specific to algorithm SRMOverAdaHedge.

| Notation | Meaning |
|---|---|
| $\eta_1^T = \{\eta_t\}_{t=1}^T$ | Sequence of learning-rates used in meta-algorithm update |
| $(\eta^{(h)})_1^T = \{\eta_t^{(h)}\}_{t=1}^T$ | Sequence of learning-rates used in base algorithm update |
| $\mathbf{w}_t(\eta_t)$ | Distribution of meta-algorithm at round $t$ on prediction |
| $\mathbf{w}_t^{(\mathsf{tree})}(\eta_t)$ | Distribution of meta-algorithm at round $t$ on tree experts |
| $\mathbf{w}_t^{(h)}(\eta_t^{(h)})$ | Distribution of base algorithm $\mathcal{A}_h$ at round $t$ on prediction |
| $(\mathbf{w}_t^{(h)})^{(\mathsf{tree})}(\eta_t^{(h)})$ | Distribution of base algorithm $\mathcal{A}_h$ at round $t$ on tree experts |
| $h_t(\eta_t)$ | Instantaneous expected loss by meta-algorithm at time $t$ |
| $H_t(\eta_1^t)$ | Cumulative expected loss by meta-algorithm at time $t$ |
| $h_t(h; \eta_t^{(h)})$ | Instantaneous expected loss by base algorithm $\mathcal{A}_h$ at time $t$ |
| $H_t(h; (\eta_t^{(h)})_1^t)$ | Cumulative expected loss by base algorithm $\mathcal{A}_h$ at time $t$ |
| $\delta_t(\eta_t)$ | Instantaneous mixability gap of meta-algorithm at time $t$ |
| $\Delta_t(\eta_1^t) := \Delta_t$ | Cumulative mixability gap of meta-algorithm at time $t$ |
| $\delta_t^{(h)}(\eta_t^{(h)})$ | Instantaneous mixability gap of base algorithm $\mathcal{A}_h$ at time $t$ |
| $\Delta_t^{(h)}((\eta^{(h)})_1^t) := \Delta_t^{(h)}$ | Cumulative mixability gap of base algorithm $\mathcal{A}_h$ at time $t$ |
| $v_t(\eta_t)$ | Instantaneous variance of loss by meta-algorithm at time $t$ |
| $V_t(\eta_1^t)$ | Cumulative variance of loss by meta-algorithm at time $t$ |
| $v_t^{(h)}(\eta_t^{(h)})$ | Instantaneous variance of loss by base algorithm $\mathcal{A}_h$ at time $t$ |
| $V_t^{(h)}((\eta_t^{(h)})_1^t)$ | Cumulative variance of loss by base algorithm $\mathcal{A}_h$ at time $t$ |

Table 2.3: Notation specific to algorithm ValidationOverAdaHedge($D$).

## Adversarial model selection

The essence of adversarial model selection for both SRMOVERADAHEDGE($D$) and VALIDATIONOVERADAHEDGE($D$) involves getting what is known as a *second-order regret bound* in terms of the cumulative variance of the algorithm. These second-order regret bounds have a long history in adaptive online learning [94, 95, 98, 100], and are relatively straightforward to prove for both our algorithms. We includes the proofs here for completeness.

### Second-order bound for SRMOVERADAHEDGE($D$)

We first obtain our second-order-regret bound, stated generally for a prior function $g : \{0, 1, \ldots, D\} \to \mathbb{R}$. Tables 2.1 and 2.2 recap the basic notation for regret minimization and important algorithmic notation, and are useful to look at while reading the proof of the second-order bound.

Recall the expression for the computationally naive update in Equation (2.18):

$$w_{t,\mathbf{f}}^{(\mathsf{tree})}(\eta_t; g) = \frac{\left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}{\sum_{\mathbf{f}\in\mathcal{F}_D}\left(\sum_{h=\mathrm{order}(\mathbf{f})}^{D} g(h)\right) e^{-\eta_t L_{t,\mathbf{f}}}}.$$

and the expression for the initial distribution on tree experts based on Definition 2.4.1:

$$w_{1,\mathbf{f}}^{(\mathsf{tree})}(g) = \frac{\sum_{h=\mathrm{order}(f)}^{D} g(h)}{Z(g)}$$

where $Z(g) > 0$ is the initial normalizing factor. The explicit expression for the normalizing factor is $Z(g) = \sum_{h=0}^{D} 2^{2^h} g(h)$.

**Lemma 2.6.1.** SRMOVERADAHEDGE($D$) *with prior function* $g(\cdot)$ *obtains regret*

$$R_{T,d} \leq \left(\sqrt{V_T \ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2}\right)$$

*for every* $d \in \{0, 1, \ldots, D\}$.

*Proof.* Recall that $\widehat{F}_d(T)$ denotes the best $d^{th}$-order tree expert at round $T$ for the given loss sequence. We denote $\widehat{L}_{T,d} := L_{t,\widehat{F}_d(t)}$ as the actual loss incurred by this expert. We start with the computationally naive update in probability distribution over tree experts as in Equation (2.18), and the proof proceeds in a very similar manner to the variance-based

regret bound for vanilla AdaHedge [98]. We denote

$$\begin{aligned}
h_t(\eta_t; g) &:= \langle \mathbf{w}_t(\eta_t; g),\, \mathbf{l}_t \rangle = \langle \mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g),\, \mathbf{l}_t^{(\mathsf{tree})} \rangle \\
H_T(\eta_1^T; g) &:= \sum_{t=1}^{T} h_t(\eta_t; g) \\
m_t(\eta_t; g) &:= \frac{1}{\eta_t} \ln \langle \mathbf{w}_t(\eta_t; g),\, e^{-\eta_t \mathbf{l}_t} \rangle = \frac{1}{\eta_t} \ln \langle \mathbf{w}_t^{(\mathsf{tree})}(\eta_t; g),\, e^{-\eta_t \mathbf{l}_t^{(\mathsf{tree})}} \rangle \\
M_T(\eta_1^T; g) &:= \sum_{t=1}^{T} m_t(\eta_t; g).
\end{aligned}$$

Recall that the mixability gap $\delta_t(\eta_t; g) = h_t(\eta_t; g) - m_t(\eta_t; g)$ and $\Delta_T(\eta_1^T; g) = \sum_{t=1}^{T} \delta_t(\eta_t; g)$. Since the instantaneous losses are bounded between 0 and 1, it is easy to show that
$0 \leq \delta_t(\eta_t; g) \leq 1$.

A standard argument tells us that

$$\begin{aligned}
R_{T,d} &= H_T(\eta_1^T; g) - L_{T,d}^* \\
&= H_T(\eta_1^T; g) - M_T(\eta_1^T; g) + M_T(\eta_1^T; g) - L_{T,d}^* \\
&= M_T(\eta_1^T; g) - L_{T,d}^* + \Delta_T(\eta_1^T; g).
\end{aligned}$$

Recall that the sequence $\eta_1^T$ is decreasing as an automatic consequence of the update in Equation (2.19), and non-negativity of $\delta_t$. Handling a time-varying, data-dependent learning rate is well known to be challenging [95, 98]. We invoke a simple lemma from the original proof of AdaHedge [98] that helps us effectively substitute the final learning rate.

**Lemma 2.6.2** ([98])**.** *For any exponential-weights update with a decreasing learning rate $\eta_1^T$ and prior function $g(\cdot)$, we have $M_T(\eta_1^T; g) \leq M_T(\{\eta_T\}_{t=1}^T; g)$.*

Thus, we get

$$R_{T,d} \leq M_T(\{\eta_T\}_{t=1}^T; g) - L_{T,d}^* + \Delta_T(\eta_1^T; g). \tag{2.30}$$

We also have the following simple intermediate result for $M_T(\{\eta_T\}_{t=1}^T; g)$, which is simply a slightly more general version of the lemma in [98] that can apply to non-uniform priors.

**Lemma 2.6.3.**

$$M_T(\{\eta_T\}_{t=1}^T; g) \leq L_{T,d}^* + \frac{1}{\eta_T} \ln \left( \frac{Z(g)}{g(d)} \right).$$

*Proof.* We note that

$$\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}} \rangle \geq w_{1, f_{T,d}^*}^{(\mathsf{tree})}(g) e^{-\eta_T L_{T,d}^*}.$$

Because the initial distribution $\mathbf{w}_1^{(\mathsf{tree})}$ is normalized to sum to 1, a simple telescoping argument gives $M_T(\{\eta_T\}_{t=1}^T; g) = \sum_{t=1}^T m_t(\{\eta_T\}_{t=1}^T; g) = -\frac{1}{\eta_T} \ln\left(\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}} \rangle\right)$.

This automatically tells us that

$$\begin{aligned}
M_T(\{\eta_T\}_{t=1}^T; g) &= -\frac{1}{\eta_T} \ln\left(\langle \mathbf{w}_1^{(\mathsf{tree})}(g),\, e^{-\eta_T \mathbf{L}_T^{(\mathsf{tree})}} \rangle\right) \\
&\leq -\frac{1}{\eta_T} \ln(w_{1,f^*_{T,d}}^{(\mathsf{tree})}(g)) + L^*_{T,d} \\
&= L^*_{T,d} + \frac{1}{\eta_T} \ln\left(\frac{1}{w_{1,f^*_{T,d}}^{(\mathsf{tree})}(g)}\right) \\
&= L^*_{T,d} + \frac{1}{\eta_T} \ln\left(\frac{Z(g)}{\sum_{h=d}^D g(h)}\right) \\
&\leq L^*_{T,d} + \frac{1}{\eta_T} \ln\left(\frac{Z(g)}{g(d)}\right)
\end{aligned}$$

thus proving the lemma. □

Now, Equation (2.30) and Lemma 2.6.3 together with the definition of $\eta_t$ in Equation (2.19) give us

$$\begin{aligned}
R_{T,d} &\leq \frac{1}{\eta_T} \ln\left(\frac{Z(g)}{g(d)}\right) + \Delta_T(\eta_1^T; g) \\
&= \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2} \Delta_{T-1}(\eta_1^{T-1}; g) + \Delta_T(\eta_1^T; g).
\end{aligned}$$

From non-negativity of $\delta_t$, we have $\Delta_{T-1}(\eta_1^T; g) \leq \Delta_T(\eta_1^T; g)$ and so

$$R_{T,d} \leq \Delta_T(\eta_1^T; g)(1 + \frac{\ln\left(\frac{Z(g)}{g(d)}\right)}{\ln 2}). \tag{2.31}$$

It now remains to bound the quantity $\Delta_T$ in terms of variance. In fact, it will be useful to define slightly more generic quantities

$$\begin{aligned}
\Delta_{T_0}^T(\eta_{T_0}^T; g) &:= \sum_{t=T_0}^T \delta_t(\eta_t; g) \\
V_{T_0}^T(\eta_{T_0}^T; g) &:= \sum_{t=T_0}^T v_t(\eta_t; g) \text{ where} \\
v_t(\eta_t; g) &:= \mathrm{var}_{K_t \sim \mathbf{w}_t(\eta_t; g)} \left[l_{t,K_t}\right].
\end{aligned}$$

The bound is described below.

**Lemma 2.6.4.** *We have*

$$\Delta_{T_0}^T(\eta_{T_0}^T; g) \leq \sqrt{V_{T_0}^T(\eta_{T_0}^T; g)\ln 2} + \left(\frac{2}{3}\ln 2 + 1\right).$$

*Proof.* The argument is similar to the original AdaHedge proof [98] and proceeds below. We use a telescoping sum to get

$$\begin{aligned}
\Big(\Delta_{T_0}^T\big(\eta_{T_0}^T; g\big)\Big)^2 &= \sum_{t=T_0+1}^{T} \Big(\Delta_{T_0}^t\big(\eta_{T_0}^t; g\big)\Big)^2 - \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1}; g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1}; g\big) + \delta_t\big(\eta_t; g\big)\Big)^2 - \Big(\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1}; g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t; g\big)\Delta_{T_0}^{t-1}\big(\eta_{T_0}^{t-1}; g\big) + \Big(\delta_t\big(\eta_t; g\big)\Big)^2 \\
&\leq \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t; g\big)\Delta_{t-1}\big(\eta_1^{t-1}; g\big) + \Big(\delta_t\big(\eta_t; g\big)\Big)^2 \\
&= \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t; g\big)\frac{\ln 2}{\eta_t} + \Big(\delta_t\big(\eta_t; g\big)\Big)^2) \\
&\leq \sum_{t=T_0}^{T} 2\delta_t\big(\eta_t; g\big)\frac{\ln 2}{\eta_t} + \delta_t\big(\eta_t; g\big) \text{ since } \delta_t(\eta_t; g) \leq 1 \\
&\leq (2\ln 2)\sum_{t=T_0}^{T} \frac{\delta_t(\eta_t; g)}{\eta_t} + \Delta_{T_0}^T\big(\eta_{T_0}^T; g\big).
\end{aligned}$$

We also recall the following lemma from the original proof of AdaHedge [98]. The proof of this lemma involves a Bernstein tail bounding argument.

**Lemma 2.6.5** ([98])**.** *We have*

$$\frac{\delta_t\big(\eta_t; g\big)}{\eta_t} \leq \frac{1}{2}v_t\big(\eta_t; g\big) + \frac{1}{3}\delta_t\big(\eta_t; g\big).$$

Using Lemma 2.6.5, we then get

$$\Big(\Delta_{T_0}^T\big(\eta_{T_0}^T; g\big)\Big)^2 \leq V_{T_0}^T\big(\eta_{T_0}^T; g\big)\ln 2 + \left(\frac{2}{3}\ln 2 + 1\right)\Delta_{T_0}^T\big(\eta_{T_0}^T; g\big) \tag{2.32}$$

which is an inequality for the quantity $\Delta_{T_0}^T\big(\eta_{T_0}^T; g\big)$ in quadratic form. We now solve Equation (2.32), and use Fact 2.7.11 from Appendix 2.7 to get

$$\Delta_{T_0}^T\big(\eta_{T_0}^T; g\big) \leq \sqrt{V_{T_0}^T\big(\eta_{T_0}^T; g\big)\ln 2} + \frac{2}{3}\ln 2 + 1. \tag{2.33}$$

□

Now we complete the proof of Lemma 2.6.1 by combining Equations (2.31) and (2.33) for the special case of $T_0 = 1$. □

Now, noting that $V_T(\eta_1^T; g) \leq \frac{T}{4}$ and substituting the expression for $g = g_{\mathsf{prop}}$ from Equation (2.23) directly proves Equation (2.24) from Lemma 2.6.1. To see this, we substitute $g = g_{\mathsf{prop}}$ into the statement of Lemma 2.6.1 to get

$$\begin{aligned}
R_{T,d} &\leq \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z(g_{\mathsf{prop}})}{g_{\mathsf{prop}}(d)}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{\sum_{h=0}^{D} 2^{2^h} 2^{-2^{h+1}}}{2^{-2^{d+1}}}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{\sum_{h=0}^{D} 2^{-2^h}}{2^{-2^{d+1}}}\right)}{\ln 2}\right) \\
&\leq \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(2 \cdot 2^{2^{d+1}}\right)}{\ln 2}\right) \\
&= \left(\sqrt{V_T(\eta_1^T; g)\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(2 + 2^{d+1}\right) \\
&\leq \left(\frac{1}{2}\sqrt{T\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(2 + 2^{d+1}\right)
\end{aligned}$$

which is precisely Equation (2.24) when expressed in big-$\mathcal{O}$ notation.

**Second-order bound for ValidationOverAdaHedge($D$)**

**Lemma 2.6.6.** ValidationOverAdaHedge($D$) *gives second-order regret bound*

$$R_{T,d} = \mathcal{O}\left(\sqrt{V_T \ln 2} \cdot \ln D + \sqrt{V_T^{(d)} \cdot 2^d \ln 2} + \ln D + 2^d \ln 2\right).$$

*for every* $d \in \{0, 1, \ldots, D\}$.

*Proof.* The principal ingredient in this proof is essentially a chaining argument[8]: we observe that

$$\begin{aligned}
R_{T,d} &= H_T(\eta_1^T) - \widehat{L}_{t,d} \\
&= H_T(\eta_1^T) - H_T(d; (\eta_t^{(h)})_1^T) + H_T(d; (\eta_t^{(h)})_1^T) - \widehat{L}_{t,d}.
\end{aligned}$$

[8]This step is spiritually similar to Theorem 9 of Hutter and Poland [93], who did this for a different algorithm FTPL and with learning rate schedules that were not data-dependent.

We start with the first term, which is effectively a bound on the regret of the meta-algorithm.

**Lemma 2.6.7.** *For every $d \in \{0, 1, \ldots, D\}$, we have*

$$H_T(\eta_1^T) - H_T(d; (\eta_t^{(h)})_1^T) \leq \Delta_T \left(1 + \ln D\right).$$

*Proof.* We denote

$$m_t(\eta_t) := -\frac{1}{\eta_t} \ln \left( \sum_{h=0}^{D} q_t(h; \eta_t) e^{-\eta_t h_t(h; \eta_t^{(h)})} \right)$$
$$M_t(\eta_1^t) := \sum_{s=1}^{t-1} m_t(\eta_t).$$

A standard argument tells us that

$$H_T(\eta_1^T) - H_T(d; \eta_1^T) = \underbrace{H_T(\eta_1^T) - M_T(\eta_1^T)}_{T_1} + \underbrace{M_T(\eta_1^T) - H_T(d; \eta_1^T)}_{T_2}.$$

We first bound the term $T_2$. Observing that $\eta_1^T$ is a decreasing sequence, a simple adaptation of Lemma 2 from [98] gives us $M_T(\{\eta_t\}_{t=1}^T) \leq M_T(\{\eta_T\}_{t=1}^T) := M_T(\eta_T)$ for shorthand. Then, since $\eta_T < \infty$, we can apply Lemma 1, part 2 from [98] to get

$$\begin{aligned} M_T(\eta_T) &= -\frac{1}{\eta_T} \ln \left( \sum_{h=0}^{D} q_1(h; \eta_T) e^{-\eta_t H_T(h; (\eta_t^{(h)})_{t=1}^T)} \right) \\ &= -\frac{1}{\eta_T} \ln \left( \frac{1}{D} \sum_{h=0}^{D} e^{-\eta_t H_T(h; (\eta_t^{(h)})_{t=1}^T)} \right), \end{aligned}$$

and then we note that

$$\begin{aligned} -\frac{1}{\eta_T} \ln \left( \frac{1}{D} \sum_{h=0}^{D} e^{-\eta_t H_T(h; (\eta_t^{(h)})_{t=1}^T)} \right) &\leq -\frac{1}{\eta_T} \ln \left( \frac{1}{D} e^{-\eta_t H_T(d; (\eta_t^{(d)})_{t=1}^T)} \right) \\ &= \frac{\ln D}{\eta_T} - H_T(d; (\eta_t^{(d)})_1^T) \\ \implies T_2 = M_T(\eta_1^T) - H_T(d; (\eta_t^{(d)})_1^T) &\leq \frac{\ln D}{\eta_T} = \Delta_{T-1} \cdot \ln D \leq \Delta_T \cdot \ln D. \end{aligned}$$

Next, we bound the term $T_1$. Noting that $f(\cdot) = e^{-\eta_t(\cdot)}$ is convex for all $\eta_t$, applying Jensen's inequality tells us that

$$\left\langle \mathbf{w}_t^{(h)}, e^{-\eta_t \mathbf{l}_t} \right\rangle \geq e^{-\eta_t \langle \mathbf{w}_t^{(h)}, \mathbf{l}_t \rangle} = e^{-\eta_t h_t(h; \eta_t^{(h)})}$$

and so, we have

$$\begin{aligned} m_t(\eta_t) &= -\frac{1}{\eta_t} \ln \left( \sum_{h=0}^{D} q_t(h;\eta_t) e^{-\eta_t h_t(h;\eta_t^{(h)})} \right) \\ &\geq -\frac{1}{\eta_t} \ln \left( \sum_{h=0}^{D} q_t(h;\eta_t) \langle \mathbf{w}_t^{(h)},\, e^{-\eta_t \mathbf{l}_t} \rangle \right) \\ &= -\frac{1}{\eta_t} \ln \left( \langle \mathbf{w}_t(\eta_t),\, e^{-\eta_t \mathbf{l}_t} \rangle \right) \\ &= h_t(\eta_t) - \delta_t(\eta_t) \end{aligned}$$

where the last equality follows from the definition of $\delta_t(\eta_t)$ in Equation (2.33). Thus, we have

$$\begin{aligned} T_1 = H_T(\eta_1^T) - M_T(\eta_1^T) &= \sum_{t=1}^{T} h_t(\eta_t) - m_t(\eta_t) \\ &\leq \sum_{t=1}^{T} \delta_t(\eta_t) \\ &= \Delta_T. \end{aligned}$$

Finally, we have

$$H_T(\eta_1^T) - H_T(d;\eta_1^T) = T_1 + T_2 = \Delta_T + \Delta_T \ln D = \Delta_T \left(1 + \ln D\right),$$

which completes the proof.

□

Next, we bound the second term, which is simply the regret of the base algorithm $\mathcal{A}_d$ under decreasing learning rate schedule $\{\eta_t^{(d)}\}_{t=1}^T$. This proof is a simple adaptation of the proof of AdaHedge($d$) to contextual experts.

**Lemma 2.6.8.** *For every $d \in \{0, 1, \ldots, D\}$, we have*

$$H_T(d;\eta_1^T) - \widehat{L}_{t,d} \leq 2\Delta_T^{(d)}.$$

*Proof.* For this proof, it will be convenient to work with the naive update as defined in Equation (2.18). We defined (instantaneous and cumulative) mix loss

$$m_t^{(d)}(\eta_t^{(d)}) := -\frac{1}{\eta_t^{(d)}} \ln \left( \langle \mathbf{w}_t^{(d)},\, e^{-\eta_t^{(d)} \mathbf{l}_t} \rangle \right)$$

$$M_t^{(d)}((\eta_t^{(d)})_1^t) := \sum_{s=1}^{t} m_t^{(d)}(\eta_t^{(d)}).$$

Then, we have

$$H_T(d;(\eta_t^{(d)})_1^T) - \widehat{L}_{t,d} = \underbrace{H_T(d;(\eta_t^{(d)})_1^T) - M_T^{(d)}((\eta^{(d)})_1^T)}_{T_1} + \underbrace{M_T^{(d)}((\eta_t^{(d)})_1^T) - \widehat{L}_{t,d}}_{T_2}.$$

We first bound the term $T_2$. Notice that $\eta_t^{(d)}$ is also a decreasing sequence, so we can apply Lemmas 1 and 2 from [98] to get

$$\begin{aligned}
M_T^{(d)}((\eta_t^{(d)})_1^T) &\leq M_T^{(d)}(\{\eta_T^{(d)}\}_{t=1}^T) \\
&= -\frac{1}{\eta_T^{(d)}} \ln\left(\langle \mathbf{w}_1^{(d)},\, e^{-\eta_T^{(d)}\mathbf{L}_T}\rangle\right) \\
&\leq -\frac{1}{\eta_T^{(d)}} \ln\left(\frac{1}{2^{2^d}} \cdot e^{-\eta_T^{(d)}\widehat{L}_{T,d}}\right) \\
&= \widehat{L}_{T,d} + \frac{2^d \ln 2}{\eta_T^{(d)}} \\
&= \widehat{L}_{T,d} + \Delta_{T-1}^{(d)}((\eta_t^{(d)})_1^{T-1}) \leq \widehat{L}_{T,d} + \Delta_T^{(d)}((\eta_t^{(d)})_1^T),
\end{aligned}$$

where the last line follows from the definition of $\eta_t^{(d)}$ in Equation (2.19).
Thus we have

$$T_2 = M_T^{(d)}((\eta^{(d)})_1^T) - \widehat{L}_{t,d} \leq \Delta_T^{(d)}.$$

Next, by definition we note that $T_1 = \Delta_T^{(d)}$. This completes the proof.

□

Assuming Lemmas 2.6.7 and 2.6.8, we have

$$R_{T,d} \leq \Delta_T(\eta_1^T)\,(1 + \ln D) + 2\Delta_T^{(d)}((\eta_t^{(d)})_1^T). \tag{2.34}$$

It remains to bound the quantities $\Delta_T$ and $\Delta_T^{(d)}$ in terms of variance. In fact, it will be useful to define slightly more generic quantities

$$\begin{aligned}
\Delta_{T_0}^T(\eta_{T_0}^T) &:= \sum_{t=T_0}^T \delta_t(\eta_t) \\
V_{T_0}^T(\eta_{T_0}^T) &:= \sum_{t=T_0}^T v_t(\eta_t) \text{ where} \\
v_t(\eta_t) &:= \mathrm{var}_{K_t \sim \mathbf{w}_t(\eta_t)}\left[l_{t,K_t}\right].
\end{aligned}$$

We describe these second-order bounds below.

**Lemma 2.6.9.** *We have*

$$\Delta_{T_0}^T(\eta_{T_0}^T) \le \sqrt{V_{T_0}^T(\eta_{T_0}^T)\ln 2} + \left(\frac{2}{3}\ln 2 + 1\right), \tag{2.35}$$

*and we have*

$$\Delta_T^{(d)} \le \sqrt{V_T^{(d)} \cdot 2^d \ln 2} + \left(\frac{4}{3}\cdot 2^d \ln 2 + 2\right). \tag{2.36}$$

*Proof.* Since it suffices to prove the statement of Equation (2.36) for the naive update, this statement follows as a special case of Theorem 6 in [98] with $K = 2^{2^d}$ experts; we refer the reader to that proof.

For the statement of Equation (2.35), the proof is *similar* to the argument in Theorem 6 as well (with different constants, however, so we reproduce it here). We use a telescoping sum to get

$$\begin{aligned}
\left(\Delta_{T_0}^T(\eta_{T_0}^T)\right)^2 &= \sum_{t=T_0+1}^T \left(\Delta_{T_0}^t(\eta_{T_0}^t)\right)^2 - \left(\Delta_{T_0}^{t-1}(\eta_{T_0}^{t-1})\right)^2 \\
&= \sum_{t=T_0}^T \left(\Delta_{T_0}^{t-1}(\eta_{T_0}^{t-1}) + \delta_t(\eta_t)\right)^2 - \left(\Delta_{T_0}^{t-1}(\eta_{T_0}^{t-1})\right)^2 \\
&= \sum_{t=T_0}^T 2\delta_t(\eta_t)\Delta_{T_0}^{t-1}(\eta_{T_0}^{t-1}) + \left(\delta_t(\eta_t)\right)^2 \\
&\le \sum_{t=T_0}^T 2\delta_t(\eta_t)\Delta_{t-1}(\eta_1^{t-1}) + \left(\delta_t(\eta_t)\right)^2 \\
&= \sum_{t=T_0}^T 2\delta_t(\eta_t)\frac{\ln 2}{\eta_t} + \left(\delta_t(\eta_t)\right)^2) \\
&\le \sum_{t=T_0}^T 2\delta_t(\eta_t)\frac{\ln 2}{\eta_t} + \delta_t(\eta_t) \text{ since } \delta_t(\eta_t) \le 1 \\
&\le (2\ln 2)\sum_{t=T_0}^T \frac{\delta_t(\eta_t)}{\eta_t} + \Delta_{T_0}^T(\eta_{T_0}^T).
\end{aligned}$$

We also recall the following lemma from the original proof of the AdaHedge learning rate choice $\eta_t = 1/\Delta_{t-1}$ [98]. The proof of this lemma involves a Bernstein tail bounding argument.

**Lemma 2.6.10** ([98])**.** *We have*

$$\frac{\delta_t(\eta_t)}{\eta_t} \le \frac{1}{2}v_t(\eta_t) + \frac{1}{3}\delta_t(\eta_t).$$

Using Lemma 2.6.10, we then get

$$\left(\Delta_{T_0}^T(\eta_{T_0}^T)\right)^2 \leq V_{T_0}^T(\eta_{T_0}^T)\ln 2 + \left(\frac{2}{3}\ln 2 + 1\right)\Delta_{T_0}^T(\eta_{T_0}^T) \tag{2.37}$$

which is an inequality for the quantity $\Delta_{T_0}^T(\eta_{T_0}^T)$ in quadratic form. We now solve Equation (2.37), and use Fact 2.7.11 from Appendix 2.7 to get

$$\Delta_{T_0}^T(\eta_{T_0}^T) \leq \sqrt{V_{T_0}^T(\eta_{T_0}^T)\ln 2} + \frac{2}{3}\ln 2 + 1. \tag{2.38}$$

□

Substituting Equations (2.35) and (2.36) into Equation (2.34), we get the second-order bound

$$R_{T,d} \leq \sqrt{V_T(\eta_1^T)\ln 2}\,(1+\ln D) + 2\sqrt{V_T^{(d)}((\eta_t^{(d)})_1^T)\cdot 2^d \ln 2} + \left(\frac{2}{3}\ln 2 + 1\right)(1+\ln D) \tag{2.39}$$
$$+ \frac{4}{3}\cdot 2^d \ln 2 + 2,$$

which completes the proof of Lemma 2.6.6 when expressed in big-$\mathcal{O}$ notation. Further, noting that $V_T(\eta_1^T) \leq \frac{T}{4}$ gives us Equation (2.27)when expressed in big-$\mathcal{O}$ notation.

□

## Stochastic model selection

We now provide the proofs for stochastic model selection for both algorithms: SRMOverAdaHedge($D$) and ValidationOverAdaHedge($D$). We begin by providing basic notation for stochastic contextual prediction that is common to both algorithms.

### Notation for contextual prediction

First, we define a couple of convenient counts for the number of appearances of a particular context, and the number of contexts that have so far appeared.

**Definition 2.6.11.** *The **appearance frequency** of a particular context $x(h) \in \mathcal{X}^h$ at time $t$ is given by*

$$N_t(x(h)) := \sum_{s=1}^{t-1} \mathbb{I}[X_s(h) = x(h)],$$

*The **fraction of times the value** $y \in \mathcal{X}$ seen after a particular context is given by*

$$\widehat{P}_t(y|x(h)) := \frac{\sum_{s=1}^{t-1}\mathbb{I}[X_s(h)=x(h), Y_s = y]}{\sum_{s=h}^{t-1}\mathbb{I}[X_s(h)=x(h)]}$$
$$\left(= 1 - \frac{L_{x(h),t-1,y}}{N_t(x(h))}\right)$$

| Notation | Meaning/Interpretation |
|---|---|
| $N_t(x(h))$ | Appearance frequency of a sub-context $x(h) \in \mathcal{X}^h$ |
| $\widehat{P}_t(h\|x(h))$ | Fraction of times that we observed $X_t(h) = x(h), Y_t = y$ |
| $S_{t,h}$ | Number-of-seen sub-contexts of length $h$ at time $t$ |
| $\widehat{\pi}_h(t)$ | Estimated unpredictability based on $h^{th}$-order tree expert predictors |
| $D_t(h)$ | Gap between correct and incorrect predictors at time $t$ |
| $\mathbf{w}_t^{(h)}$ | Probability distribution on predictions |
| $q_t(h) \propto Q_t(h)$ | Posterior probability that the $h^{th}$-order model is the right model |
| $d$ | True model order of data $(X_t, Y_t)_{t=1}^T$ |
| $Q_h^*(\cdot), h \leq d$ | Marginal distribution on $X_t(h), h \leq d$ |
| $P^*(\cdot\|x(h))$ | Conditional distribution on $Y_t$ given $X_t = x(h)$ |
| $\beta(x(d)), \beta^*$ | Average prediction accuracy with context $x(d)$ |
| $\pi_h^*, h \leq D$ | Asymptotic unpredictability under $h^{th}$-order model. |

Table 2.4: Notation for analysis.

*The **number-of-seen-contexts** is given by*

$$S_{t,h} := \sum_{x(h) \in \mathcal{X}^h} \mathbb{I}[N_t(x(h)) > 0].$$

Next, we define our estimates for unpredictability, effectively an estimate for the approximation error, under various model orders.

**Definition 2.6.12** ([110])**.** *For every value of $h \geq 0$ and a sequence $\{(X_t, Y_t)\}_{t \geq 1}$, we define its estimated unpredictability*

$$\begin{aligned}\widehat{\pi}_h(t) &:= \sum_{x(h) \in \mathcal{X}^h} \frac{N_t(x(h))}{t} \left(1 - \max_{y \in \mathcal{X}} \{\widehat{P}_t(y|x(h))\}\right) \\ &= \sum_{x(h) \in \mathcal{X}^h} \frac{1}{t} \min_{y \in \mathcal{X}} \{L_{x(h),t,y}\}.\end{aligned}$$

This definition is inspired by the information-theoretic perspective on universal sequence prediction [110]. In this line of work, the quantity $\widehat{\pi}_h(t)$ represents the estimated unpredictability of a binary sequence under a $h$-memory Markov model. We will see that this is the natural estimate of *approximation error of the $h^{th}$-order model* that is used to carry out data-driven model selection under the $d^{th}$-order stochastic condition on responses (Definition 2.1.1), and all three generative assumptions we have made on contextual information (Definitions 2.1.2, 2.1.3 and 2.1.4).

Finally, we denote the *true prediction* (the one we would make if we had oracle knowledge of the best predictor $f^*(\cdot)$) as

$$Y_t^* := f^*(X_t(d)).$$

Then, for every $h \geq d$ we define

$$D_t(h) := L_{X_t(h),t,1-Y_t^*} - L_{X_t(h),t,Y_t^*} \tag{2.40}$$

represents the "gap" between the correct predictor $Y_t^*$ and the worse predictor $1 - Y_t^*$ at time $t$, and pertaining to the current context $X_t(h)$.

## Probabilistic model selection

As hinted in the algorithm design, both SRMOVERADAHEDGE($D$) and VALIDATIONOVERADAHEDGE($D$) can be interpreted as explicitly performing *probabilistic* model selection using the principles of SRM and validation respectively. We now show this explicitly, starting with SRMOVERADAHEDGE($D$).

### Probabilistic model selection for SRMOVERADAHEDGE($D$)

To effectively bound regret for the "easier" stochastic instances, we need finer control on the cumulative mixability gap term $\Delta_T(\eta_1^T; g)$. Our starting point is the following thresholding lemma.

**Lemma 2.6.13.** *Fix $t_0 > 0$. Let $T_0 := \max\{0 < t \leq T : \eta_t > \frac{\ln 2}{t_0}\}$. Then, we have*

$$\Delta_T(\eta_1^T; g) \leq t_0 + 1 + \sqrt{V_{T_0}^T(\eta_{T_0}^T; g) \ln 2} + \frac{2}{3}\ln 2 + 1. \tag{2.41}$$

*Proof.* From the definition of $T_0$, we observe that

$$\begin{aligned} \eta_{T_0} = \frac{\ln 2}{\Delta_{T_0-1}(\eta_1^{T_0-1}; g)} &> \frac{\ln 2}{t_0} \\ \implies \Delta_{T_0-1}(\eta_1^{T_0-1}; g) &< t_0 \\ \implies \Delta_{T_0}(\eta_1^{T_0}; g) &< t_0 + 1. \end{aligned}$$

Then, using $\Delta_T(\eta_1^T; g) = \Delta_{T_0}(\eta_1^{T_0}; g) + \Delta_{T_0}^T(\eta_{T_0}^T; g)$ and Lemma 2.6.4 directly gives us the statement in Equation (2.41) and completes the proof. □

We observe that the threshold $T_0$ depends on the choice of $t_0$ as well as the data (in fact, it is a random variable when the process $\{(X_t, Y_t)\}_{t=1}^T$ is stochastic). We have the freedom to choose $t_0 > 0$ for our analysis. Conceptually, in the stochastic regime, the choice of $t_0$ thresholds the number of rounds $T_0$ below which we can make few, if any, statistical guarantees, and will become clear in subsequent sections. Effectively, Lemma 2.6.13 uses the elegant inverse relationship between learning rate and mixability (in Equation (2.19)) to show that a minimal amount of regret, precisely, in terms of $t_0$, is accumulated *even before we can make high-probability statistical guarantees*.

Now, we have stated the problem of wanting to exploit the structure of a $d^{th}$-order stochastic sequence $\{(X_t, Y_t)\}_{t\geq 1}$ in an online fashion, as a model selection problem. This has been implicitly clear in the choice of prior function in Equation (2.23): more complex experts are down-weighted. Now, we make the connection clear.

As a reminder, we evaluate the performance of the algorithm SRMOVERADAHEDGE($D$) with prior function $g_{\mathsf{prop}}(\cdot)$), and using Equation (2.33) as a jumping point, we are concerned with bounding the cumulative variance $V_{T_0}^T(\eta_{T_0}^T; g)$.

First, we observe that

$$\begin{aligned} V_{T_0}^T(\eta_{T_0}^T; g_{\mathsf{prop}}) &= \sum_{t=T_0}^{T} v_t(\eta_t; g_{\mathsf{prop}}) \\ &= \sum_{t=T_0}^{T} w_{t,Y_t^*}(\eta_t; g_{\mathsf{prop}}) \left(1 - w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})\right) \text{ since } l_{t,K_t} \text{ i.i.d } \sim \mathrm{Ber}(w_{t,1}) \\ &\leq \sum_{t=T_0}^{T} w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}}) \end{aligned}$$

and thus, it is sufficient to control the evolution of the term $w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})$ with $t$. This is the probability with which we select the prediction $1 - Y_t^*$ that is more likely to be wrong under the stochastic model for the data.

The first step is to express the update in this probability in terms of a posterior probability on the effective *order of the model* the algorithm is selecting. Explicitly, we can re-write Equation (2.22a) as

$$w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}}) = \sum_{h=0}^{D} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}^{(h)}(\eta_t)$$

where we have defined the shorthand notation for the update used by SRMOVERADAHEDGE($h$) with uniform prior,

$$w_{t,1-Y_t^*}^{(h)}(\eta_t) := w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{unif}}) = \frac{e^{-\eta_t D_t(h)}}{1 + e^{-\eta_t D_t(h)}},$$

where $D_t(h)$ is according to Equation (2.40) and the quantities $\{q_t(h; \eta_t, g_{\mathsf{prop}})\}$ are explicitly written as

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \propto Q_t(h; \eta_t, g_{\mathsf{prop}}) := g_{\mathsf{prop}}(h) \prod_{x(h)\in\mathcal{X}^h} \left( \sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \tag{2.42}$$

where the proportionality constant is set such that $\sum_{h'=0}^{D} q_t(h; \eta_t, g_{\mathsf{prop}}) = 1$. The quantity $q_t(h; \eta_t, g_{\mathsf{prop}})$ is exactly the *posterior probability* that SRMOVERADAHEDGE($D$) selects a

$h^{th}$-order model. We will see that controlling the posterior on model order selection is crucial to bounding the variance in our desired manner.

First, we state a simple lemma that bounds Equation (2.42) in terms of more intuitive quantities.

**Lemma 2.6.14.** *We have*

$$\exp\{-\eta_t \widehat{\pi}_h(t)t + \ln g_{\mathsf{prop}}(h)\} \leq Q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \exp\{-\eta_t \widehat{\pi}_h(t)t + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h)\}. \tag{2.43}$$

*Proof.* For the upper bound, we have

$$\begin{aligned}
Q_t(h; \eta_t, g_{\mathsf{prop}}) &:= g_{\mathsf{prop}}(h) \prod_{x(h)\in\mathcal{X}^h} \left( \sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) \\
&= \exp\left\{ \sum_{x(h)\in\mathcal{X}^h} \ln \left( \sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&\leq \exp\left\{ \sum_{x(h)\in\mathcal{X}^h} \ln \left( 2e^{-\eta_t \min_{y\in\mathcal{X}}\{L_{x(h),t,y}\}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ - \sum_{x(h)\in\mathcal{X}^h} \eta_t \min_{y\in\mathcal{X}}\{L_{x(h),t,y}\} + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ -\eta_t \widehat{\pi}_h(t)t + 2^h \ln 2 + \ln g_{\mathsf{prop}}(h) \right\}
\end{aligned}$$

and for the lower bound, we have

$$\begin{aligned}
Q_t(h; \eta_t, g_{\mathsf{prop}}) &:= \exp\left\{ \sum_{x(h)\in\mathcal{X}^h} \ln \left( \sum_{y\in\mathcal{X}} e^{-\eta_t L_{x(h),t,y}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&\geq \exp\left\{ \sum_{x(h)\in\mathcal{X}^h} \ln \left( e^{-\eta_t \min_{y\in\mathcal{X}}\{L_{x(h),t,y}\}} \right) + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ - \sum_{x(h)\in\mathcal{X}^h} \eta_t \min_{y\in\mathcal{X}}\{L_{x(h),t,y}\} + \ln g_{\mathsf{prop}}(h) \right\} \\
&= \exp\left\{ -\eta_t \widehat{\pi}_h(t)t + \ln g_{\mathsf{prop}}(h) \right\}
\end{aligned}$$

□

Substituting $\ln g_{\mathsf{prop}}(h) = -2^{h+1} \ln 2 = -2 \cdot 2^h \ln 2$, we get

$$\exp\{-\eta_t \widehat{\pi}_h(t)t - 2 \cdot 2^h \ln 2\} \leq Q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \exp\{-\eta_t \widehat{\pi}_h(t)t - 2^h \ln 2\}. \tag{2.44}$$

Equation (2.44) effectively makes the trade-off between approximation error (reflected by the quantity $\widehat{\pi}_h(t)$) and model complexity (reflected by the quantity $2^h \ln 2$ clear in the model-order selection problem. We can think of the model orders as "meta-experts" that are being randomized over. Note that the learning rate that is being used to randomize their selection is still $\eta_t$!

**Probabilistic model selection for ValidationOverAdaHedge($D$)**

We now express the analysis of the cumulative variance of the algorithm ValidationOverAdaHedge($D$) as a stochastic model selection problem.

First, we observe that

$$\begin{aligned} V_T(\eta_1^T) &= \sum_{t=1}^{T} v_t(\eta_t) \\ &= \sum_{t=T_0}^{T} w_{t,Y_t^*}(\eta_t; g_{\mathsf{prop}}) \left(1 - w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})\right) \text{ since } l_{t,K_t} \text{ i.i.d } \sim \mathrm{Ber}(w_{t,1}) \\ &\leq \sum_{t=T_0}^{T} w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}}) \end{aligned}$$

and thus, it is sufficient to control the evolution of the term $w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}})$ with $t$. This is the probability with which we select the prediction $1 - Y_t^*$ that is more likely to be wrong under the stochastic model for the data.

The first step is to express the update in this probability in terms of a posterior probability on the effective *order of the model* the algorithm is selecting. Recall from Equation (2.22a) that

$$w_{t,1-Y_t^*}(\eta_t; g_{\mathsf{prop}}) = \sum_{h=0}^{D} q_t(h; \eta_t) w_{t,1-Y_t^*}^{(h)}(\eta_t^{(h)}),$$

where recall that $w_{t,1-Y_t^*}^{(h)}(\eta_t^{(h)}$ was defined as the weight vector used by base algorithm $\mathcal{A}_h$, and the "model selecting" probabilities are defined as

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \propto Q_t(h; \eta_t) := e^{-\eta_t H_{t-1}(h)}, \tag{2.45}$$

where $H_{t-1}(h)$ represents the cumulative *loss* experienced by base algorithm $\mathcal{A}_h$ at time $(t-1)$. The quantity $q_t(h; \eta_t)$ is exactly the *posterior probability* that the algorithm ValidationOverAdaHedge($D$) selects a $h^{th}$-order model, and so this is a kind of *probabilistic* model selection. Clearly, controlling the posterior on model order selection is crucial to bounding the variance in our desired manner.

Note that we can decompose

$$H_t(h) = \widehat{L}_{t,h} + R_t^{(h)} = t \cdot \widehat{\pi}_h(t) + R_t^{(h)},$$

where recall that $R_t^{(h)}$ is the regret incurred by the base algorithm $\mathcal{A}_h$. Thus, our probabilistic model selection procedure will pick a base algorithm that minimizes the additive combination of approximation error (reflected by the quantity $\widehat{\pi}_h(t)$) and estimation error (reflected by the regret accumulated) in the model-order selection problem. More directly, this can be thought of as *online validation*.

## Analysis for a higher-than-needed model order

Before getting into the essence of model selection, we recap guarantees on regret with respect to a particular model order to illustrate the perils of picking an over-fitting model formally. These guarantees are slightly different for the SRM algorithm owing to the choice of learning rate (which is actually sub-optimal), and so we describe the guarantee separately for both algorithms.

### Analysis for a higher-than-needed model order for SRMOverAdaHedge($D$)

Here, we analyze the contribution of a specific selected model order to the variance, an important intermediate step. Formally, we consider the algorithm SRMOverAdaHedge($h$) equipped with the uniform prior function $g_{\mathsf{unif}}(h') = \mathbb{I}[h' = h]$. The regret guarantee is given by the following proposition.

**Proposition 2.6.15.** *1. For any sequence $\{X_t, Y_t\}_{t=1}^T$, SRMOverAdaHedge($h$) with uniform prior gives us regret rate*

$$R_{T,d} = \mathcal{O}\left(\sqrt{T} 2^h\right) \tag{2.46}$$

*with respect to the best $d^{th}$-order tree expert in hindsight, and for every $d \leq h$.*

*2. SRMOverAdaHedge($h$) with uniform prior gives regret with probability greater than $(1-\epsilon)$:*

$$R_{T,d} = \mathcal{O}\Big(\frac{2^{2h}}{(2\beta^* - 1)^2}\left(h + \ln\left(\frac{1}{\epsilon(2\beta^* - 1)}\right)\right)\Big).$$

*on a sequence $(X_t, Y_t)_{t \geq 1}$ that satisfies the $d^{th}$-order stochastic condition on responses (Definition 2.1.1) with parameter $\beta^*$.*

Observe the sub-optimal scaling in terms of $2^{2h}$ in the regret bound for the case where $d < h$. We now proceed to prove Proposition 2.6.15.

Formally, the algorithm SRMOVERADAHEDGE($h$) equipped with the uniform prior function $g_{\mathsf{unif}}(h') = \mathbb{I}[h' = h]$ gives us $q_t(h'; \eta_t, g_{\mathsf{unif}}) = \mathbb{I}[h' = h]$, and we would get

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h'=0}^{D} q_t(h';\eta_t,g_{\mathsf{unif}})w_{t,1-Y_t^*}^{(h)}(\eta_t) &= \sum_{t=1}^{T} w_{t,1-Y_t^*}^{(h)} \\ &= \sum_{t=1}^{T}\frac{e^{-\eta_t D_t(h)}}{1+e^{-\eta_t D_t(h)}} \\ &\le \sum_{t=1}^{T}\min\{e^{-\eta_t D_t(h)},1\} \\ &\le \sum_{t=1}^{T}\min\{e^{-\eta_T D_t(h)},1\}\end{aligned}$$

where $D_t(h)$ is the gap between predictions as in Equation (2.40), and the last inequality is because $\eta_1^T$ is a decreasing sequence according to the update in Equation (2.19).

Therefore, we have

$$V_T(\eta_1^T; g_{\mathsf{unif}}) \le \sum_{t=1}^{T}\min\{e^{-\eta_t D_t(h)},1\}. \tag{2.47}$$

We observe that Equation (2.47) can be effectively unraveled to get a closed-form variance bound for particular evolutions of $\{D_t(h)\}_{t\ge 1}$. Particularly, we care about $D_t(h)$ as a function of $N_t(X_t(h))$, the number of appearances so far of the current context. We show this result in the following lemma.

**Lemma 2.6.16.** *Let the following condition hold for some $t_0(h) > 0$ and $\alpha > 0$.*

$$D_t(h) \ge \alpha N_t(X_t(h)) \text{ for all } t \text{ such that } N_t(X_t(h)) \ge t_0(h) \tag{2.48}$$

*for some $\alpha > 0$.*
*Then, we have*

$$\sum_{t=1}^{\infty} w_{t,1-Y_t^*}^{(h)}(\eta_t) \le 2^h\left(t_0(h) + \frac{1}{\eta_T\alpha}\right). \tag{2.49}$$

*Proof.* We can directly use the condition in Equation (2.48). For values of $t$ such that $N_t(X_t(h)) < t_0(h)$, we apply $w_{t,1-Y_t^*}^{(h)}(\eta_t) \le 1$. Otherwise, we use $w_{t,1-Y_t^*}^{(h)}(\eta_t) \le e^{-\eta_T\alpha N_t(X_t(h))}$.

Combining the two gives us

$$\begin{aligned}\sum_{t=1}^{\infty} w_{t,1-Y_t^*}^{(h)}(\eta_t) &\leq \sum_{x(h)\in\mathcal{X}^h}\left(t_0 + \sum_{s=t_0(h)}^{N_T(x(h))} e^{-\eta_T\alpha s}\right)\\ &\leq 2^h t_0(h) + \sum_{x(h)\in\mathcal{X}^h}\sum_{s=t_0(h)}^{\infty} e^{-\eta_T\alpha s}\\ &\leq 2^h\left(t_0(h) + \sum_{s=t_0(h)}^{\infty} e^{-\eta_T\alpha s}\right)\\ &\leq 2^h\left(t_0(h) + \frac{e^{-\eta_T\alpha}}{1-e^{-\eta_T\alpha}}\right).\end{aligned}$$

Now, we have

$$\begin{aligned}\frac{e^{-\eta_T\alpha}}{1-e^{-\eta_T\alpha}} &= \frac{1}{e^{\eta_T\alpha}-1}\\ &\leq \frac{1}{\eta_T\alpha}\end{aligned}$$

by the inequality $e^a \geq 1+a$ for $a \geq 0$. Substituting this above gives us our required result. □

It remains to show that the condition in Equation (2.48) is met with high probability for $(X_t, Y_t)_{t\geq 1}$ satisfying the $d^{th}$-order *realizability condition* on responses (Definition 2.1.1) with parameter $\beta^*$, and for any $d \leq h$. We use a standard Hoeffding-bounding technique to show this.

**Lemma 2.6.17.** *Let $\epsilon \in (0,1]$. For a process $\{(X_t,Y_t)\}_{t\geq 1}$ satisfying the $d^{th}$-order realizability condition with parameter $\beta^* > 1/2$, the condition in Equation (2.48) holds for all $h \geq d$ for parameter values*

$$\alpha := \frac{(2\beta^*-1)}{2} \tag{2.50}$$

$$t_0(h) = t_{\mathsf{high}}(h) := \frac{2}{\alpha^2}\ln\left(\frac{4(D-d)\cdot 2^{h+1}}{\alpha^2\epsilon}\right) \tag{2.51}$$

*with probability greater than or equal to $(1-\epsilon/2)$.*

*Proof.* Essentially, we need to obtain to bound properties of the gap sequence $\{D_{t,(h)}\}_{t=1}^T$ so defined in Equation (2.40) – we use the Hoeffding bound for this. This proof is a simple adaptation of the proof in the original AdaHedge paper [95] to the case of contextual prediction.

We denote the $p^{th}$ epoch of arrival of context $x(h) \in \mathcal{X}^h$ by $T_p(x(h))$. Showing that the condition in Equation (2.48) holds with probability greater than or equal to $(1 - \epsilon/2)$ is exactly equivalent to showing that the probability of the following bad event

$$\left\{ \cup_{h=d}^{D} \cup_{\mathbf{x}_{(h)} \in \mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))} \left\{ D_{T_p(x(h))}(h) < \alpha p \right\} \right\} \tag{2.52}$$

is less than or equal to $\frac{\epsilon}{2}$. We proceed by showing exactly this.

From the definition of a $d^{th}$-order stochastic process, we have $Y_t | \{X_t, (X_s, Y_s)_{s=1}^{t-1}\}$ i.i.d $\sim P^*(\cdot | X_t(d))$. This means that $Y_t$ is independent of $(X_t(D, \ldots, D_d), X_s, Y_s)_{s=1}^{t-1}$ conditioned on $X_t(d)$, and we can write

$$D_{T_p(x(h))}(h) = \sum_{s'=1}^{p} 2Z_{s'}$$

where

$$\{Z'_s\}_{s' \geq 1} \text{ i.i.d } \sim \begin{cases} 1 \text{ w. p. } \beta(x(d)) \\ -1 \text{ otherwise .} \end{cases}$$

Denote $\alpha := \frac{2\beta^* - 1}{2}$. We have $\mathbb{E}[Z_s] = 2\beta(x(d)) - 1 \geq (2\beta^* - 1) = 2\alpha$ and so we have $\mathbb{E}\left[D_{T_p(x(h))}(h)\right] \geq 2\alpha p$. Noting that $Z_s \in \{-1, 1\}$, we can directly use the Hoeffding bound to get

$$\begin{aligned} \Pr\left[D_{T_p(x(h))}(h) < \alpha p\right] &\leq \Pr\left[D_{T_p(x(h))}(h) < \left(\frac{2\beta(x(d)) - 1}{2}\right) p\right] \\ &\leq \exp\{-\frac{(2\beta(x(d)) - 1)^2 p}{8}\} \\ &\leq \exp\{-\frac{\alpha^2 p}{2}\}, \end{aligned}$$

and so, for any $t_0(h) \geq 1$ and $x(h) \in \mathcal{X}^h$, we can use the union bound to get

$$\begin{aligned} \Pr\left[\cup_{p=t_0(h)}^{N_T(x(h))} \left\{ D_{T_p(x(h))}(h) < \alpha p \right\}\right] &\leq \sum_{p=t_0(h)}^{N_T(x(h))} \exp\{-\frac{\alpha^2 p}{2}\} \\ &\leq \sum_{p=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 p}{2}\} \\ &\leq \int_{u=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 u}{2}\} du \\ &= \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}. \end{aligned}$$

We need to bound the probability that the above bad event happens *for any* context $x(h) \in \mathcal{X}^h$ and model order $h \geq d$. To do this, we apply the union bound twice more, to get

$$\begin{aligned}\Pr\left[\cup_{h=d}^{D} \cup_{x(h)\in\mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))} \left\{D_{T_p(x(h))}(h) < \alpha p\right\}\right] &\leq \sum_{h=d}^{D} \sum_{x(h)\in\mathcal{X}^h} \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2} \\ &= \left(\sum_{h=d}^{D} \frac{2\cdot 2^h \cdot e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}\right) \\ &\leq \epsilon/2\end{aligned}$$

if $t_0(h) \geq t_{\mathsf{high}}(h) = \frac{2}{(\alpha)^2}\ln\left(\frac{4(D-d)\cdot 2^h}{\epsilon(\alpha)^2}\right)$.

Setting $t_0(h) = t_{\mathsf{high}}(h)$ bounds the probability of the bad event as defined in Equation (2.52), and completes our proof.

□

Now, the proof of Proposition 2.6.15 follows directly from Lemmas 2.6.1 and 2.6.16. We denote as shorthand the following:

$$\begin{aligned}\Delta_T^{(h)} &= \Delta_T(\eta_1^T; g_{\mathsf{unif}}) \\ V_T^{(h)} &= V_T(\eta_1^T; g_{\mathsf{unif}})\end{aligned}$$

Substituting $g(\cdot) = g_{\mathsf{unif}}(\cdot)$ into Lemma 2.6.1, we have

$$\begin{aligned}R_{T,d} \leq R_{T,h} &\leq \left(\sqrt{V_T^{(h)}\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + \frac{\ln\left(\frac{Z}{g_{\mathsf{unif}}(h)}\right)}{\ln 2}\right) \\ &\leq \left(\sqrt{V_T^{(h)}\ln 2} + \frac{2}{3}\ln 2 + 1\right)\left(1 + 2^h\right)\end{aligned}$$

Thus, it remains to bound the variance term $V_T^{(h)}$. We denote the final learning rate as

$$\eta_T^{(h)} = \frac{\ln 2}{\Delta_{T-1}^{(h)}} \geq \frac{\ln 2}{\Delta_T^{(h)}}$$

and from [98] that

$$\begin{aligned}\Delta_T^{(h)} &\leq \sqrt{V_T^{(h)}\ln 2} + \frac{2}{3}\ln 2 + 1 \\ &\leq \sqrt{V_T^{(h)}}\left(\sqrt{\ln 2} + \frac{4}{3}\ln 2 + 2\right)\Big(\text{ as } \sqrt{V_T^{(h)}} \geq \sqrt{v_1^{(h)}} = \frac{1}{2}\Big) \\ &\leq 6\sqrt{V_T^{(h)}}\ln 2.\end{aligned}$$

Together, these give us

$$\eta_T^{(h)} \geq \frac{1}{6\sqrt{V_T^{(h)}}}$$

and therefore, we have with probability greater than or equal to $(1-\epsilon)$,

$$\begin{aligned}
V_T^{(h)} &\leq \sum_{t=1}^{T} w_{t,1-X_t^*}^{(h)} \\
&\leq 2^h \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T^{(h)}(2\beta^* - 1)} \right) \\
&\leq 2^h \left( t_{\mathsf{high}}(h) + \frac{6\sqrt{V_T^{(h)}}}{(2\beta^* - 1)} \right) \\
&\leq 2^h \left( \frac{8}{(2\beta^* - 1)^2} \ln\left( \frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right) + \frac{6\sqrt{V_T^{(h)}}}{(2\beta^* - 1)} \right)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\sqrt{V_T^{(h)}} &\leq \frac{8 \cdot 2^h}{(2\beta^* - 1)^2} \ln\left( \frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right) + \frac{6 \cdot 2^h}{(2\beta^* - 1)} \\
&\leq \frac{14 \cdot 2^h}{(2\beta^* - 1)^2} \ln\left( \frac{8 \cdot 2^h}{\epsilon(2\beta^* - 1)^2} \right)
\end{aligned}$$

This gives us

$$R_{T,d} = \mathcal{O}\Big( \frac{2^{2h}}{(2\beta^* - 1)^2} \left( h + \ln\left( \frac{1}{\epsilon(2\beta^* - 1)} \right) \right) \Big).$$

with probability greater than or equal to $(1-\epsilon)$. This completes the proof.

### Analysis for a higher-than-needed model order for ValidationOverAdaHedge($D$)

Here, we consider the regret accumulated by the base algorithm $\mathcal{A}_h$ for any $h \geq d$. The regret guarantee is given by the following proposition.

**Proposition 2.6.18.** *Let $h \geq d$. For any sequence $\{X_t, Y_t\}_{t=1}^T$ the base algorithm $\mathcal{A}_h$ gives regret with respect to the best $d^{th}$-order tree expert in hindsight*

$$R_{T,d} = \mathcal{O}\Big( \frac{2^h}{(2\beta^* - 1)} \sqrt{\left( h + \ln\left( \frac{(D-d)}{\epsilon(2\beta^* - 1)^2} \right) \right)} \Big)$$

*with probability at least* $(1-\epsilon)$ *on any sequence* $(X_t, Y_t)_{t\geq 1}$ *that satisfies the* $d^{th}$*-order stochastic condition on responses (Definition 2.1.1) with parameter* $\beta^*$*.*

Observe the sub-optimal scaling in terms of $2^h$ in the regret bound for the case where $d < h$. We now proceed to prove Proposition 2.6.18, which is really a simple adaptation of the stochastic AdaHedge proof [95] to the contextual prediction case. (Note that the result cannot be applied out-of-the-box as if for the naive update, as it is linear in the number of experts $K$, which would be prohibitively large here ($2^{2^h}$).)

Observe from the second-order bound on the base algorithm $\mathcal{A}_h$ that we have

$$R_{T,d} = 2\Delta_T^{(h)} = 2\sqrt{V_T^{(h)} \cdot 2^h \ln 2} + \frac{8}{3} \cdot 2^h \ln 2 + 4,$$

and thus it suffices to bound the variance of the algorithm $V_T^{(h)}$. By a similar argument as before, we have

$$\begin{aligned} V_T^{(h)} &\leq \sum_{t=1}^{T} w_{t,1-Y_t^*}^{(h)}(\eta_t^{(h)}) \\ &= \sum_{t=1}^{T} \frac{e^{-\eta_t^{(h)} D_t(h)}}{1 + e^{-\eta_t^{(h)} D_t(h)}} \\ &\leq \sum_{t=1}^{T} \min\{e^{-\eta_t^{(h)} D_t(h)}, 1\} \\ &\leq \sum_{t=1}^{T} \min\{e^{-\eta_T^{(h)} D_t(h)}, 1\} \end{aligned}$$

where $D_t(h)$ is the gap between predictions as in Equation (2.40), and the last inequality is because $\eta_1^T$ is a decreasing sequence according to the update in Equation (2.19).

Therefore, we have

$$V_T((\eta^{(h)})_1^T) \leq \sum_{t=1}^{T} \min\{e^{-\eta_t^{(h)} D_t(h)}, 1\}. \tag{2.53}$$

We observe that Equation (2.53) can be effectively unraveled to get a closed-form variance bound for particular evolutions of $\{D_t(h)\}_{t\geq 1}$. Particularly, we care about $D_t(h)$ as a function of $N_t(X_t(h))$, the number of appearances so far of the current context. For a given model order $h$, we define the following statistical event of the statistics $D_t(h)$ increasing linearly as a function of $N_t(X_t(h))$ after a sufficient number of appearances of the context $X_t(h)$:

$$\Upsilon(h; t_0(h), \alpha) := \{D_t(h) \geq \alpha N_t(X_t(h)) \text{ for all } t \text{ such that } N_t(X_t(h)) \geq t_0(h)\}. \tag{2.54}$$

We will subsequently show (in Lemma 2.6.20) that the *intersection* of the events $\{\Upsilon(h; t_0(h), \alpha)\}_{h=d}^D$ holds with high probability for appropriately chosen parameters $\{t_0(h)\}_{h=d}^D$ and $\alpha$.

We will now show that, given this event, we can bound the variance of the algorithm.

**Lemma 2.6.19.** *Given the event $\Upsilon(h; t_0(h), \alpha)$ in Equation (2.54) for a given value of $h$ and choice of parameters $(t_0(h), \alpha)$, we have*

$$\sum_{t=1}^{\infty} w_{t,1-Y_t^*}^{(h)}(\eta_t^{(h)}) \leq 2^h \left( t_0(h) + \frac{1}{\eta_T^{(h)}\alpha} \right). \tag{2.55}$$

*Proof.* We can directly use the condition in Equation (2.54). For values of $t$ such that $N_t(X_t(h)) < t_0(h)$, we apply $w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq 1$. Otherwise, we use $w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq e^{-\eta_T \alpha N_t(X_t(h))}$.

Combining the two gives us

$$\begin{aligned}
\sum_{t=1}^{\infty} w_{t,1-Y_t^*}^{(h)}(\eta_t^{(h)}) &\leq \sum_{x(h)\in\mathcal{X}^h} \left( t_0 + \sum_{s=t_0(h)}^{N_T(x(h))} e^{-\eta_T^{(h)}\alpha s} \right) \\
&\leq 2^h t_0(h) + \sum_{x(h)\in\mathcal{X}^h} \sum_{s=t_0(h)}^{\infty} e^{-\eta_T^{(h)}\alpha s} \\
&\leq 2^h \left( t_0(h) + \sum_{s=t_0(h)}^{\infty} e^{-\eta_T^{(h)}\alpha s} \right) \\
&\leq 2^h \left( t_0(h) + \frac{e^{-\eta_T^{(h)}\alpha}}{1 - e^{-\eta_T^{(h)}\alpha}} \right).
\end{aligned}$$

Now, we have

$$\begin{aligned}
\frac{e^{-\eta_T^{(h)}\alpha}}{1 - e^{-\eta_T^{(h)}\alpha}} &= \frac{1}{e^{\eta_T^{(h)}\alpha} - 1} \\
&\leq \frac{1}{\eta_T^{(h)}\alpha}
\end{aligned}$$

by the inequality $e^a \geq 1 + a$ for $a \geq 0$. Substituting this above gives us our required result. □

It remains to show that the event in Equation (2.54) is met with high probability for $(X_t, Y_t)_{t\geq 1}$ satisfying the $d^{th}$-order *realizability condition* with parameter $\beta^*$, and for any $d \leq h$. We use a standard Hoeffding-bounding technique to show this.

**Lemma 2.6.20.** *Let $\epsilon \in (0, 1]$. For a process $\{(X_t, Y_t)\}_{t\geq 1}$ satisfying the $d^{th}$-order realizability condition with parameter $\beta^* > 1/2$, the event $\cap_{h=d}^{D} \Upsilon(h; t_0(h), \alpha)$ holds for parameter values*

$$\alpha := \frac{(2\beta^* - 1)}{2} \tag{2.56}$$

$$t_0(h) = t_{\mathsf{high}}(h) := \frac{2}{\alpha^2} \ln \left( \frac{4(D-d) \cdot 2^{h+1}}{\alpha^2 \epsilon} \right) \tag{2.57}$$

*with probability greater than or equal to $(1 - \epsilon/2)$.*

*Proof.* Essentially, we need to obtain to bound properties of the gap sequence $\{D_{t,(h)}\}_{t=1}^{T}$ so defined in Equation (2.40) – we use the Hoeffding bound for this. This proof is a simple adaptation of the proof in the original AdaHedge paper [95] to the case of contextual prediction.

We denote the $p^{th}$ epoch of arrival of context $x(h) \in \mathcal{X}^h$ by $T_p(x(h))$. Showing that the event $\cap_{h=d}^{D} \Upsilon(h; t_0(h), \alpha)$ holds with probability greater than or equal to $(1 - \epsilon/2)$ is exactly equivalent to showing that the probability of the following bad event

$$\left\{ \cup_{h=d}^{D} \cup_{\mathbf{x}_{(h)} \in \mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))} \left\{ D_{T_p(x(h))}(h) < \alpha p \right\} \right\} \tag{2.58}$$

is less than or equal to $\frac{\epsilon}{2}$. We proceed by showing exactly this.

We condition on the partition of the time horizon $[T]$ into the subsets

$$\left\{ \mathcal{T}(\mathbf{x}_{(h)}) := \{ t \in [T] : \mathbf{X}_{(h)}(t) = \mathbf{x}_{(h)} \} \right\}_{\mathbf{x}_{(h)} \in \mathcal{X}^h}$$

From the definition of a $d^{th}$-order stochastic process, we have $Y_t | \{X_t, (X_s, Y_s)_{s=1}^{t-1}\}$ i.i.d $\sim P^*(\cdot | X_t(d))$. This means that $Y_t$ is independent of $(X_t(D, \ldots, D_d), X_s, Y_s)_{s=1}^{t-1}$ conditioned on $X_t(d)$, and we can write

$$D_{T_p(x(h))}(h) = \sum_{s'=1}^{p} 2Z_{s'}$$

where

$$\{Z'_s\}_{s' \geq 1} \text{ i.i.d } \sim \begin{cases} 1 \text{ w. p. } \beta(x(d)) \\ -1 \text{ otherwise .} \end{cases}$$

Denote $\alpha := \frac{2\beta^* - 1}{2}$. We have $\mathbb{E}[Z_s] = 2\beta(x(d)) - 1 \geq (2\beta^* - 1) = 2\alpha$ and so we have $\mathbb{E}\left[D_{T_p(x(h))}(h)\right] \geq 2\alpha p$. Noting that $Z_s \in \{-1, 1\}$, we can directly use the Hoeffding bound to get

$$\begin{aligned} \Pr\left[D_{T_p(x(h))}(h) < \alpha p\right] &\leq \Pr\left[D_{T_p(x(h))}(h) < \left(\frac{2\beta(x(d)) - 1}{2}\right) p\right] \\ &\leq \exp\{-\frac{(2\beta(x(d)) - 1)^2 p}{8}\} \\ &\leq \exp\{-\frac{\alpha^2 p}{2}\}, \end{aligned}$$

and so, for any $t_0(h) \geq 1$ and $x(h) \in \mathcal{X}^h$, we can use the union bound to get

$$\begin{aligned}
\Pr\left[\cup_{p=t_0(h)}^{N_T(x(h))} \left\{D_{T_p(x(h))}(h) < \alpha p\right\}\right] &\leq \sum_{p=t_0(h)}^{N_T(x(h))} \exp\{-\frac{\alpha^2 p}{2}\} \\
&\leq \sum_{p=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 p}{2}\} \\
&\leq \int_{u=t_0(h)}^{\infty} \exp\{-\frac{\alpha^2 u}{2}\} du \\
&= \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}.
\end{aligned}$$

We need to bound the probability that the above bad event happens *for any* context $x(h) \in \mathcal{X}^h$ and model order $h \geq d$. To do this, we apply the union bound twice more, to get

$$\begin{aligned}
\Pr\left[\cup_{h=d}^{D} \cup_{x(h)\in\mathcal{X}^h} \cup_{p=t_0(h)}^{N_T(x(h))} \left\{D_{T_p(x(h))}(h) < \alpha p\right\}\right] &\leq \sum_{h=d}^{D} \sum_{x(h)\in\mathcal{X}^h} \frac{2e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2} \\
&= \left(\sum_{h=d}^{D} \frac{2 \cdot 2^h \cdot e^{-\frac{\alpha^2 t_0(h)}{2}}}{\alpha^2}\right) \\
&\leq \frac{2 \cdot 2^{D+1} \cdot e^{-\frac{(\alpha)^2 t_0}{2}}}{(\alpha)^2} \leq \epsilon/2
\end{aligned}$$

if $t_0(h) \geq t_{\mathsf{high}}(h) = \frac{2}{(\alpha)^2} \ln\left(\frac{4(D-d)\cdot 2^h}{\epsilon(\alpha)^2}\right)$.

Since the expression is independent of the partitioning $\{\mathcal{T}(\mathbf{x}_{(h)})\}$ on which we conditioned, we have

$$\Pr\left[\cup_{h=d}^{D} \cup_{\mathbf{x}_{(h)}\in\mathcal{X}^h} \cup_{p=t_0}^{N_T(\mathbf{x}_{(h)})} \left\{D_{t_p(\mathbf{x}_{(h)})} < \alpha p\right\}\right] \leq \epsilon/2,$$

Setting $t_0(h) = t_{\mathsf{high}}(h)$ bounds the probability of the bad event as defined in Equation (2.58), and completes our proof.

□

Now, the proof of Proposition 2.6.18 follows directly from Lemmas 2.6.6 and 2.6.19. We denote the final learning rate as

$$\eta_T^{(h)} = \frac{2^h \ln 2}{\Delta_{T-1}^{(h)}} \geq \frac{2^h \ln 2}{\Delta_T^{(h)}}$$

and thus we have

$$\begin{aligned} V_T^{(h)} &\le 2^h \left( t_{\mathsf{high}}(h) + \frac{2\Delta_T^{(h)}}{2^h \ln 2 \cdot (2\beta^* - 1)} \right) \\ &= 2^h \cdot t_{\mathsf{high}}(h) + \frac{2\Delta_T^{(h)}}{\ln 2 \cdot (2\beta^* - 1)}. \end{aligned}$$

Further, recall that

$$\begin{aligned} \Delta_T^{(h)} &\le \sqrt{V_T^{(h)} \cdot 2^h \ln 2} + \frac{4}{3} \cdot 2^h \ln 2 + 2 \\ &\le \sqrt{2^{2h} \ln 2 \cdot t_{\mathsf{high}}(h) + \frac{2^h \ln 2 \cdot \Delta_T^{(h)}}{(2\beta^* - 1)}} + \frac{4}{3} \cdot 2^h \ln 2 + 2. \\ &= \sqrt{A + B\Delta_T^{(h)}} + C \end{aligned}$$

where

$$\begin{aligned} A &= 2^{2h} \ln 2 \cdot t_{\mathsf{high}}(h) \\ B &= \frac{2^h \ln 2}{(2\beta^* - 1)} \\ C &= \frac{4}{3} \cdot 2^h \ln 2 + 2. \end{aligned}$$

Thus, we have

$$\begin{aligned} \Delta_T^{(h)} &\le \sqrt{A + B\Delta_T^{(h)}} + C \\ \implies \left(\Delta_T^{(h)} - C\right)^2 &\le A + B\Delta_T^{(h)}. \end{aligned}$$

Noting that $(\Delta_T^{(h)})^2 - C^2 - 2\Delta_T^{(h)} C \le (\Delta_T^{(h)})^2 + C^2 - 2\Delta_T^{(h)} C = \left(\Delta_T^{(h)} - C\right)^2$, this implies

$$\begin{aligned} \implies (\Delta_T^{(h)})^2 - (B + 2C)\Delta_T^{(h)} - (A + C^2) &\le 0 \\ \implies \Delta_T^{(h)} &\le \sqrt{A + C^2} + (B + 2C) \\ &\le \sqrt{A} + B + 3C \end{aligned}$$

where the second-to-last inequality follows from Fact 2.7.11, and the last inequality follows because for any two numbers $a, b \ge 0$, we have $\sqrt{a + b} \le \sqrt{a} + \sqrt{b}$.

Substituting back, we get

$$\begin{aligned}\Delta_T^{(h)} &\leq 2^h \ln 2 \cdot \sqrt{t_{\mathsf{high}}(h)} + \frac{2^h \ln 2}{(2\beta^* - 1)} + 4 \cdot 2^h \ln 2 + 6 \\ &= 2\sqrt{2} \cdot \frac{2^h \ln 2}{(2\beta^* - 1)} \cdot \sqrt{\ln\left(\frac{4(D-d) \cdot 2^{h+1}}{(2\beta^* - 1)^2 \epsilon}\right)} + \frac{2^h \ln 2}{(2\beta^* - 1)} + 4 \cdot 2^h \ln 2 + 6 \\ &= \mathcal{O}\Big(\frac{2^h}{(2\beta^* - 1)} \sqrt{\left(h + \ln\left(\frac{(D-d)}{\epsilon(2\beta^* - 1)^2}\right)\right)}\Big)\end{aligned}$$

with probability greater than or equal to $(1 - \epsilon)$. This completes the proof.

## Ruling out higher-order models that over-fit

Now, we get into the essence of provable model selection guarantees, starting by showing that we can effectively limit the contribution of higher-order models to the algorithmic variance. Owing to the explicit complexity penalty in SRM, this is a much easier task for SRMOverAdaHedge($D$) than for ValidationOverAdaHedge($D$). The latter algorithm relies on algorithmic errors to approximately capture over-fitting error and successfully rule out higher-order models in an online fashion.

### Ruling out higher-order models for SRMOverAdaHedge($D$)

We can make two clear inferences from Lemma 2.6.16:

1. SRMOverAdaHedge($d$) gives us a regret scaling only as a function of $d$ in terms of $\mathcal{O}(2^{2d}\left(d + \ln\left(\frac{1}{\epsilon}\right)\right))$.

2. For $h > d$, SRMOverAdaHedge($h$) gives us sub-optimal scaling $Oh(2^{2h}\left(h + \ln\left(\frac{1}{\epsilon}\right)\right))$. The reason for sub-optimality is because of sample splitting: for every true context $x(d) \in \mathcal{X}^d$, we are unnecessarily splitting the data into $2^{d-h}$ extra contexts and treating the best predictors for these contexts as independent.

It is clear, particularly from the second inference, that we would like to control the posterior probability with which we select overly complex models. This quantity is expressed as $q_t(h; \eta_t, g_{\mathsf{prop}})$ for all $h > d$. Now, we consider an explicit upper bound on $q_t(h; \eta_t, g_{\mathsf{prop}})$ and show how it decreases with $t$.

Using Equation (2.44), it is convenient to consider the following upper bound on the quantity $q_t(h; \eta_t, g_{\mathsf{prop}})$ for $h > d$:

$$\begin{aligned} q_t(h; \eta_t, g_{\mathsf{prop}}) &= \frac{Q_t(h; \eta_t, g_{\mathsf{prop}})}{\sum_{h'=0}^{D} Q_t(h'; \eta_t, g_{\mathsf{prop}})} \\ &\leq \frac{Q_t(h; \eta_t, g_{\mathsf{prop}})}{Q_t(d; \eta_t, g_{\mathsf{prop}})} \\ &\leq \exp\{\eta_t(\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t - 2^h \ln 2 + 2 \cdot 2^d \ln 2\} \end{aligned}$$

We should expect that as $t$ becomes large the difference in estimated approximation errors is negligible, i.e. we will observe that $\widehat{\pi}_h(t) = \widehat{\pi}_d(t)$ with high probability. We would then get a scaling of $q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \exp\{-2^h \ln 2\}$. However, we can say $\widehat{\pi}_h(t) = \widehat{\pi}_d(t)$ with high probability only after $\mathcal{O}(2^h)$ rounds. Before this, and particularly for times between $\mathcal{O}(2^d)$ and $\mathcal{O}(2^h)$, we have to worry about the difference in approximation errors, $\eta_t(\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t$. This is the *over-fitting regime* in which the $h$th order model may look deceptively better. Luckily, we can cap this quantity as well owing to already established statistical guarantees on the sequence $\{X_t\}_{t \geq 1}$. The following lemma expresses this.

**Lemma 2.6.21.** *The process $\{(X_t, Y_t)\}_{t \geq 1}$ satisfying Equation (2.48) for all $h \geq d$ and for*

$$t_0(h) = t_{\mathsf{high}}(h)$$

*directly implies*

$$(\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t \leq \min\{\frac{t}{2}, 2^{h-1} t_{\mathsf{high}}(h)\}. \tag{2.59}$$

The two quantities on the right hand side of Equation (2.59) have different operational meaning. The bound in terms of $\frac{t}{2}$ will be used to show that for a small number of rounds, the doubly exponential prior on model order $h$ will weigh this model order down and prevent it from being selected prematurely *even if it could be leveraged for more accurate prediction in later rounds, as would be the case when the data is out-of-model.* On the other hand, the bound in terms of $2^{h-1} t_{\mathsf{high}}(h)$ is useful to conclusively rule out the $h^{th}$-order model even in later rounds *for the case where data is realized from a $d^{th}$-order model*, by which time it is clear that the higher-order model does not lead to any improvement in approximability.

*Proof.* It suffices to prove the following two inequalities separately:

$$\begin{aligned} (\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t &\leq \frac{t}{2} \\ (\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t &\leq 2^{h-1} t_{\mathsf{high}}(h). \end{aligned}$$

Recall the notation we defined for the best $d^{th}$-order tree expert at time $t$, $\widehat{F}_d(t)$, as well as the number of appearances of context $x(h)$ at time $t$, denoted by $N_t(x(h))$.

From Definition 2.6.12, we have

$$
\begin{aligned}
&(\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t \\
&= \sum_{x(d)\in\mathcal{X}^d} N_t(x(d)) \left(1 - \max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(d))\}\right) - \sum_{x(h)\in\mathcal{X}^h} N_t(x(h)) \left(1 - \max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\}\right) \\
&= \sum_{x(d)\in\mathcal{X}^d} \underbrace{\left(\sum_{x(h):x(d)\subset x(h)} N_t(x(h)) \left(\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(\widehat{F}_d(t)(x(d))|x(d))\right)\right)}_{T_1}
\end{aligned}
$$

Let $T_1$ be the quantity under the brace (for shorthand). We also define the number of *super-contexts* of length $h$ that contain $x(d)$,

$$S_{t,h-d}(x(d)) := \sum_{x(h):x(d)\subset x(h)} \mathbb{I}[N_t(x(h)) > 0].$$

Now, we have one of two cases:

1. We have $N_t(x(d)) \leq t_{\mathsf{high}}$. In this case, we have $T_1 \leq \frac{t_{\mathsf{high}}}{2}$.

2. $N_t(x(d)) > t_{\mathsf{high}}$. In this case, we have $\widehat{F}_d(t)(x(d)) = f^*(x(d))$ from Equation (2.48), and we directly get

$$T_1 = \sum_{x(h)\in\mathcal{X}_{-d}(h)} N_t(x(h)) \left(\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(f^*_d(x(d))|x(d))\right)$$

   where $\mathcal{X}_{-d}(h) := \{x(h) : x(d) \subset x(h) \text{ and } \arg\max\{\widehat{P}_t(y|x(h))\} \neq f^*(x(d))\}$.

   Clearly, the over-fitting effect is created *only* by the set of contexts $x(h)$ for which the best predictor does not match $f^*(x(d))$. From Lemma 2.6.17, Equation (2.48) is satisfied for all $h \geq d$ and for $N_t(x(h)) \geq t_{\mathsf{high}}(h)$. It is easy to see that Equation (2.48) implies a non-negative separation between the truly correct predictor $f^*(x(d))$ and its alternative, and so we have

$$\arg\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\} = f^*(x(d)) \text{ if } N_t(x(h)) \geq t_{\mathsf{high}}(h).$$

   Substituting this directly, and noting that

$$\max_{y\in\mathcal{X}}\{\widehat{P}_t(y|x(h))\} - \widehat{P}_t(f^*_d(x(d))|x(d)) \leq 1/2$$

gives us

$$\begin{aligned} T_1 &\leq \sum_{x(h):x(d)\subset x(h) \text{ and } N_t(x(h))\leq t_{\mathsf{high}}(h)} \frac{\min\{N_t(x(h), t_{\mathsf{high}}(h)\}}{2}\} \\ &\leq \sum_{x(h):x(d)\subset x(h) \text{ and } N_t(x(h))\leq t_{\mathsf{high}}(h)} \frac{t_{\mathsf{high}}(h)}{2} \\ &\leq S_{t,h-d}(x(d))\frac{t_{\mathsf{high}}(h)}{2}. \end{aligned}$$

Noting that $1 \leq 2^{h-d}$ and $S_{t,h-d}(x(d)) \leq 2^{h-d}$ gives us

$$T_1 \leq 2^{h-d}\frac{t_{\mathsf{high}}(h)}{2},$$

and substituting back this expression yields

$$\begin{aligned} (\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t &\leq \sum_{x(d)\in\mathcal{X}^d} T_1 \\ &\leq 2^{h-1}t_{\mathsf{high}}(h). \end{aligned}$$

This completes our proof.

□

Recall that for all $t > T_0(h)$ where $T_0(h)$ is as defined in Lemma 2.6.13 with respect to $t_0(h) = t_{\mathsf{high}}(h)$, we have $\eta_t < \frac{\ln 2}{t_0}$. Under this condition, the explicit cap on the over-fitting effect as defined in Lemma 2.6.21, together with the adaptive regularization of AdaHedge, ensures that we can sufficiently restrict the contribution of higher-order models.

We use Equation (2.59) to get

$$\begin{aligned} q_t(h;\eta_t,g_{\mathsf{prop}}) &\leq \exp\{\eta_t(\widehat{\pi}_d(t) - \widehat{\pi}_h(t))t - 2^h\ln 2 + 2\cdot 2^d \ln 2\} \\ &\leq \exp\{\frac{2^{h-1}t_{\mathsf{high}}(h)\ln 2}{t_{\mathsf{high}}(h)} - 2^h\ln 2 + 2^{d+1}\ln 2\} \\ &\leq \exp\{-2^{h-1}\ln 2 + 2^{d+1}\ln 2\} \\ &= 2^{-2^{h-1}+2^{d+1}}. \end{aligned}$$

Therefore, we can apply Lemma 2.6.16 to get

$$\begin{aligned} \sum_{t=T_0}^{T} q_t(h;\eta_t,g_{\mathsf{prop}})w^{(h)}_{t,1-Y^*_t}(\eta_t) &\leq 2^{-2^{h-1}+2^{d+1}}\sum_{t=T_0}^{T} w^{(h)}_{t,1-Y^*_t} \\ &\leq 2^{h-2^{h-1}+2^{d+1}}\left(t_{\mathsf{high}}(h) + \frac{1}{\eta_T\alpha}\right). \end{aligned}$$

It is now easy to check that

$$\begin{aligned} & 2h \leq 2^{h-1} - 2^{d+1} \text{ for all } h \geq d+4 \text{ and } d \geq 0 \\ \implies & h - 2^{h-1} + 2^{d+1} \leq -h \\ \implies & 2^{h-2^{h-1}+2^{d+1}} \leq 2^{-h}. \end{aligned}$$

Therefore, for $h \geq d+4$, we get

$$\sum_{t=T_0}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq 2^{-h} \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T \alpha} \right).$$

For $h < d+4$, we do not try to non-trivially bound $q_t(h; \eta_t, g_{\mathsf{prop}})$. We directly use Lemma 2.6.16 to get

$$\sum_{t=T_0}^{T} q_t(h) w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq 2^{h} \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T \alpha} \right).$$

We have thus guaranteed that the contribution from the higher-order models (particularly for $h \geq d+4$) not only has no exponential dependence on $h$, but is in fact exponentially decaying in $h$! Ultimately, we will see that we get a very weak linear dependence on $D$, the maximum model order, in our regret bound.

### Ruling out higher-order models for ValidationOverAdaHedge($D$)

Using Equation (2.45), it is convenient to consider the following upper bound on the quantity $q_t(h; \eta_t)$ for $h > d$:

$$\begin{aligned} q_t(h; \eta_t) &= \frac{Q_t(h; \eta_t)}{\sum_{h'=0}^{D} Q_t(h'; \eta_t)} \\ &\leq \frac{Q_t(h; \eta_t)}{Q_t(d; \eta_t)} \\ &\leq \exp\{-\eta_t (H_t(h; \{\eta_s\}_{s=1}^t) - H_t(d; \{\eta_s\}_{s=1}^t))\}. \end{aligned}$$

Thus, it suffices to obtain a uniform lower bound on the excess over-fitting loss, $H_t(h; \{\eta_s\}_{s=1}^t) - H_t(d; \{\eta_s\}_{s=1}^t)$ for any $h > d$. This is a highly technical quantity to analyze, owing to the dependencies between contexts, responses, and algorithmic updates. Definition 2.1.4 makes a major simplification on the evolution of the contexts that allows us to still highlight the key ideas in characterizing validation error of over-fitting models. Without loss of generality, we condition[9] on $X_1 = \mathbf{0}$, which is the only randomness in the context process. In this case, the contexts become

$$X_t := x_t \equiv t \mod 2^D. \tag{2.60}$$

[9]The analysis will be identical for any choice of $x_1 \in \mathcal{X}^D$; we only make this choice for convenience.

Note that this automatically implies that $x_t(h) \equiv (t \mod 2^h)$ for any $h \in \{0, 1, \ldots, D\}$. For convenience, we will define $k_0(t,h) := \lfloor t/2^h \rfloor$ as shorthand for the number of times the context $X_t$ has been seen at round $t$.

Under the above assumption, we state and prove the following main lemma that bounds the excess over-fitting loss.

**Lemma 2.6.22.** *Let the contexts be periodic according to Definition 2.1.4. Then, for all $h \in \{1, \ldots, D\}$, and given the event $\Upsilon(h, t_{\mathsf{high}}(h); \alpha^*)$, we have*

$$H_t(h; \{\eta_s\}_{s=1}^t) - H_t(d; \{\eta_s\}_{s=1}^t) \geq \begin{cases} \frac{\alpha^* \cdot t}{4k_0(\tau(h),h)} - \frac{15 \cdot 2^d \cdot t_{\mathsf{high}}(d)}{2} \text{ if } t'_{\mathsf{high}}(h) \leq t \leq \tau(h) \\ \frac{\alpha^* \cdot 2^h}{2} - c'(\alpha^*) \cdot \ln\left(\frac{(D-d)}{(\alpha^*)^2 \epsilon}\right) - \frac{15 \cdot 2^d \cdot t_{\mathsf{high}}(d)}{2} \text{ if } t > \tau(h). \end{cases} \tag{2.61}$$

*with probability at least $(1-\epsilon)$. Here, we define*

$$t'_{\mathsf{high}}(h) := \frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)}{(\alpha^*)^2} \cdot \ln\left(\frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)(D-d)}{(\alpha^*)^2 \epsilon}\right) \text{ and}$$
$$\tau(h) := 13h \cdot \sqrt{t_{\mathsf{high}}(h)} \cdot 2^h,$$

*and $c(\alpha^*), c'(\alpha^*)$ are universal positive constants that only depend on $\alpha^*$.*

We first prove this lemma, and then subsequently use it to bound the contribution coming from higher-order models. Notice the scaling of $t'_{\mathsf{high}}(h) = \mathcal{O}(h^3)$.

*Proof.* We consider the sub-sequences

$$\mathcal{T}(x(d)) := \{k \cdot 2^d + x(d)\}_{k \geq 0},$$

for all $x(d) \in \mathcal{X}(d)$. By the periodic assumption on contexts, we have

$$Y_t \sim \text{Ber}(P^*(1|x(d))) \text{ for all } t \in \mathcal{T}(x(d)).$$

Thus, we get

$$\begin{aligned} H_{t-1}(h) - H_{t-1}(d) = &\sum_{x(d) \in \mathcal{X}(d)} \sum_{k=0}^{k_0(t,d)-1} (w^{(h)}_{k,1-f^*(x(d))} - w^{(d)}_{k,1-f^*(x(d))}) \cdot Z_k(\mathbf{x}(d)) \\ &+ \sum_{x(d)=0}^{x_t(d)} (w^{(h)}_{k_0(t,d),1-f^*(x(d))} - w^{(d)}_{k_0(t,d),1-f^*(x(d))}) W_{k_0(t,d)}(x(d)), \end{aligned}$$

where we define

$$W_k(x(d)) = \begin{cases} 2Y_k(x(d)) - 1 \text{ if } f^*(x(d)) = 1 \\ 1 - 2Y_k(x(d)) \text{ if } f^*(x(d)) = 0. \end{cases} \tag{2.62}$$

For the case of AdaHedge($d$), we have for all $t$,

$$\begin{aligned}
&\sum_{x(d)\in\mathcal{X}(d)} \sum_{k=0}^{k_0(t,d)-1} w^{(d)}_{k,1-f^*(x(d))} \cdot W_k(x(d)) + \sum_{x(d)=0}^{x_t(d)} w^{(d)}_{k_0(t,d),1-f^*(x(d))} \cdot W_{k_0(t)}(x(d)) \\
&\leq \sum_{s=1}^{t-1} w^{(d)}_{s,1-Y^*_s} \\
&\leq 2^d \left( t_{\mathsf{high}}(d) + \frac{\sqrt{t_{\mathsf{high}}(d)}}{(2\beta^*-1)} + \frac{1}{(2\beta^*-1)^2} + \frac{4}{(2\beta^*-1)} + \frac{6}{(2\beta^*-1)} \right) \\
&\leq \frac{15 \cdot 2^d \cdot t_{\mathsf{high}}(d)}{2},
\end{aligned}$$

where the last inequality can be verified by substituting expressions for $t_{\mathsf{high}}(d)$ and noting that $t_{\mathsf{high}}(d)/(2\beta^*-1)^2 \geq 1$.

Thus, we will henceforth focus on the term coming from AdaHedge($h$), which is given by:

$$T_A(t,h) := \sum_{x(d)\in\mathcal{X}(d)} \sum_{k=0}^{k_0(t,d)-1} w^{(h)}_{k,1-f^*(x(d))} \cdot W_k(x(d)) + \sum_{x(d)=0}^{x_t(d)} w^{(h)}_{k_0(t,d),1-f^*(x(d))} \cdot W_{k_0(t,d)}(x(d))$$

Alternatively, recalling the definition of $Y^*_t$, we can also write

$$T_A(t,h) = \sum_{s=1}^{t-1} w^{(h)}_{s,1-Y^*_s} \cdot W_{k_0(s,d)}(x_s(d)),$$

where recall that $x_s(d) \equiv (s \mod 2^d)$. We state and prove the following technical lemma on $T_A(t,h)$ that uses a martingale argument.

**Lemma 2.6.23.** *We define the event*

$$\Pi(h, t_0(h), \alpha) := \left\{ T_A(t,h) \geq \frac{\alpha^* \cdot t}{4 \cdot k_0(\tau(h))} \text{ for all } t_0(h) \leq t \leq \tau(h) \right\}.$$

*Then, the events $\Pi(h, t'_{\mathsf{high}}(h), \alpha^*)$ hold for all $h = d, \ldots, D$ with probability at least $(1-\epsilon)$, where we define*

$$t'_{\mathsf{high}}(h) := \frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)}{(\alpha^*)^2} \cdot \ln\left( \frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)(D-d)}{(\alpha^*)^2 \epsilon} \right),$$

*and $c(\alpha^*)$ is a positive constant that depends on $\alpha^*$.*

Given Lemma 2.6.23, or more precisely, assuming the event $\Pi(h, t'_{\mathsf{high}}(h), \alpha^*)$ to be given for the moment, we can complete the proof of Lemma 2.6.22. This is because we can use the event $\Upsilon(h, t_{\mathsf{high}}(h); \alpha^*)$ to lower bound the process $T_A(t,h)$ for all $t > \tau(h)$. Since the random variables $|W_k| \leq 1$, we have $T_A(t,h) - T_A(t-1,h) \geq -w^{(h)}_{t,1-Y^*_t}$, and thus we get, for any $t > \tau(h)$,

$$\begin{aligned} T_A(t,h) &\geq \frac{\alpha^* \cdot \tau(h)}{4 \cdot k_0(\tau(h))} - \sum_{s=\tau(h)+1}^{t} w^{(h)}_{s,1-Y^*_s} \\ &\geq \frac{\alpha^* \cdot 2^h}{4} - \sum_{s=\tau(h)+1}^{t} w^{(h)}_{t,1-Y^*_t}. \end{aligned}$$

Furthermore, we recall that

$$w^{(h)}_{t,1-Y^*_t} \leq e^{-\eta^{(h)}_T D_t(h)}.$$

It is easy to verify from the proof of Proposition 2.6.18 that under the event $\Upsilon(h, t_{\mathsf{high}}(h); \alpha^*)$, we have

$$\eta^{(h)}_T \geq \frac{1}{13\sqrt{t_{\mathsf{high}}(h)}}.$$

Further, by the periodic Definition 2.1.4, the number of appearances of each context is $N_t(x(h)) \geq k_0(\tau(h), h)$ for all $t > \tau(h)$ and all $x(h) \in \mathcal{X}(h)$. Putting all of this together, we get (under the event $\Upsilon(h, t_{\mathsf{high}}(h); \alpha^*)$),

$$\begin{aligned} \sum_{s=\tau(h)+1}^{t} w^{(h)}_{s,1-Y^*_s} &\leq \sum_{s=\tau(h)}^{\infty} w^{(h)}_{s,1-Y^*_s} \\ &= 2^h \cdot \sum_{k=k_0(\tau(h),h)}^{\infty} e^{-\frac{\alpha^* k}{13\sqrt{t_{\mathsf{high}}(h)}}} \\ &\leq \frac{2^h \cdot e^{-\frac{\alpha^* \tau(h)}{13 \cdot 2^h \cdot \sqrt{t_{\mathsf{high}}(h)}}}}{1 - e^{-\frac{\alpha^*}{13\sqrt{t_{\mathsf{high}}(h)}}}}. \end{aligned}$$

Substituting the definition of $\tau(h)$, the numerator of the above becomes $2^h \cdot e^{-h} = \left(\frac{2}{e}\right)^h$, while the denominator becomes

$$1 - e^{-\frac{\alpha^*}{13\sqrt{t_{\mathsf{high}}(h)}}} \geq 0.82 \left( \frac{\alpha^*}{13\sqrt{t_{\mathsf{high}}(h)}} \right).$$

Thus, we get

$$\sum_{s=\tau(h)+1}^{t} w_{t,1-Y_t^*}^{(h)} \leq c(\alpha^*) \cdot \sqrt{t_{\mathsf{high}}(h)} \cdot \left(\frac{2}{e}\right)^h \leq c'(\alpha^*) \cdot \ln\left(\frac{(D-d)}{(\alpha^*)^2\epsilon}\right),$$

where the last inequality easily follows as the function $\sqrt{h} \cdot (2/e)^h$ is decreasing in $h$. From the above, we get

$$T_A(t,h) \geq \frac{\alpha^* \cdot 2^h}{4} - c'(\alpha^*) \cdot \ln\left(\frac{(D-d)}{(\alpha^*)^2\epsilon}\right) \text{ for all } t \geq \tau(h).$$

Plugging the three cases for $T_A(t,h)$ back into the definition of $H_{t-1,h} - H_{t-1,d}$ completes the proof of Lemma 2.6.22.

Thus, it only remains to prove the technical Lemma 2.6.23, which we do below.

*Proof.* We define the sequence

$$Z_t(h) := T_A(t,h) - T_A(t-1,h) - \frac{\alpha^* \cdot e^{-k_0(t,h)}}{2}.$$

It turns out that $\{Z_t(h)\}_{t\geq 1}$ is a sub-martingale difference sequence. This follows because we have

$$\begin{aligned}\mathbb{E}_{t-1}\left[T_A(t,h) - T_A(t-1,h)\right] &= \mathbb{E}_{t-1}\left[w_{t,1-Y_t^*}^{(h)} \cdot Z_{k_0(t,d)}(x_t(d))\right] \\ &\geq \alpha^* \cdot w_{t,1-Y_t^*}^{(h)},\end{aligned}$$

where the last step follows because of the periodic assumption on the contexts. Since $Y_t^*$ is deterministic, the quantity $w_{t,1-Y_t^*}^{(h)}$ is a deterministic function of the past, and since $Y_t$ is independent of the past, we get that $Y_t$ is conditionally independent of $w_{t,1-Y_t^*}^{(h)}$ at time $t$. Then, by Equation (2.62) we have $\mathbb{E}\left[Z_{k_0(t,d)}(x_t(d))\right] = (2\beta(x(d)) - 1) \geq (2\beta^* - 1) = \alpha^*$.

Moreover, by the periodic context assumption (Definition 2.1.4), the context $x_t(h)$ has appeared $k_0(t,h)$ times by round $t$. Therefore, we get

$$\begin{aligned}w_{t,1-Y_t^*}^{(h)} &= \frac{e^{-\eta_t^{(h)} \cdot D_t(h)}}{1 + e^{-\eta_t^{(h)} \cdot D_t(h)}} \\ &\overset{\text{(i)}}{\geq} \frac{e^{-\eta_t^{(h)} \cdot k_0(t,h)}}{1 + e^{-\eta_t^{(h)} \cdot k_0(t,h)}} \\ &\geq \frac{e^{-\eta_t^{(h)} \cdot k_0(t,h)}}{2} \\ &\overset{\text{(ii)}}{\geq} \frac{1}{2} e^{-k_0(t,h)}.\end{aligned}$$

Here, inequality (i) follows because we have $D_t(\mathbf{x}_t(h)) \geq k_0(t,h)$, and inequality (ii) follows because we can trivially bound $\eta_t^{(h)} \leq 1$ for all $t \geq 1$.

Thus, we have shown that $\{Z_t(h)\}_{t\geq 1}$ is a sub-martingale difference sequence. For shorthand, we define $\underline{c}(t,h;\alpha^*) := \frac{\alpha^*}{2}\cdot e^{-k_0(t,h)}$, and write $Z_t(h) = T_A(t,h) - T_A(t-1,h) - \underline{c}(t,h;\alpha^*)$. We also note that $|Z_t(h)| \leq 2$ for all $t$. Thus, we can use the Azuma-Hoeffding inequality to show that for any $t \geq 1$, we have

$$\Pr\left[\sum_{s=1}^{t} Z_s(h) < -z_t\right] \leq e^{-\frac{z_t^2}{8t}}$$

for any $z_t > 0$.

We will consider $z_t := \sum_{s=1}^{t} \frac{\underline{c}(s,h;\alpha^*)}{2}$. With this choice, we can substitute the definition of $Z_t(h)$ to get

$$\Pr\left[T_A(t,h) < \sum_{s=1}^{t} \frac{\underline{c}(s,h;\alpha^*)}{2}\right] \leq e^{-\frac{(\sum_{s=1}^{t}\frac{\underline{c}(s,h;\alpha^*)}{2})^2}{8t}}. \tag{2.63}$$

We will show using the above concentration bound that the event $\Pi(h, t'_{\mathsf{high}}(h);\alpha^*)$ holds with probability at least $(1-\epsilon)$ for the given choice of $t'_{\mathsf{high}}(h)$. Note that $k_0(\tau(h),h) \leq c(\alpha^*)\cdot 13h\sqrt{t_{\mathsf{high}}(h)}$ for appropriately chosen constant $c(\alpha^*) > 0$. Further, we note that for all values of $t$ such that $k_0(t,h) \geq 1$, we have

$$\begin{aligned}\sum_{s=1}^{t} \underline{c}(s,h;\alpha^*) &\geq \sum_{k=0}^{k_0(t,h)} \alpha^* \cdot 2^h \cdot \frac{e^{-k}}{2} \\ &\geq \frac{\alpha^* \cdot k_0(t,h)\cdot 2^h}{2k_0(t,h)} \geq \frac{\alpha^*\cdot t}{4k_0(t,h)},\end{aligned}$$

where the last inequality follows for $k_0(t,h) \geq 1$, noting that

$$k_0(t,h)\cdot 2^h \leq t \leq \lceil \frac{t}{2^h} \rceil \cdot 2^h \leq 2\lfloor \frac{t}{2^h} \rfloor \cdot 2^h \leq 2k_0(t,h)\cdot 2^h.$$

On the other hand, for $k_0(t,h) = 0$ (i.e. $t < 2^h$), we have

$$\sum_{s=1}^{t} \underline{c}(s,h;\alpha^*) \geq \alpha^* \cdot \frac{t}{2} \geq \frac{\alpha^* t}{4k_0(\tau(h),h)}$$

where the last inequality follows because $k_0(\tau(h),h) \geq 1/2$. Thus in both cases, we get

$$\sum_{s=1}^{t} \underline{c}(s,h;\alpha^*) \geq \frac{\alpha^*\cdot t}{4k_0(\tau(h),h)} \text{ for all } t \leq \tau(h). \tag{2.64}$$

We then plug in Equation (2.64) into the tail bound of Equation (2.63), we get

$$\Pr\left[T_A(t,h) < \frac{\alpha^* \cdot t}{4k_0(\tau(h),h)}\right] \leq e^{-\frac{(\alpha^*)^2 t^2}{32t \cdot k_0^2(\tau(h),h)}} = e^{-\frac{(\alpha^*)^2 t}{32 \cdot k_0^2(\tau(h),h)}}$$

for every $t \in \{1, \ldots, \tau(h)\}$. Thus, the probability of the complement of $\Pi(h, t_0(h); \alpha^*)$ is upper bounded by

$$\sum_{t=t_0(h)}^{\infty} e^{-\frac{(\alpha^*)^2 t}{32 \cdot k_0^2(\tau(h),h)}} \leq \frac{32 \cdot k_0^2(\tau(h),h)}{(\alpha^*)^2} e^{-\frac{(\alpha^*)^2 t_0(h)}{32 \cdot k_0^2(\tau(h),h)}} \leq \frac{\epsilon}{(D-d)},$$

provided that

$$t_0(h) \geq \frac{32k_0^2(\tau(h),h)}{(\alpha^*)^2} \cdot \ln\left(\frac{32k_0(\tau(h),h)^2(D-d)}{(\alpha^*)^2\epsilon}\right).$$

Note that $k_0(\tau(h),h) \leq c(\alpha^*) \cdot 13h\sqrt{t_{\mathsf{high}}(h)}$ for appropriately chosen constant $c(\alpha^*)$, and so for the choice

$$t'_{\mathsf{high}}(h) := \frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)}{(\alpha^*)^2} \cdot \ln\left(\frac{c(\alpha^*) \cdot h^2 \cdot t_{\mathsf{high}}(h)(D-d)}{(\alpha^*)^2\epsilon}\right), \tag{2.65}$$

the events $\Pi(h; t'_{\mathsf{high}}(h); \alpha^*)$ hold *for all* $h = d, \ldots, D$ with probability at least $(1-\epsilon)$. This completes the proof of Lemma 2.6.23. □

Now that we have completed the proof of Lemma 2.6.23, we have completed the proof of Lemma 2.6.22. □

We now use Lemma 2.6.22 to characterize the contribution coming from higher-order models. First, note that the bounds in Equation (2.61) are primarily useful for large enough $t$ so that:

$$\frac{\alpha^* \cdot t}{4k_0(\tau(h),h)} - \frac{15 \cdot 2^d \cdot t_{\mathsf{high}}(d)}{2} \geq \frac{\alpha^* \cdot t}{8k_0(\tau(h),h)} \tag{2.66}$$

$$\implies t \geq t''_{\mathsf{high}}(d,h) := \max\left\{\frac{60 \cdot 2^d \cdot k_0(\tau(h),h) \cdot t_{\mathsf{high}}(d)}{\alpha^*}, t'_{\mathsf{high}}(h)\right\}. \tag{2.67}$$

Thus, we can consider any $t \geq t''_{\mathsf{high}}(D,d)$. Note that $t''_{\mathsf{high}}(D,d) = \widetilde{\mathcal{O}}(D^{3/2} \cdot \sqrt{d} \cdot 2^d + D^3)$. Then, we get

$$\sum_{h=d+1}^{D} \sum_{t=1}^{t''_{\mathsf{high}}(D,d)} q_t(h) w^{(h)}_{t,1-Y^*_t} \leq \sum_{t=1}^{t''_{\mathsf{high}}(D,d)} w_{t,1-Y^*_t} \leq t''_{\mathsf{high}}(D,d).$$

Next, we require a lower bound on $h$, above which we apply our bounds. We note that

$$\begin{aligned}\frac{\alpha^* \cdot 2^h}{4} - \frac{15 \cdot 2^d \cdot t_{\mathsf{high}}(d)}{2} - c'(\alpha^*) \cdot \ln\left(\frac{(D-d)}{(\alpha^*)^2\epsilon}\right) &\geq \frac{\alpha^* \cdot 2^h}{8} \\ \implies h \geq h_0(d) &:= \ln C(\alpha^*) + d + \ln(t_{\mathsf{high}}(d)) + \ln\ln\left(\frac{(D-d)}{(\alpha^*)^2\epsilon}\right).\end{aligned}$$

Note that $h_0(d) = \widetilde{\mathcal{O}}\left(d + \ln d + \ln\ln D\right)$, which is useful to keep in mind when we put the pieces together in the final bound.

Then, for all $h \geq h_0(d)$, we get:

$$\sum_{t=t''_{\mathsf{high}}(D,d)+1}^{\infty} q_t(h) w^{(h)}_{t,1-Y^*_t} \leq \underbrace{\sum_{t=1}^{\infty} e^{-\frac{\eta_T \cdot \alpha^* \cdot t}{4k_0(\tau(h),h)}}}_{A} + \underbrace{e^{-\frac{\eta_T \cdot \alpha^* \cdot 2^h}{4}} \cdot \sum_{t=\tau(h)}^{\infty} w^{(h)}_{t,1-Y^*_t}}_{B} \cdot$$

To bound the term $A$, we note, from a similar argument as in the proof of Proposition 2.6.18, that

$$\begin{aligned}A &\leq \frac{e^{-\frac{\eta_T \cdot \alpha^*}{4k_0(\tau(h),h)}}}{1 - e^{-\frac{\eta_T \cdot \alpha^*}{4}}} \\ &\leq \frac{4k_0(\tau(h),h)}{\alpha^* \eta_T},\end{aligned}$$

where the last inequality follows from $e^x \geq 1 + x$.

To bound the term $B$, a direct substitution of the proof argument of Proposition 2.6.18 gives

$$\begin{aligned}\sum_{t=\tau(h)}^{\infty} w^{(h)}_{t,1-Y^*_t} &\leq \frac{1}{\eta^{(h)}_T \alpha^*} \\ &= \frac{\Delta^{(h)}_T}{\alpha^*} \\ &\leq \frac{2^h \ln 2 \cdot t_{\mathsf{high}}(h)}{(\alpha^*)^3},\end{aligned}$$

and so we get

$$\begin{aligned}B &\leq \frac{e^{-\eta_T \cdot \alpha^* \cdot 2^h/8} \cdot 2^h \ln 2 \cdot t_{\mathsf{high}}(h)}{(\alpha^*)^3} \\ &\leq \frac{c(\alpha^*) \cdot t_{\mathsf{high}}(h) \cdot \ln 2}{\eta_T},\end{aligned}$$

where the last inequality follows from the inequality $e^{-x} \leq 1/x$ which holds for any $x > 0$. Thus, the total contribution from the higher-order models becomes

$$\sum_{t=t''_{\mathsf{high}}(D,d)}^{\infty} q_t(h) w_{t,1-Y_t^*}^{(h)} \leq \frac{4k_0(\tau(h),h)}{\alpha^* \eta_T} + \frac{c(\alpha^*) \cdot t_{\mathsf{high}}(h) \cdot \ln 2}{\eta_T} = \mathcal{O}\left(\frac{h^{3/2}}{\eta_T}\right) \tag{2.68}$$

$$\sum_{h=d+1}^{D} \sum_{t=1}^{t''_{\mathsf{high}}(D,d)} q_t(h) w_{t,1-Y_t^*}^{(h)} \leq t''_{\mathsf{high}}(D,d) = \widetilde{\mathcal{O}}(D^{3/2} \cdot \sqrt{d} \cdot 2^d + D^3). \tag{2.69}$$

## Ruling our *bad* lower models

We now turn to the second component of provable model order selection, which involves ruling out lower-order models that incur sizable approximation error. This proof utilizes empirical process theory in an online fashion and is quite similar for both algorithms. It is worth noting that the optimal choice of learning rate as afforded by the design of ValidationOverAdaHedge($D$) allows for a better overall contribution from lower-order models.

### Ruling out lower-order models for SRMOverAdaHedge($D$)

Using Equation (2.44), it is convenient to consider the following upper bound on the quantity $q_t(h)$ for $h < d$:

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \frac{Q_t(h; \eta_t, g_{\mathsf{prop}})}{Q_t(d; \eta_t, g_{\mathsf{prop}})} \tag{2.70a}$$

$$\leq \exp\{-\eta_t(\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t + 2 \cdot 2^d \ln 2 - 2^h \ln 2\} \tag{2.70b}$$

Ruling out lower-order models actually stems from the fact that we can make concrete statements about the sequence's unpredictability (poor approximability) under these models.

The kind of concrete statement that we would like is detailed in the lemma below.

**Lemma 2.6.24.** *Let $h < d$. Consider a sequence $\{x_t\}_{t \geq 1}$ such that we have*

$$(\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t \geq \alpha_{h,d} t \text{ for all } t \geq t_0(h) > 0 \tag{2.71}$$

*for some $\alpha_{h,d} > 0$.*
*Then, we have*

$$\sum_{t=1}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}} \tag{2.72}$$

*where*

$$t'_{\mathsf{low}}(h) = \max\{t_0(h), \frac{2 \cdot 2^d \ln 2}{\eta_T \alpha_{h,d}}\}. \tag{2.73}$$

*Proof.* The condition in Equation (2.71) is essentially the same as the condition on gaps between losses in the original AdaHedge paper [95] used to prove constant regret bounds. We use a similar argument here.

First, we substitute the condition in Equation (2.71) into Equation (2.70b) to get the upper bound

$$\begin{aligned} q_t(h; \eta_t, g_{\mathsf{prop}}) &\leq \exp\{-\eta_t \alpha_{h,d} t + 2 \cdot 2^d \ln 2 - 2^h \ln 2\} \\ &\leq \exp\{-\eta_t \alpha_{h,d} t + 2 \cdot 2^d\} \\ &= \exp\{2 \cdot 2^d \ln 2 - \eta_t \alpha_{h,d} t\} \\ &\leq \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\}. \end{aligned}$$

where the last inequality applies because $\eta_1^T$ is a decreasing sequence. Putting this together with the trivial bound $q_t(h; \eta_t, g_{\mathsf{prop}}) \leq 1$ gives us

$$q_t(h; \eta_t, g_{\mathsf{prop}}) \leq \begin{cases} 1 \text{ for } t \leq t'_{\mathsf{low}}(h) \\ \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\} \text{ for } t > t'_{\mathsf{low}}(h). \end{cases}$$

where we have

$$t'_{\mathsf{low}} = \max\{t_0(h), \frac{2 \cdot 2^d \ln 2}{\eta_T \alpha_{h,d}}\}.$$

From this, using the trivial bound $w_{t,1-Y_t^*}(\eta_t) \leq 1$ we get

$$\begin{aligned} \sum_{t=1}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}(\eta_t) &\leq t'_{\mathsf{low}}(h) + \sum_{t=t'_{\mathsf{low}}(h)+1}^{\infty} \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t\} \\ &\leq t'_{\mathsf{low}}(h) + \exp\{2 \cdot 2^d \ln 2 - \eta_T \alpha_{h,d} t'_{\mathsf{low}}(h)\} \left( \sum_{t=1}^{\infty} e^{-\eta_T \alpha_{h,d} t} \right) \\ &= t'_{\mathsf{low}}(h) + \sum_{t=1}^{\infty} e^{-\eta_T \alpha_{h,d} t} \\ &\leq t'_{\mathsf{low}}(h) + \int_{u=0}^{\infty} e^{-\eta_T \alpha_{h,d} u} du \\ &= t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}} \int_{v=0}^{\infty} e^{-v} dv \\ &= t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}}, \end{aligned}$$

This completes the proof. □

From Lemma 2.6.24, we can clearly bound the contribution of lower-order models to cumulative variance by a constant term. This is because the difference in estimated unpredictability between the right model and the bad lower-order model remains as the number of rounds increase – leading to an exponentially decaying likelihood of selecting the lower-order model. (We do not even need to use any information about whether the online learning algorithm would ensure low regret when selecting a lower-order model, although this is sometimes the case in practice[10].)

It is of interest to characterize when the condition in Equation (2.71) holds. We show that this holds under the $d^{th}$-order stochastic condition on responses, and the iid and Markov assumptions on contexts (Definitions 2.1.2 and 2.1.3). The informal statement is stated below; for a formal statement and proof see Appendix 2.7.

**Lemma 2.6.25** (Informal.)**.** *The condition in Equation* (2.71) *holds for* $\alpha_{h,d} = \frac{\pi^*_h - \pi^*_d}{2}$*, some constant* $c > 0$*, and*

$$t_0(h) = t_{\mathsf{low}}(h) := \frac{32}{\alpha^2_{h,d}} \left( d \cdot 2^h \ln 2 + \ln \left( \frac{64d}{\epsilon \alpha^2_{h,d}} \right) \right). \tag{2.74}$$

*with probability greater than equal to* $(1 - \epsilon)$ *when* $Y_t | X_t$ *satisfies the* $d^{th}$*-order stochastic condition with unpredictability factors* $\{\pi^*_h\}_{h=0}^d$*, and the contexts are one of iid (Definition 2.1.2) or Markov (Definition 2.1.3).*

### Ruling out bad lower models for ValidationOverAdaHedge($D$)

Using Equation (2.45), it is convenient to consider the following upper bound on the quantity $q_t(h)$ for $h < d$:

$$\begin{aligned} q_t(h; \eta_t) &\leq \frac{Q_t(h; \eta_t)}{Q_t(d; \eta_t)} && \text{(2.75a)} \\ &\leq \exp\{-\eta_t((\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t + R^{(h)}_{t-1} - R^{(d)}_{t-1})\} && \text{(2.75b)} \\ &\leq \exp\{-\eta_t((\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t - R^{(d)}_{t-1})\}, && \text{(2.75c)} \end{aligned}$$

where the last inequality follows because $R^{(h)}_{t-1} \geq 0$. Ruling out lower-order models actually stems from the fact that we can make concrete statements about the sequence's unpredictability (poor approximability) under these models.

The kind of concrete statement that we would like is detailed in the lemma below.

**Lemma 2.6.26.** *Let* $h < d$*. Consider a sequence* $\{x_t\}_{t \geq 1}$ *such that we have*

$$(\widehat{\pi}_h(t) - \widehat{\pi}_d(t))t \geq \alpha_{h,d} t \text{ for all } t \geq t_0(h) > 0 \tag{2.76}$$

[10]In fact, models that are close in approximability to the true model will suffer less regret. Ideally, our analysis should consider this nuance, but doing so is likely to be technically challenging because of the data-dependent learning rate.

*for some* $\alpha_{h,d} > 0$.
*Then, we have*

$$\sum_{t=1}^{T} q_t(h; \eta_t, g_{\mathsf{prop}}) w_{t,1-Y_t^*}^{(h)}(\eta_t) \leq t'_{\mathsf{low}}(h) + \frac{1}{\eta_T \alpha_{h,d}} \tag{2.77}$$

*where*

$$t'_{\mathsf{low}}(h) = \max\{t_0(h), \frac{2 \cdot 2^d \ln 2}{\alpha_{h,d}(2\beta^* - 1)^2}\}. \tag{2.78}$$

*Proof.* The condition in Equation (2.76) is essentially the same as the condition on gaps between losses in the original AdaHedge paper [95] used to prove constant regret bounds. We use a similar argument here.

First, we substitute the condition in Equation (2.76) into Equation (2.75c) to get the upper bound

$$q_t(h; \eta_t) \leq \exp\{-\eta_t(\alpha_{h,d} t - R_{t-1}^{(d)})\}.$$

Recall that the regret effectively measures the estimation error under the correct model order $d$. We can get a slightly cruder (in terms of constants) upper bound on the regret from Proposition 2.6.18. In particular, we get

$$R_{t-1}^{(d)} \leq 6 \cdot 2^d \ln 2 \cdot \sqrt{t_{\mathsf{high}}(d)}.$$

Substituting that above, we get for all $t \geq t_0(h)$ that

$$\begin{aligned} q_t(h; \eta_t) &\leq \exp\{-\eta_t(\alpha_{h,d} t - 6 \cdot 2^d \ln 2 \cdot \sqrt{t_{\mathsf{high}}(d)})\} \\ &\leq \begin{cases} q_t(h; \eta_t) \text{ if } t < t'_{\mathsf{low}}(h) \\ \exp\{-\eta_t \alpha_{h,d}(t - t'_{\mathsf{low}}(h))\} \text{ otherwise} \end{cases} \end{aligned}$$

where we have defined

$$t'_{\mathsf{low}}(h) := \max\{t_0(h), \frac{6 \cdot 2^d \ln 2 \cdot \sqrt{t_{\mathsf{high}}(d)}}{\alpha_{h,d}}\}.$$

For $t \geq t'_{\mathsf{low}}(h)$, this gives us

$$\begin{aligned} q_t(h; \eta_t) &\leq \exp\{-\eta_t \alpha_{h,d}(t - t'_{\mathsf{low}}(h))\} \\ &\leq \exp\{-\eta_T \alpha_{h,d}(t - t'_{\mathsf{low}}(h))\} \end{aligned}$$

because $\eta_1^T$ is a decreasing sequence. For $t < t'_{\mathsf{low}}(h)$, we use the trivial bound $w_{t,1-Y_t^*}(\eta_t) \leq 1$. Thus, we get

$$\begin{aligned}
\sum_{t=1}^{T} q_t(h;\eta_t) w_{t,1-Y_t^*}(\eta_t) &\leq \sum_{t=1}^{t'_{\mathsf{low}}(h)} q_t(h;\eta_t) + \sum_{t=t'_{\mathsf{low}}(h)+1}^{\infty} \exp\{-\eta_T \alpha_{h,d}(t - t'_{\mathsf{low}}(h))\} \\
&\leq \sum_{t=1}^{t'_{\mathsf{low}}(h)} q_t(h;\eta_t) + \left(\sum_{t=1}^{\infty} e^{-\eta_T \alpha_{h,d} t}\right) \\
&\leq \sum_{t=1}^{t'_{\mathsf{low}}(h)} q_t(h;\eta_t) + \int_{u=0}^{\infty} e^{-\eta_T \alpha_{h,d} u} du \\
&= \sum_{t=1}^{t'_{\mathsf{low}}(h)} q_t(h;\eta_t) + \frac{1}{\eta_T \alpha_{h,d}} \int_{v=0}^{\infty} e^{-v} dv \\
&= \sum_{t=1}^{t'_{\mathsf{low}}(h)} q_t(h;\eta_t) + \frac{1}{\eta_T \alpha_{h,d}},
\end{aligned}$$

This completes the proof. □

From Lemma 2.6.26, we can clearly bound the contribution of lower-order models to cumulative variance by a constant term. This is because the difference in estimated unpredictability between the right model and the bad lower-order model remains as the number of rounds increase – leading to an exponentially decaying likelihood of selecting the lower-order model. (We do not even need to use any information about whether the online learning algorithm would ensure low regret when selecting a lower-order model, although this is sometimes the case in practice[11].)

It is of interest to characterize when the condition in Equation (2.76) holds. In the previous section, we showed that the condition held for the case of responses being drawn from the $d^{th}$-order stochastic condition, as well as iid or Markovian contexts. The following lemma postulates that the condition holds for the case of periodic contexts (Definition 2.1.4) as well. The informal statement is stated below; for a formal statement and proof see Appendix 2.7.

**Lemma 2.6.27** (Informal.)**.** *The condition in Equation* (2.76) *holds for* $\alpha_{h,d} = \frac{\pi_h^* - \pi_d^*}{2}$*, some constant* $c > 0$*, and*

$$t_0(h) = t_{\mathsf{low}}(h) := \frac{32}{\alpha_{h,d}^2}\left(d \cdot 2^h \ln 2 + \ln\left(\frac{64d}{\epsilon \alpha_{h,d}^2}\right)\right). \tag{2.79}$$

[11]In fact, models that are close in approximability to the true model will suffer less regret. Ideally, our analysis should consider this nuance, but doing so is likely to be technically challenging because of the data-dependent learning rate.

*with probability greater than equal to* $(1-\epsilon)$ *when* $Y_t|X_t$ *satisfies the* $d^{th}$*-order stochastic condition with unpredictability factors* $\{\pi_h^*\}_{h=0}^d$*, and the contexts are periodic (Definition 2.1.4).*

## Putting the pieces together: SRMOVERADAHEDGE($D$)

Now, we put the pieces together to complete the proof of Theorem 2.4.3.

In Section 2.6, we determined the overall contribution to the cumulative variance coming from the vicinity of the true model orders, $h \in \{d, d+1, d+2, d+3\}$. Then, in Section 2.6 + 2.6, we appropriately limited the contribution of lower-order and higher-order models to the cumulative variance. Now, we put together the pieces and characterize cumulative regret to complete the proof of Theorem 2.4.3.

First, we apply Lemma 2.6.13 setting $t_0 = t_{\mathsf{high}}(D)$. Recall that $t_{\mathsf{high}}(D)$ represents the number of appearances of a full context before which we cannot necessarily make statistical guarantees about the predictor. This gives us[12]

$$\Delta_T \leq t_{\mathsf{high}}(D) + \sqrt{V_{T_0(D)}^T \ln 2} + \frac{2}{3}\ln 2 + 2. \tag{2.80}$$

We now proceed to bound the quantity $V_{T_0(D)}^T$. Recall that

$$\begin{aligned} V_{T_0(D)}^T &\leq \sum_{h=0}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-X_t^*}^{(h)} \\ &\leq \underbrace{\sum_{h=0}^{d-1} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-X_t^*}^{(h)}}_{T_1} + \underbrace{\sum_{h=d}^{d+3} \sum_{t=T_0(D)}^{T} w_{t,1-X_t^*}^{(d)}}_{T_2} + \underbrace{\sum_{h=d+4}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-X_t^*}^{(h)}}_{T_3} \end{aligned}$$

We start with summarizing the lower-order model contribution $T_1$. From Lemmas 2.6.24 and 2.7.1, we have

$$\begin{aligned} T_1 &\leq \sum_{h=0}^{d-1} t'_{\mathsf{low}}(h) + \frac{1}{\eta_T}\left(\sum_{h=0}^{d-1} \frac{1}{\alpha_{h,d}}\right) \\ &\leq d t'_{\mathsf{low}}(d-1) + \frac{1}{\eta_T}\left(\sum_{h=0}^{d-1} \frac{1}{\alpha_{h,d}}\right). \end{aligned}$$

[12]Equation (2.80) exposes new conceptual beauty in the umbrella of approaches to varying the learning rate inversely proportional to accumulated regret so far. The only reason a high learning rate does not affect us is because it means that very little regret has been accumulated up to that point. Effectively, $t_0 = t_{\mathsf{high}}(D)$ represents the extent of cumulative mixability the algorithm is willing to tolerate in this regime before carrying out probabilistic stochastic model selection, and is the natural statistical quantity to reflect this.

*Notice that $T_1$ is a constant independent of the horizon $T$ as long as $\eta_T$ does not decay with $T$.*

Next, we move on to the vicinity of the true model order contribution, represented by model orders $\{d, d+1, d+2, d+3\}$. From Lemmas 2.6.16 and 2.6.17, we get

$$\begin{aligned} T_2 &\leq \sum_{h=d}^{d+3} 2^h \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T(2\beta^* - 1)} \right) \\ &\leq 15 \cdot 2^d \left( t_{\mathsf{high}}(d+3) + \frac{1}{\eta_T(2\beta^* - 1)} \right). \end{aligned}$$

*Notice that $T_2$ is roughly what we should expect (up-to constant factors) if we knew the model order exactly.*

Finally, we summarize the higher-order-model contribution $T_3$. From Lemma 2.6.21 and the analysis in Section 2.6, we have

$$\begin{aligned} T_3 &\leq \sum_{h=d+4}^{D} 2^{-h} \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T(2\beta^* - 1)} \right) \\ &= \sum_{h=d+4}^{D} 2^{-h} t_{\mathsf{high}}(h) + \frac{2}{\eta_T(2\beta^* - 1)}. \end{aligned}$$

Recall from Equation (2.50) that

$$\begin{aligned} t_{\mathsf{high}}(h) &= \frac{2}{(2\beta^* - 1)^2} \ln \left( \frac{(D-d) \cdot 2^h}{(2\beta^* - 1)^2 \epsilon} \right) \\ &= \frac{2h}{(2\beta^* - 1)^2} \ln 2 + \frac{2}{(2\beta^* - 1)^2} \ln \left( \frac{(D-d)}{(2\beta^* - 1)^2 \epsilon} \right) \end{aligned}$$

and since $\sum_{h=0}^{\infty} 2^{-h} \leq \sum_{h=0}^{\infty} h \cdot 2^{-h} = 4$, we get

$$\begin{aligned} T_3 &\leq \frac{8}{(2\beta^* - 1)^2} \ln 2 + \frac{8}{(2\beta^* - 1)^2} \ln \left( \frac{(D-d)}{(2\beta^* - 1)^2 \epsilon} \right) + \frac{2}{\eta_T(2\beta^* - 1)} \\ &= 8 t_{\mathsf{high}}(1) + \frac{2}{\eta_T(2\beta^* - 1)}. \end{aligned}$$

*Notice that $T_3$ is a constant that scales only logarithmically in the maximum model order $D$!*

Now combining the three equations for $T_1$,$T_2$ and $T_3$, we get

$$V_{T_0(D)}^T \leq d t'_{\mathsf{low}}(d-1) + 15 \cdot 2^d t_{\mathsf{high}}(d+3) + 8 t_{\mathsf{high}}(1) + \frac{(d+1) \cdot 2^d}{\eta_T \overline{\gamma}},$$

where

$$\frac{1}{\overline{\gamma}} := \frac{1}{d+1}\Big(\sum_{h=0}^{d-1}\frac{1}{\alpha_{h,d}} + \frac{15}{(2\beta^*-1)}\Big)$$

Next, recall from Equation (2.73) that

$$t'_{\mathsf{low}}(d-1) = \max\{t_{\mathsf{low}}(d-1), \frac{2\cdot 2^d}{\eta_T\alpha_{d-1,d}}\} \le t_{\mathsf{low}}(d-1) + \frac{2\cdot 2^d}{\eta_T\alpha_{d-1,d}}$$

using Fact 2.7.10. Substituting this expression gives us

$$V_{T_0(D)}^T \le d\cdot t_{\mathsf{low}}(d-1) + 15\cdot 2^d\cdot t_{\mathsf{high}}(d+3) + 8t_{\mathsf{high}}(1) + \frac{(d+2)\cdot 2^d}{\eta_T\overline{\gamma}}.$$

Next, we use the connection between learning rate and mixability gap from Equation (2.19) to get

$$\begin{aligned}
&\eta_T = \frac{\ln 2}{\Delta_{T-1}} \ge \frac{\ln 2}{\Delta_T}\\
\implies &\frac{1}{\eta_T} \le \frac{\Delta_T}{\ln 2}\\
&\qquad \le \frac{t_{\mathsf{high}}(D)}{\ln 2} + \frac{1}{\ln 2}\left(\sqrt{V_{T_0(D)}^T\ln 2} + \frac{2}{3}\ln 2 + 1\right)
\end{aligned}$$

where in the last step we applied Equation (2.80).
Ultimately, we get the following inequality for $V_{T_0(D)}^T$:

$$\begin{aligned}
V_{T_0(D)}^T \le\ & d\cdot t_{\mathsf{low}}(d-1) + 15\cdot 2^d\cdot t_{\mathsf{high}}(d+3) + 8t_{\mathsf{high}}(1)\\
&+ \frac{(d+2)\cdot 2^d}{\overline{\gamma}}\left(\frac{t_{\mathsf{high}}(D)}{\ln 2} + \frac{1}{\ln 2}\left(\sqrt{V_{T_0(D)}^T\ln 2} + \frac{2}{3}\ln 2 + 1\right)\right).
\end{aligned}$$

Now, we have two cases:

1. $V_{T_0(D)}^T < \frac{1}{4}$.
2. $V_{T_0(D)}^T \ge \frac{1}{4}$, in which case, we get

$$\begin{aligned}
V_{T_0(D)}^T &\le \sqrt{V_{T_0(D)}^T}\Big(2d\cdot t_{\mathsf{low}}(d-1) + 30\cdot 2^d\cdot t_{\mathsf{high}}(d+3) + 16\cdot t_{\mathsf{high}}(1)\\
&\qquad + \frac{2\cdot(d+2)\cdot 2^d\cdot t_{\mathsf{high}}(D)}{\overline{\gamma}\ln 2} + \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}\Big)\\
\implies \sqrt{V_{T_0(D)}^T} &\le 2d\cdot t_{\mathsf{low}}(d-1) + 30\cdot 2^d\cdot t_{\mathsf{high}}(d+3) + 16\cdot t_{\mathsf{high}}(1)\\
&\qquad + \frac{2\cdot(d+2)\cdot 2^d\cdot t_{\mathsf{high}}(D)}{\overline{\gamma}\ln 2} + \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}.
\end{aligned}$$

So, we have bounded the cumulative variance term $V_{T_0(D)}^T$. We now substitute back into Equation (2.80) to get

$$\begin{aligned}\Delta_T \leq t_{\mathsf{high}}(D) + \Big(&2d \cdot t_{\mathsf{low}}(d-1) + 30 \cdot 2^d \cdot t_{\mathsf{high}}(d+3) + 16 \cdot t_{\mathsf{high}}(1) + \\ &+ \frac{2 \cdot (d+2) \cdot 2^d \cdot t_{\mathsf{high}}(D)}{\overline{\gamma} \ln 2} \\ &+ \frac{1}{\sqrt{\ln 2}} + \frac{2}{3} + \frac{1}{\ln 2}\Big)\sqrt{\ln 2} + \frac{2}{3}\ln 2 + 2.\end{aligned}$$

Observe, from this inequality, that the cumulative mixability gap $\Delta_T$ is dominated by three intuitive quantities (other than the constant additive term):

1. $t_{\mathsf{low}}(d-1)$, which represents the number of rounds after which all lower-order models can be conclusively ruled out. The dependence on $t_{\mathsf{low}}(d-1)$ is saying that this much mixability could have accumulated (due to poor approximability) before then.

2. $t_{\mathsf{high}}(D)$, which represents the amount of mixability the algorithm has to accumulate before performing effective higher-order model selection to rule out the over-fitting models[13].

3. $2^d \cdot t_{\mathsf{high}}(d)$, which represents the amount of mixability accumulated by the algorithm *at the right model order*. This is the term in analysis that corresponds to standard best-of-both-worlds analysis over a fixed model order.

Now, we know from Equation (2.50) that $t_{\mathsf{high}}(h) = \frac{2}{(2\beta^*-1)^2}\ln\left(\frac{(D-d)\cdot 2^h}{(2\beta^*-1)^2\epsilon}\right)$ and from Equation (2.74) that $t_{\mathsf{low}}(d-1) = \frac{32d}{\alpha_{d-1,d}^2}\left(d \cdot 2^{d-1}\ln 2 + \ln\left(\frac{64d}{\epsilon\alpha_{d-1,d}^2}\right)\right)$. Substituting these in, we get

$$\Delta_T = \mathcal{O}\left(2^d\left(\frac{d^2}{\alpha_{d-1,d}^2}\ln\left(\frac{d}{\alpha_{d-1,d}^2\epsilon}\right) + \frac{D(d+2)}{\overline{\gamma}(2\beta^*-1)^2}\ln\left(\frac{D}{(2\beta^*-1)^2\epsilon}\right)\right)\right) \tag{2.81}$$

and substituting this into Lemma 2.6.1 gives

$$R_{T,d} = \mathcal{O}\left(2^{2d}\left(\frac{d^2}{\alpha_{d-1,d}^2}\ln\left(\frac{d}{\alpha_{d-1,d}^2\epsilon}\right) + \frac{D(d+2)}{\overline{\gamma}(2\beta^*-1)^2}\ln\left(\frac{D}{(2\beta^*-1)^2\epsilon}\right)\right)\right), \tag{2.82}$$

completing the proof. To highlight the dependence on true model order $d$ and maximum model order $D$ (as is expressed in the informal statement of Theorem 2.4.3), we can hide the constants in terms of parameters and write

$$\begin{aligned}R_{T,d} &= \Delta_T\left(1 + 2^d\right) &(2.83)\\ &= \mathcal{O}\left(2^{2d}\left(D \cdot d \cdot \ln\left(\frac{D}{\epsilon}\right)\right)\right). &(2.84)\end{aligned}$$

[13]It is also possible that the algorithm would not have accumulated even this mixability, and the model selection phase is never reached – however, we never observed this case empirically.

## Putting the pieces together: ValidationOverAdaHedge($D$)

Now, we put the pieces together to complete the proof of Theorem 2.5.2. In Section 2.6, we determined the overall contribution to the cumulative variance coming from the vicinity of the true model orders, $h \in \{d, \dots, h_0(d)\}$. Then, in Sections 2.6 + 2.6, we appropriately limited the contribution of lower-order and higher-order models to the cumulative variance. Now, we put together the pieces and characterize cumulative regret to complete the proof of Theorem 2.5.2.

We start with bounding the quantity $V_{T_0(D)}^T$. Recall that

$$\begin{aligned} V_{T_0(D)}^T &\leq \sum_{h=0}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-Y_t^*}^{(h)} \\ &\leq \underbrace{\sum_{h=0}^{d-1} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-Y_t^*}^{(h)}}_{T_1} + \underbrace{\sum_{h=d}^{h_0(d)-1} \sum_{t=T_0(D)}^{T} w_{t,1-Y_t^*}^{(h)}}_{T_2} + \underbrace{\sum_{h=h_0(d)}^{D} q_t(h) \sum_{t=T_0(D)}^{T} w_{t,1-Y_t^*}^{(h)}}_{T_3}. \end{aligned}$$

We start with summarizing the lower-order model contribution $T_1$. From Lemma 2.6.26, we have

$$\begin{aligned} T_1 &\leq \sum_{h=0}^{d-1} t'_{\mathsf{low}}(h) + \frac{1}{\eta_T} \left( \sum_{h=0}^{d-1} \frac{1}{\alpha_{h,d}} \right) \\ &\leq d t'_{\mathsf{low}}(d-1) + \frac{1}{\eta_T} \left( \sum_{h=0}^{d-1} \frac{1}{\alpha_{h,d}} \right). \end{aligned}$$

*Notice that $T_1$ is a constant independent of the horizon $T$ as long as $\eta_T$ does not decay with $T$.*

Next, we move on to the vicinity of the true model order contribution, represented by model orders $\{d, \dots, h_0(d)\}$. From Lemmas 2.6.19 and 2.6.20, we get

$$\begin{aligned} T_2 &\leq \sum_{h=d}^{h_0(d)} 2^h \left( t_{\mathsf{high}}(h) + \frac{1}{\eta_T(2\beta^* - 1)} \right) \\ &\leq h_0(d) \cdot 2^{h_0(d)} \left( t_{\mathsf{high}}(h_0(d)) + \frac{1}{\eta_T(2\beta^* - 1)} \right). \end{aligned}$$

*Notice that $T_2 = \mathcal{O}(h_0(d)^{3/2} \cdot 2^{h_0(d)}) = \mathcal{O}((d + \ln d + \ln\ln D)^{3/2}(d \ln D) \cdot 2^d)$, which is worse than $\mathcal{O}(2^d)$ by factors of $\mathcal{O}(d^{5/2} \ln D)$.*

Finally, we summarize the higher-order-model contribution $T_3$. From Lemma 2.6.22 and the analysis in Section 2.6, we got

$$\begin{aligned} T_3 &\leq t''_{\mathsf{high}}(D,d) + \sum_{h=h_0(d)}^{D} \frac{4k_0(\tau(h),h)}{\alpha^*\eta_T} + \frac{c(\alpha^*)\cdot t_{\mathsf{high}}(h)\cdot \ln 2}{\eta_T} \\ &= \widetilde{\mathcal{O}}\left(D^{3/2}\cdot\sqrt{d}\cdot 2^d + D^3 + \frac{D^{5/2}}{\eta_T}\right). \end{aligned}$$

*Notice that $T_3$ scales worst-case cubic in $D$, which is far better than the worst-case exponential dependence in $D$ that would be afforded by an algorithm that does not do model selection.*

Now combining the three equations for $T_1$,$T_2$ and $T_3$, we get

$$\begin{aligned} V^T_{T_0(D)} \leq\ & d t'_{\mathsf{low}}(d-1) + \frac{d}{\alpha_{d-1,d}\eta_T} \\ & + h_0(d)\cdot 2^{h_0(d)}\left(t_{\mathsf{high}}(h_0(d)) + \frac{1}{\eta_T(2\beta^*-1)}\right) \\ & + t''_{\mathsf{high}}(D,d) + \frac{4Dk_0(\tau(D),D)}{\alpha^*\eta_T} + \frac{c(\alpha^*)\cdot D t_{\mathsf{high}}(D)\cdot\ln 2}{\eta_T}. \end{aligned}$$

Next, recall from Equation (2.73) that

$$t'_{\mathsf{low}}(d-1) = \max\{t_{\mathsf{low}}(d-1), \frac{2\cdot 2^d}{\eta_T\alpha_{d-1,d}}\} \leq t_{\mathsf{low}}(d-1) + \frac{2\cdot 2^d}{\eta_T\alpha_{d-1,d}}$$

using Fact 2.7.10. Moreover, we have $t''_{\mathsf{high}}(D,d) \leq t'_{\mathsf{high}}(D) + \frac{60\cdot 2^d\cdot k_0(\tau(D),D)\cdot t_{\mathsf{high}}(d)}{\alpha^*}$. Substituting these expressions gives us

$$V^T_{T_0(D)} \leq A + \frac{B}{\eta_T}, \text{ where}$$

$$A = d\cdot t_{\mathsf{low}}(d-1) + h_0(d)\cdot 2^{h_0(d)}\cdot t_{\mathsf{high}}(h_0(d)) + t'_{\mathsf{high}}(D) + \frac{60\cdot 2^d\cdot k_0(\tau(D),D)\cdot t_{\mathsf{high}}(d)}{\alpha^*}$$

$$B = \frac{d}{\alpha_{d-1,d}} + \frac{h_0(d)\cdot 2^{h_0(d)} + 4D\cdot k_0(\tau(D),D)}{\alpha^*} + c(\alpha^*)\cdot D t_{\mathsf{high}}(D)\cdot\ln 2.$$

Next, we use the connection between learning rate and mixability gap from Equation (2.19) to get

$$\begin{aligned} \eta_T &= \frac{\ln 2}{\Delta_{T-1}} \geq \frac{\ln 2}{\Delta_T} \\ \implies \frac{1}{\eta_T} &\leq \frac{\Delta_T}{\ln 2} \\ &\leq \frac{1}{\ln 2}\left(\sqrt{V^T_{T_0(D)}\ln 2} + \frac{2}{3}\ln 2 + 1\right). \end{aligned}$$

Thus, we get the following inequality for $V^T_{T_0(D)}$:

$$V^T_{T_0(D)} \leq A + B\sqrt{\frac{V^T_{T_0(D)}}{\ln 2}} + \frac{5B}{3}.$$

Now, we have two cases:

1. $V^T_{T_0(D)} < \frac{1}{4}$, in which case we are done.
2. $V^T_{T_0(D)} \geq \frac{1}{4}$, in which case, we get

$$\begin{aligned} V^T_{T_0(D)} &\leq \sqrt{V^T_{T_0(D)}}\left(A + \frac{8B}{3}\right) \\ \implies \sqrt{V^T_{T_0(D)}} &\leq A + \frac{8B}{3} \end{aligned}$$

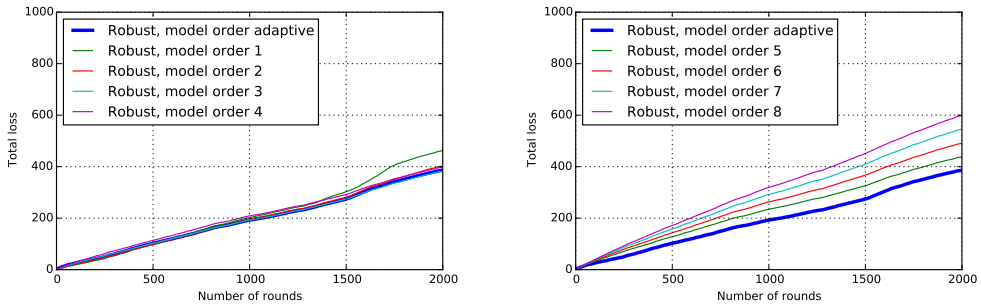
and thus Equation (2.39) together with Proposition 2.6.18 gives us

$$\begin{aligned} R_{T,d} &\leq \left(A + \frac{8B}{3}\right)(1 + \ln D) + \mathcal{O}(2^d \cdot t_{\mathsf{high}}(d)) \\ &= \mathcal{O}\Big(d \cdot t_{\mathsf{low}}(d-1) + h_0(d) \cdot 2^{h_0(d)} \cdot t_{\mathsf{high}}(h_0(d)) + t'_{\mathsf{high}}(D) \\ &+ \frac{60 \cdot 2^d \cdot k_0(\tau(D), D) \cdot t_{\mathsf{high}}(d)}{\alpha^*} \\ &+ \frac{d}{\alpha_{d-1,d}} + \frac{h_0(d) \cdot 2^{h_0(d)} + 4D \cdot k_0(\tau(D), D)}{\alpha^*} + c(\alpha^*) \cdot D t_{\mathsf{high}}(D) \cdot \ln 2\Big) \\ &= \mathcal{O}\Big(d^2 \cdot 2^d + d \ln\left(\frac{d}{\epsilon}\right) + d^2\left(\ln\left(\frac{D}{\epsilon}\right)\right)^2 \cdot 2^d + D^3 \ln\left(\frac{D^2}{\epsilon}\ln\left(\frac{D}{\epsilon}\right)\right) \\ &+ D^{3/2} \cdot d \cdot 2^d \cdot \left(\ln\left(\frac{D}{\epsilon}\right)\right)^{3/2}\Big) \\ &= \widetilde{\mathcal{O}}\left(d^2 \cdot 2^d + d \ln d + d^2 (\ln D)^2 + D^3 \ln(D^2 \ln D) + D^{3/2} \cdot d \cdot 2^d (\ln D)^{3/2}\right), \end{aligned}$$

and this completes the proof of Theorem 2.5.2.

## 2.7 Future work

In this chapter, we motivated online model selection in full-information environments as an important goal in the context of the broader goal of adaptively maximizing reward (minimizing loss) in an unknown environment that could be *stochastic or adversarial*. This methodology does not yet explicitly consider competitive or cooperative environments for adaptivity.




(a) Total loss as a function of $T$ compared to lower-model orders.

(b) Total loss as a function of $T$ compared to higher-model orders.

Figure 2.4: Comparison of model-adaptive SRMOVERADAHEDGE($D$) with uniform-prior SRMOVERADAHEDGE($d$) for fixed model orders on a HMM with slowly transitioning states. Figures from [88].

However, full-information online model selection even between stochastic and adversarial environments is a practically important objective, and has several independently interesting future directions. In particular, the three essential ingredients for practical applicability — computational efficiency beyond binary prediction, effectiveness in mis-specified models, and applicability beyond traditional statistical learning — are not yet fully established. We now discuss future directions along these lines.

## Mis-specification

As mentioned in Section 2.3, the paradigm of data-driven model selection is most broadly applied in *mis-specified* stochastic environments: that is, the data is stochastic, but not realizable by any of the model orders. It is of substantial interest to obtain the corresponding online model selection results in a mis-specified environment. Indeed, meaningful theoretical guarantees do not even exist for the purely stochastic case in online learning. This is because, unlike in the $d^{th}$-order realizable stochastic environment, none of the benchmarks for regret correspond to an optimal guarantee on reward! A meaningful guarantee in mis-specified environments would constitute an upper bound on *time-averaged prediction error* rather than any fixed notion of regret.

We motivate the pursuit of this guarantee through a preliminary empirical evaluation of SRMOVERADAHEDGE($D$) in a mis-specified environment. Our representative example of mis-specification is that of a hidden Markov model (HMM) with the following parameters:

$$\text{Hidden state evolution } W_{t+1} \sim \text{Ber}\left(|W_t - 0.001|\right)$$
$$Y_t|W_t = 0 \sim \text{Ber}(0.2) \text{ and } Y_t|W_t = 1 \sim \text{Ber}(0.9).$$

This is an interesting example of a HMM with very slowly transitioning hidden states, that has long-range dependencies. From the simulation results in Figure 2.4, it appears that

the best model fit is of order 3 or 4; we observe that our adaptive algorithm naturally tracks the performance of such a model fit in this example as well. If we do not select models of roughly this order, we either over-fit or under-fit as seen in the simulations. It is worth noting that depending on the parameters of the HMM, different model orders could be considered as optimal fits for increasing numbers of rounds; it is notable that SRMOVERADAHEDGE($D$) adapts to a suitable model order for different choices of parameters.

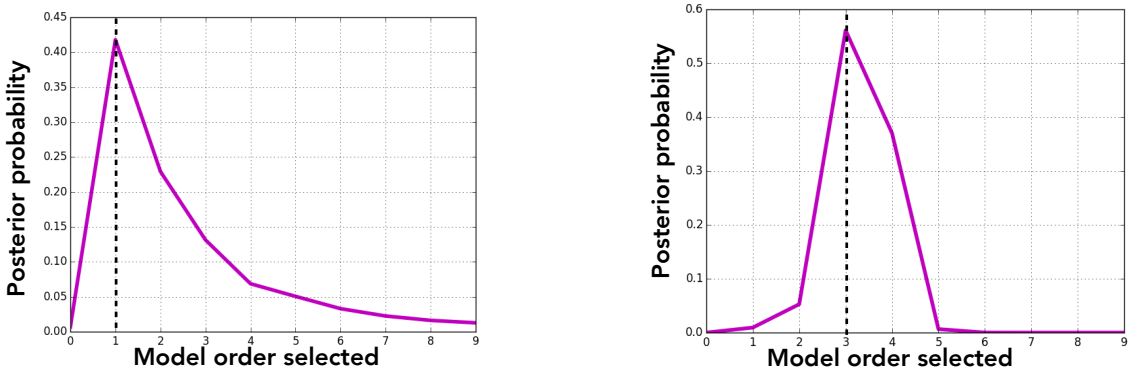



Figure 2.5: Posterior probabilities of model order selected by SRMOVERADAHEDGE($D$) at different numbers of rounds, $t = 250$ and $t = 1500$, when played against a "sticky" HMM. We see that the algorithm is most likely to select increasing model orders as more rounds of the game are played. Figures from [88].

Moreover, Figure 2.5 shows that the model order most likely to be selected by SRMOVERADAHEDGE($D$) increases with the number of rounds $T$; eventually, the largest model order would be selected. These pictures show that a direct analysis of the average prediction error is necessary for any non-trivial guarantee; any regret-based analysis would necessarily scale as $\mathcal{O}(2^D)$, which is extremely pessimistic for this class of models. This is an interesting and non-trivial direction for future work. On one hand, the algorithms SRMOVERADAHEDGE($D$) and VALIDATIONOVERADAHEDGE($D$) make explicit connections to the statistical methodologies of SRM and cross-validation, both of which are heavily used in mis-specified environments. Their proven success in the realizable case thus bodes well for an eventual guarantee under mis-specification. On the other hand, obtaining simultaneously adversarial and stochastic bounds on average prediction error, rather than regret, remains a challenge. Doing this likely requires a new "second-order-bounding" technique, as the current second-order bounding techniques [94, 95, 98, 100] are critically framed in the context of regret.

## Computational efficiency

The question of computational efficiency has critically under-lied the novelty of our results as restricted to the binary contextual prediction setup with $d^{th}$-order realizable stochastic data. Achieving efficiency in two-fold adaptive algorithms between stochastic and adversarial model selection is a significant challenge: even the most sophisticated algorithms like AdaNormalHedge and Squint suffer from double-exponential in $D$ complexity (in the case of Squint, primarily because it departs significantly from the exponential weighting framework), or highly non-trivial model selection analysis (in the case of the sleeping-experts implementation of AdaNormalHedge). Our algorithms, while admitting a more involved analysis for stochastic settings, are efficient per iteration and thus simulate-able.

Obtaining computationally efficient adaptive online learning algorithms remains a significant, and largely open, challenge. The primary difficulty lies in obtaining the adversarial guarantee: in a generic online supervised learning setup with function class $\{\mathcal{F}_h\}_{h=0}^{D}$, the computational complexity-per-iteration of most standard online learning algorithms (e.g. Hedge) scales as $\mathcal{O}(|\mathcal{F}_D|)$, which is far worse than the complexity of purely stochastic approaches based on empirical risk minimization. Until recently, efficient approaches took advantage of special structure in the function classes[14], and were not known more generally. More recently, there has been promising progress in the online learning community in understanding the computational trade-offs as well as explicit efficient algorithm design. On one hand, there do exist worst-case function classes for which a $\mathcal{O}(\sqrt{|\mathcal{F}_D|})$ complexity-per-iteration is unavoidable [131]. On the other hand, it was shown recently [39] that certain function classes afford *oracle-efficient* online learning, i.e. online learning algorithms with the optimal worst-case guarantee whose computational complexity per iteration is equivalent to the computational complexity per iteration of empirical risk minimization. This oracle-efficient algorithms uses classical perturbation-based techniques in online learning [80, 132] with shared randomness across multiple actions/functions to achieve an efficient implementation. In fact, these algorithms underlie the design of repeated auctions with worst-case guarantees against a stream of (myopic) bidders.

This recent progress provides a possible road-map for recovering adaptive guarantees in the online supervised learning setup while retaining computational efficiency. This goal requires non-trivial work in itself, as the analysis of the oracle-efficient framework that was developed in [39] was restricted to the "worst-case" choice of learning rate $\eta_t = 1/\sqrt{t}$. It would be very interesting to understand whether oracle-efficient approaches can be made to work with data-adaptive learning rates.

As a final remark, we note that meta-experts-based approaches to model selection, like ValidationOverAdaHedge($D$), only incur a computational overhead of $\mathcal{O}(D)$ over the complexity of the base algorithms; thus, model selection by itself adds minimal computational complexity even in a generic supervised learning framework. The more difficult aspect of computationally efficient adaptivity is between adversarial and stochastic environments.

[14] As in the tree-experts case, which was studied in this chapter. The computationally efficient implementations that we use were first proposed for the worst case by [109, 130].

## Online model selection beyond traditional statistical learning

Finally, as already mentioned in Section 2.3, all existing theoretical guarantees for online model selection algorithms are for the traditional complexity hierarchy, where *we only wish to select more complex models when we have enough data.* While this is true of the worst-case generalization bounds, such bounds are not always indicative of true model performance, as evidenced by the modern success of over-parameterized models [124, 125]. It would be interesting to see whether online model selection is successful under these modern environments, whose offline guarantees are still not well-understood (see Chapter 6 for further delving into this point). While the very principle of SRM is antithetical to a possible desire to select more complex models early, it would be especially interesting to see whether data-driven validation is more successful. Viewed alternatively, the success of over-parameterized models already suggests that a data-driven validation approach could be the better approach to online model selection in real-world settings.

From a theoretical standpoint, the last few years have seen a flurry of activity in analyzing the algorithmic generalization error of over-parameterized models [133–136]. These analyses decompose the algorithmic generalization error along quantities that are starkly different from the standard quantities of bias (approximation error) and variance (estimation error). Our hope is that these perspectives can be leveraged to show that adaptive online validation, with its reliance on precisely the algorithmic generalization error, continues to provably select the optimal model even in these non-standard, modern ML regimes.

## Appendix: Stochastic model selection guarantees

In the analyses of both SRMOverAdaHedge($D$) and ValidationOverAdaHedge($D$), we required the estimates of approximation error to concentrate sufficiently quickly – in particular, we required that the difference in approximability between a higher-order model and lower-order model not look too small – in order to rule out lower-order models when appropriate. This was encapsulated in Lemmas 2.6.24 and 2.6.26. It is therefore of interest to understand when the conditions in Equations (2.71) and (2.76) holds, and in particular, characterize $t'_{\mathsf{low}}(h)$ in both cases.

Recall the definition of asymptotic unpredictability for the cases of iid contexts (Definition 2.1.2) and Markov process (Definition 2.1.3), i.e.

$$\pi^*_h := \sum_{x(h)\in\mathcal{X}^h} Q^*(x(h)) \left[1 - \max_{y\in\mathcal{X}}\{P^*(y|x(h))\}\right],$$

and for periodic contexts (Definition 2.1.4), i.e.

$$\pi^*_h := \frac{1}{2^h} \sum_{x(h)\in\mathcal{X}^h} \left[1 - \max_{y\in\mathcal{X}}\{P^*(y|x(h))\}\right].$$

Under all three cases, we have $\pi^*_h = \pi^*_d$ for $h > d$; and $\pi^*_h > \pi^*_d$ for $h < d$. It is also well-known [110] that under all three cases,

$$\widehat{\pi}_h(t) \xrightarrow{\text{prob.}} \pi^*_h \text{ for all } h \in \{0, 1, \ldots, D\}.$$

So the intuition is that for a large enough value of $t$, we should also start to see a *strict* decaying in the estimated unpredictability as $h$ increases to $d$ – and we should be able to rule out the poorly performing $h^{th}$-order models when $h < d$. That is,

$$\widehat{\pi}_h(t) > \widehat{\pi}_d(t) \text{ for all } h < d.$$

In this section, we show that this condition holds for all three cases we have considered. Proving concentration bounds for the cases of iid and periodic contexts is straightforward by Hoeffding's inequality. To prove concentration bounds for the Markov case, we invoke results from the information theory community on transportation-cost inequalities, used to establish concentration of measure for weakly dependent random variables.

**Sufficient condition for concentration of estimate of approximability**

We start by expressing our estimate for approximability for the $h^{th}$-order model, $\widehat{\pi}_h(t)$, as a minimum of $|\mathcal{F}_h|$ Lipschitz functions as below:

$$t\widehat{\pi}_h(t) = \min_{\mathbf{f} \in \mathcal{F}_h} \left\{ f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) \right\} \text{ where}$$

$$\begin{aligned} f_{(h)}(\{(X_s, Y_s)\}_{s=1}^t; \mathbf{f}) &:= \sum_{s=1}^{t} \mathbb{I}[Y_s \neq \mathbf{f}(X_s(h))] \\ &= \sum_{s=1}^{t} Z_s \end{aligned}$$

where $Z_s = \mathbb{I}[Y_s \neq \mathbf{f}(X_s(h))]$. Note that for the cases of iid and periodic contexts, the random variables $\{Z_s\}_{s=1}^t$ are independent and take values in $\{0, 1\}$.

We now state the following technical lemma:

**Lemma 2.7.1.** *Let the following condition hold for every $f \in \mathcal{F}^h$, $t \geq h + 1$ and $\delta > 0$:*

$$\Pr\left[|f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}) - \mathbb{E}\left[f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f})\right]| > t\delta\right] \leq 2\exp\{-ct\delta^2\} \tag{2.85}$$

*for some constant $c > 0$ (that can depend linearly on $d$ as well as $h$).*
*Then, the conditions in Equations* (2.71) *and* (2.76) *hold for $\alpha_{h,d} = \frac{\pi^*_h - \pi^*_d}{2}$ and*

$$t_0(h) = t_{\mathsf{low}}(h) := \frac{32}{c \cdot \alpha^2_{h,d}} \left( d \cdot 2^h \ln 2 + \ln\left(\frac{64d}{c \cdot \epsilon\alpha^2_{h,d}}\right)\right).$$

*with probability greater than equal to $(1 - \epsilon)$.*

*Proof.* Observe that $\widehat{\pi}_h(t)$ itself is not an unbiased estimate of $\pi^*_h$. But, for all three contextual models, it is easy to show that

$$\mathbb{E}\left[t\widehat{\pi}_h(t)\right] = \mathbb{E}\left[\min_{\mathbf{f}\in\mathcal{F}_h} f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f})\right] \leq \mathbb{E}\left[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f}^*_h)\right] = t\pi^*_h$$

for all $\mathbf{f}\in\mathcal{F}_h$. The upper tail bound therefore follows easily – from Equation (2.85), we have

$$\begin{aligned}\Pr\left[t\widehat{\pi}_h(t) - t\pi^*_h > \delta t\right] &\leq \Pr\left[f_{(h)}((X_s,Y_s)_{s=1}^t;\mathbf{f}^*_h) - \mathbb{E}\left[f_{(h)}((X_s,Y_s)_{s=1}^t;\mathbf{f}^*_h)\right] > \delta t\right] \\ &\leq \exp\{-ct\delta^2\}.\end{aligned}$$

To get the lower tail bound, we need to use the union bound.

$$\begin{aligned}\Pr\left[t\pi^*_h - t\widehat{\pi}_h(t) > \delta t\right] &= \Pr\left[t\widehat{\pi}_h(t) < t\pi^*_h - \delta t\right] \\ &\leq \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\left[f_{(h)}((X_s,Y_s)_{s=1}^t;\mathbf{f}) < t\pi^*_h - \delta t\right] \\ &= \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\Big[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f}) - \mathbb{E}\left[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f})\right] < \\ &\qquad t\pi^*_h - \mathbb{E}\left[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f})\right] - \delta t\Big] \\ &\leq \sum_{\mathbf{f}\in\mathcal{F}^h} \Pr\left[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f}) - \mathbb{E}\left[f_{(h)}(\{(X_s,Y_s)\}_{s=1}^t;\mathbf{f})\right] < -\delta t\right] \\ &\leq 2^{2^h}\exp\{-ct\delta^2\}.\end{aligned}$$

Next, we plug in $\delta = \frac{\alpha_{h,d}}{2} = \frac{\pi^*_h - \pi^*_d}{4}$ and re-apply the union bound to get

$$\begin{aligned}&\Pr\left[\cup_{h=0}^{d-1}\{(\widehat{\pi}_h(t) - \widehat{\pi}_d(t)) \leq \alpha_{h,d} \text{ for some } t \geq t_0(h)\}\right] \\ &\leq \Pr\left[\cup_{h=0}^{d-1}\{\pi^*_h - \widehat{\pi}_h(t) \leq \frac{\alpha_{h,d}}{2}\} \cup \{\widehat{\pi}_d(t) - \pi^*_d \leq \frac{\alpha_{h,d}}{2}\} \text{ for some } t \geq t_0(h)\right] \\ &\leq \sum_{h=0}^{d-1}\sum_{t\geq t_0(h)} \Pr\left[\pi^*_h - \widehat{\pi}_h(t) \leq \frac{\alpha_{h,d}}{2}\right] + \Pr\left[\widehat{\pi}_d(t) - \pi^*_d \leq \frac{\alpha_{h,d}}{2}\right] \\ &\leq \sum_{h=0}^{d-1} \frac{32\cdot 2^{2^h}}{c\cdot\alpha^2_{h,d}} e^{-\frac{c\cdot\alpha^2_{h,d}t_0(h)}{32}} + \frac{32}{c\cdot\alpha^2_{h,d}} e^{-\frac{c\cdot\alpha^2_{h,d}t_0(h)}{32}} \\ &\leq \epsilon/2 \text{ when}\end{aligned}$$

$$t_0(h) \geq t_{\mathsf{low}}(h) := \frac{32}{c\cdot\alpha^2_{h,d}}\left(d\cdot 2^h\ln 2 + \ln\left(\frac{64d}{c\cdot\epsilon\alpha^2_{h,d}}\right)\right).$$

This completes our proof. □

Clearly, the condition in Equation (2.85) holds for the cases of iid contexts as well as periodic contexts. In the case of iid contexts the $Z_s$'s are iid, and in the case of periodic contexts the $Z_s$'s are *independent* for any value of $X_1$. In both cases, we can apply the Hoeffding bound directly to get Equation (2.85).

We now proceed to show that Equation (2.85)also holds for finite-memory Markov models using the transportation cost method.

### Concentration for finite-memory Markov models

The concentration of sums of random variables to its mean is a classical topic in statistics and probability theory. The special case when the random variables are iid is well-understood. Intuitively, a Markov process that is well-approximated by an iid process should follow similar concentration laws – the transportation cost argument uses this to prove concentration bounds on sums of random variables following a Markov process.

Formally, this notion of approximability by a product distribution is captured by the *contractivity* of a Markov process which we define below.

**Definition 2.7.2.** *1. For a $d^{th}$-memory Markov process on $Y_1, \ldots, Y_t$ on state space $\mathcal{X}$, we define the **aggregated state** at time $s$ by*

$$W_s = (W_{s,1}, \ldots, W_{s,d}) = (Y_{(s-1)d+1}, Y_{(s-1)d+2}, \ldots, Y_{sd}). \tag{2.86}$$

*Then, clearly, for any $s \geq 1$, we have $W_s \perp W_{s-2} | W_{s-1}$ and so that the states $\{W_s \in \mathcal{X}^d\}_{s \geq 1}$ satisfy the 1-memory Markov property. Explicitly, we have for any $w, w' \in \mathcal{X}^d$,*

$$\mathbb{P}(w) = P_{(d)}(w)$$
$$\mathbb{P}_{-1}(w'|w) = \prod_{i=1}^{d} P(w'_i | (w_i, \ldots, w_d, \ldots, w'_{i-1})).$$

*2. A $d^{th}$-memory Markov process on $Y_1, \ldots, Y_t$ is $\gamma$-**contractive** if for every $w, w' \in \mathcal{X}^d$, we have*

$$\|\mathbb{P}_{-1}(.|w) - \mathbb{P}_{-1}(.|w')\|_{TV} \leq \gamma < 1. \tag{2.87}$$

The *transportation cost method* can then be easily leveraged to show that Equation (2.85) holds, as we show in the following minor lemma.

**Lemma 2.7.3.** *Equation* (2.85) *holds for a $d^{th}$-memory $\gamma$-contractive Markov process with constant $c = \frac{(1-\gamma)}{d}$.*

*Proof.* We invoke Marton's concentration theorem for $\gamma$-contractive Markov processes as described in Theorem 2.7.7 (details about the transportation cost method are provided in Section 2.7).

We recall the definition of functions

$$f_{(h)}((X_s, Y_s)_{s=1}^t; \mathbf{f}) = f_{(h)}((Y_s)_{s=1}^t; \mathbf{f}) := \sum_{s=1}^{t} Z_s$$

where $Z_s = \mathbb{I}[Y_s \neq \mathbf{f}(Y_s(h))]$. To obtain concentration bounds from Theorem 2.7.7, it remains to rewrite $f((Y_s)_{s=1}^t; \mathbf{f})$ as a sum of indicator functions on $\{W_s\}_{s\geq 1}$ and show the Lipschitz property.

Let $t = \lfloor t/d \rfloor + k$ for some $k \in \{0, \ldots, d-1\}$. Then, we can write (with a slight abuse of notation) for any $h \in \{0, 1, \ldots, d\}$,

$$f_{(h)}(Y^t; \mathbf{f}) := df_{(h)}(\{W_s\}_{s=1}^{\lceil t/d \rceil}; \mathbf{f}) = \Big( \sum_{s=1}^{\lfloor t/d \rfloor} \sum_{i=h+1}^{d} \mathbb{I}[W_{s,i} \neq f(W_{s,i-h}, \ldots, W_{s,i-1})] + \sum_{i=1}^{h} \mathbb{I}[W_{s+1,i} \neq f(W_{s,d-(h-i)}, \ldots, W_{s,d}, W_{s+1,1}, \ldots, W_{s+1,i-1})] \Big)$$

Now, it's easy to verify that that a change in $W_s$ will only affect two terms in the sum over $\lceil t/d \rceil$ terms, and some simple algebra tell us that the Lipschitz constant of function $f_{(h)}$ is at most 2. We now apply Theorem 2.7.7 directly to get

$$\Pr\left[ |f_{(h)}(W^{\lceil \frac{t}{d} \rceil}; \mathbf{f}_d) - \mathbb{E}\left[ f_{(h}(W^{\lceil \frac{t}{d} \rceil}; \mathbf{f}_d) \right] | > \delta \lceil \frac{t}{d} \rceil \right] \leq 2 \exp\{ -\frac{2\delta^2 (1-\gamma)^2 t}{4d} \} \tag{2.88}$$

and so, we finally get

$$\Pr\left[ f_{(h)}(Y^t; \mathbf{f}_d) - \mathbb{E}\left[ f_{(h)}(Y^t; \mathbf{f}_d) \right] > (t-h-1)\delta \right] \leq \exp\{ -\frac{\delta^2 (1-\gamma)^2 (t-h-1)}{2d} \} \text{ and}$$

$$\Pr\left[ \mathbb{E}\left[ f_{(h)}(Y^t; \mathbf{f}_d) \right] - f_{(h)}(Y^t; f) > (t-h-1)\delta \right] \leq \exp\{ -\frac{\delta^2 (1-\gamma)^2 (t-h-1)}{2d} \},$$

completing the proof of Lemma 2.7.3.

□

**Technical details about the transportation cost method**

Let $t > 0$. Consider a metric space $\mathcal{X}^t$ with metric $\rho$.

We will consider functions of the form $f : \mathcal{X}^t \to \mathbb{R}$ that are Lipschitz with respect to metric $\rho$; that is, there exists some $L > 0$ such that

$$|f(X_1^t) - f(X_2^t)| \leq L\rho(X_1^t, X_2^t).$$

We denote the Lipschitz constant of the function by $\|f\|_{\mathsf{Lip}}$.

Now we define a useful notion of distance called the Wasserstein distance.

**Definition 2.7.4.** *The Wasserstein distance between distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}^t$ with respect to metric $\rho$ is defined as*

$$W_\rho(\mathbb{P},\mathbb{Q}) = \sup_{f:\|f\|_{\mathsf{Lip}}\le 1} \int f d\mathbb{P} - f d\mathbb{Q} = \inf_{\mathbb{M}\ couples\ \mathbb{P}\ and\ \mathbb{Q}\ on\ (X_1^t, X_2^t)} \mathbb{E}\left[\rho(X_1^t, X_2^t)\right]$$

We will consider $X^t \in \mathcal{X}^t$ be distributed according to $\mathbb{P}$. For a function $f$ such that $\|f\|_{\mathsf{Lip}} = L$, we care about the concentration of the quantity $f(X^t)$ around its mean, $\mathbb{E}\left[f(X^t)\right]$, as a function of $t$.

In our case, $\mathcal{X}^t = \{0,1\}^t$ is finite. We consider the *additive Hamming metric*

$$\rho(X_1^t, X_2^t) := \sum_{s=1}^{t} \mathbb{I}[X_{1,s} \neq X_{2,s}]. \tag{2.89}$$

(For the special case of $t = 1$ the Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ corresponding to this metric is the total variation distance, denoted by $\|\mathbb{P} - \mathbb{Q}\|_{TV}$.

Our basic ingredient is a transportation cost inequality, which we define below.

**Definition 2.7.5.** *We say that the distribution $\mathbb{P}$ satisfies a transportation cost inequality if, for every distribution $\mathbb{Q}$, we have*

$$W_\rho(\mathbb{P},\mathbb{Q}) \le \sqrt{2\gamma D(\mathbb{Q} \parallel \mathbb{P})} \tag{2.90}$$

Marton showed [137] that a transportation inequality on the underlying distribution $\mathbb{P}$ on $X^t$ implied nice concentration bounds on $f(X^t)$ around its mean, when $f(\cdot)$ is Lipschitz with respect to the metric $\rho$. This technique is powerful because we can establish transportation cost inequalities for a much broader class of distributions $\mathbb{P}$ than just product distributions; in particular, we can handle weak dependencies. In the special case of the Wasserstein metric corresponding to *total variation distance*, the classical Pinsker's inequality is a special case of the transportation cost inequality (2.90). It turns out we can adapt Pinsker's inequality together with the chain rule on KL-divergence to prove a transportation cost inequality on the additive Hamming distance for product distributions [138]. We can also do this more generally for the case where $\mathbb{P}$ is a Markov distribution on $\mathcal{X}^t$, provided the Markov chain satisfies an important *contractivity condition*. Consider the Markov process with stationary distribution $\mathbb{P}_1(.)$, and transition probabilities $\mathbb{P}_{-1}(.|x)$ for all $x \in \mathcal{X}$. We again define $\gamma$-contractivity for a general-state-space Markov process below.

**Definition 2.7.6.** *A Markov chain is $\gamma$-contractive if for every two states $x, x' \in \mathcal{X}$, we have*

$$\|\mathbb{P}_{-1}(.|x) - \mathbb{P}_{-1}(.|x')\|_{TV} \le \gamma < 1. \tag{2.91}$$

Under this condition, the Markov distribution satisfies a transportation cost inequality, as shown by the following theorem.

**Theorem 2.7.7** ( [137]). *Let $\mathbb{P}$ be a Markov distribution on $\mathcal{X}^t$ that satisfies Equation* (2.91) *with parameter $\gamma < 1$. Then, we have*

$$W_\rho(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{1-\gamma}\sqrt{\frac{t}{2}D(\mathbb{Q} \parallel \mathbb{P})} \tag{2.92}$$

*This directly implies a concentration bound of the form*

$$\Pr[|f(X^t) - \mathbb{E}\left[f(X^t)\right]| > \delta t] \leq 2\exp\{-\frac{2\delta^2(1-\gamma)^2 t}{L^2}\} \tag{2.93}$$

## Appendix: Algorithmic benefits of SRMOverAdaHedge($D$)

In this section, we expound on the algorithmic benefits of SRMOverAdaHedge($D$) equipped with prior function $g(\cdot)$, as well as ValidationOverAdaHedge($D$): in particular, we formally show the reduced computational complexity of the algorithm, and the equivalence of the computationally efficient update in Equation (2.22a) and the computationally naive update in Equation (2.18). The equivalence was originally proved for the multiplicative weights algorithm with a fixed learning rate [109]: here, we generalize the argument to include the family of exponential-weights updates with a time-varying, data-dependent learning rate.

**Proposition 2.7.8.** *The run-time of* SRMOverAdaHedge($D$) *per prediction round is $\mathcal{O}(2^D)$.*

*Proof.* Consider round $t$ of prediction. To carry out the efficient update in Equation (2.22a), we need to visit every node in the path of the context $X_t$. Since the full context is of length $D$, the update runs in $\mathcal{O}(D)$. To perform the prediction, we must calculate the probability distribution $\mathbf{w}_t$, which has 2 entries. To calculate $\mathbf{w}_t$, we must visit every node in the single complete height $D$ tree to access the cumulative loss vectors $\{\mathbf{L}_{x(D),t}\}_{x(D)\in\mathcal{X}^D}$.

Since there are $2^D$ such loss vectors (i.e. $2^D$ nodes to visit), this operation takes $\mathcal{O}(2^D)$ time. For a general prior, these cumulative contextual losses are accessed for every value of $h \in \{0, 1, \ldots, D\}$. Thus, the total computational complexity of performing an update is

$$\sum_{h=0}^{D} 2^h = 2^{D+1} - 1 \in \mathcal{O}(2^D).$$

After performing prediction and receiving loss feedback, we need to access all these nodes again and update the cumulative losses. By a similar argument as above, this is also a $\mathcal{O}(2^D)$ operation. Therefore, the total computational complexity per round is $\mathcal{O}(2^D)$. □

**Proposition 2.7.9.** *The run-time of* ValidationOverAdaHedge($D$) *per prediction round is $\mathcal{O}(D)$.*

*Proof.* In the meta-algorithm ValidationOverAdaHedge($D$), we use the efficient implementations of the base algorithms $\mathcal{A}_h = \text{AdaHedge}(h)$ for $h = 0, \ldots, D$. Then, it suffices to show that each of the algorithms AdaHedge($h$) has run-time $\mathcal{O}(1)$ per round, for every $h = 0, \ldots, D$. This is because the additional step of storing the loss incurred by AdaHedge($h$) adds $\mathcal{O}(1)$ complexity per round, and then the meta-algorithm adds complexity $\mathcal{O}(D)$ per round as it uses exactly $D$ meta-experts.

Thus, we show that the complexity of Equation (2.22a), i.e. AdaHedge($h$) with uniform prior $g_{\mathsf{unif}}(\cdot)$, is $\mathcal{O}(1)$ per iteration. To do this, recall that under the uniform prior $g_{\mathsf{unif}}(\cdot)$, Equation (2.22a) becomes

$$\mathbf{w}_t^{(h)} \propto e^{-\eta_t^{(h)} \mathbf{L}_{X_t(h),t}}. \tag{2.94}$$

Thus, with knowledge of the learning rate $\eta_t^{(h)}$, the complexity of this update is clearly $\mathcal{O}(1)$, as it only needs access to one context vector $X_t(h) \in \mathcal{X}^h$. (Note that this inference can only be made for the special case of $g(\cdot) = g_{\mathsf{unif}}(\cdot)$ — for non-uniform priors, as we noted in the proof of Proposition 2.7.8, all the contexts $x(D) \in \mathcal{X}^D$ need to be accessed.)

Therefore, it remains to evaluate the complexity per iteration of updating the learning rate, $\eta_t^{(h)}$. Recall from Equation (2.11) that we have

$$\eta_t^{(h)} = \frac{2^d \ln 2}{\Delta_{t-1}^{(d)}((\eta^{(d)})_1^{t-1})},$$

where the *cumulative* and *instantaneous* mixability gaps are defined as below:

$$\Delta_t^{(h)}((\eta^{(h)})_1^t) := \sum_{s=1}^{t} \delta_s^{(h)}(\eta_s^{(h)}), \text{ where}$$
$$\delta_s^{(h)}(\eta_s^{(h)}) := \langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, \mathbf{l}_s^{(\mathsf{tree})} \rangle + \frac{1}{\eta_s} \ln \langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, e^{-\eta_s \mathbf{l}_s^{(\mathsf{tree})}} \rangle$$

Thus, at round $(t-1)$, we have access to the quantity $\Delta_{t-1}^{(h)}$, and the complexity of computing $\eta_t^{(h)}$ is exactly equal to the complexity of computing $\delta_s^{(h)}(\eta_s^{(h)})$. It is easy to see that $\langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, \mathbf{l}_s^{(\mathsf{tree})} \rangle = \langle \mathbf{w}_s(\eta_s),\, \mathbf{l}_s \rangle$ and $\langle \mathbf{w}_s^{(\mathsf{tree})}(\eta_s),\, e^{-\eta_s \mathbf{l}_s^{(\mathsf{tree})}} \rangle = \langle \mathbf{w}_s(\eta_s),\, e^{-\eta_s \mathbf{l}_s} \rangle$. Therefore, we get

$$\delta_t^{(h)}(\eta_t^{(h)}) := \langle \mathbf{w}_t(\eta_t),\, \mathbf{l}_t \rangle + \frac{1}{\eta_t} \ln \langle \mathbf{w}_t(\eta_t),\, e^{-\eta_t \mathbf{l}_t} \rangle,$$

and since the size of all the vectors is equal to 2, the complexity of computing $\delta_t^{(h)}(\eta_t^{(h)})$ is $\mathcal{O}(1)$. Therefore, the complexity of computing $\eta_t^{(h)}$ is also $\mathcal{O}(1)$, and this completes the proof of the proposition. □

### Supplementary algebra

In this section, we state a couple of supplementary algebraic statements (and prove them when necessary).

**Fact 2.7.10.** *For two quantities $B, C \geq 0$, we have* $\max\{B, C\} \leq B + C$.

**Fact 2.7.11.** *For two numbers $B, C \geq 0$,*

$$x^2 - Bx - C \leq 0 \implies x \leq \sqrt{C} + B.$$

*This results from the quadratic formula, which gives us*

$$\begin{aligned} x &\leq \frac{B + \sqrt{B^2 + 4C}}{2} \\ &\leq \frac{B + B + 2\sqrt{C}}{2} = \sqrt{C} + B \end{aligned}$$

*where the last inequality is a consequence of*

$$a, b \geq 0 \implies \sqrt{a + b} \leq \sqrt{a} + \sqrt{b}.$$

# Chapter 3

# Model selection under bandit feedback

In the last chapter, we saw schemes that successfully adapt along two axes: a) between stochastic and adversarially generated data, b) offline benchmark, or model selection. These schemes critically relied on *full-information* feedback, i.e. the learner has the ability to observe the loss she would have incurred had she played any action, not just the one that she took. However, we noted in Chapter 1 that this is not a realistic assumption for several real-world applications. In real-world applications, often the learner can only receive loss/reward feedback for the action she took, and not observe feedback from any other action. For example, in wireless spectrum applications [139–141], the learner is a cognitive radio (or a decentralized set of radios), and the set of actions corresponds to a set of candidate channels or routing paths. Thus, the learner will receive reward feedback only corresponding to the action (channel or path) that she picked. We will subsequently see that models for online learning with contextual information in applications like recommender systems, advertisement placement, and mobile healthcare also receive feedback only corresponding to the action that was taken by the learner.

For these applications, we need to consider the *limited-information feedback* model, more colloquially known as the bandit model. This model introduces new challenges in the form of exploration: now actions need to be chosen not only to maximize estimated reward on a round (exploration), but also with the aim of maximizing "information gain" (exploitation). In this section, we explore the possibility of online learning algorithms that adapt in this more challenging environment, and particularly explore the problem of *model selection in contextual bandit problems*. While we provide a partial solution to this problem, several important questions remain open and we discuss these at the end of the chapter.

## 3.1 Setup

Here, we introduce the limited-information feedback (bandit problem) with side information. The sense in which we will consider model selection, is in deciding whether or not this side information matters. Before defining our setup, we provide basic notation and definitions

for this chapter below.

## Notation and definitions

Given a vector $v$, let $v_i$ denote its $i^{th}$ component. For a vector we let $\|v\|_p$ for $p \in [1, \infty]$ denote the $\ell_p$-norm. Given a matrix $M$ we denote its operator norm by $\|M\|_{op}$, and use $\|M\|_F$ to denote its Frobenius norm. Given a symmetric matrix $S$ let $\gamma_{\max}(S)$ and $\gamma_{\min}(S)$ denote its largest and smallest eigenvalues. Given a positive definite matrix $V$ we define the norm of a vector $w$ with respect to matrix $V$ as $\|w\|_V^2 = w^\top V w$. Let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration. A stochastic process $\{\xi_t\}_{t=1}^\infty$ where $\xi_t$ is measurable with respect to $\mathcal{F}_{t-1}$ is defined to be conditionally $\sigma$-sub-Gaussian for some $\sigma > 0$ if, for all $\lambda \in \mathbb{R}$, we have, $\mathbb{E}\left[e^{\lambda \xi_t} \middle| \mathcal{F}_{t-1}\right] \leq \exp(\lambda^2 \sigma^2 / 2)$.

## The multi-armed bandit problem (Simple model)

The most basic version of the problem of learning under limited-information feedback constitutes the *multi-armed bandit* problem with stochastic reward feedback. Under this model, the rewards of $K$ arms are iid over rounds, and so we have

$$g_{i,t} = \mu_i + \eta_{i,t}, \qquad \forall i \in [K]$$

where $\mu_i \in [-1, 1]$, $\{\eta_{i,t}\}_{i=1}^K$ are identical, independent, zero mean, $\sigma$-sub-Gaussian noise (defined below). The mean parameters $\mu_i$ are unknown beforehand to the learner. Critically, the learner receives reward feedback only for the arm she pulls at every round. More formally, if she plays arm $A_t$ at time $t$, she will only observe the reward $g_{t,A_t}$. Informally speaking, this necessitates the learner to trade-off *exploration* — playing arms that she has seen relatively fewer samples of thus far to learn more about the environment, and *exploitation* — playing the arm that seems to have the highest expected reward as estimated so far, to play optimality. How we define optimality is not immediately clear, and is traditionally formalized through two statistical paradigms:

1. The *Bayesian* paradigm, in which there is a prior over the reward means $\{\mu_i\}_{i=1}^K$ and rewards are discounted over time. In this paradigm, the multi-armed bandit problem is formulated as a partially observed Markov decision process (POMDP) and remarkably, its optimal policy is an *index policy* as highlighted in the seminal work of Gittins [142, 143].

2. The *frequentist* paradigm, in which we wish to asymptotically optimize the total expected reward of the policy, in the sense that the time-averaged gap to the maximal possible reward decays to 0 as the number of rounds, $T \to \infty$, and/or obtain a *minimax-optimal* guarantee on regret as a function of the number of rounds for any finite $T > 0$.

For convenience and simplicity[1] in algorithm design, we use the frequentist paradigm for the multi-armed bandit problem — asymptotically optimal algorithms, as we will describe below, include the popular algorithms UCB as well as Thompson[2] sampling [148]. Let the arm with the highest reward have mean $\mu^*$ and be indexed by $i^*$. Our certificate of optimality is minimizing the *pseudo-regret* quantity (henceforth regret for brevity), defined as

$$R_T^s := T\mu^* - \sum_{s=1}^{T} \mu_{A_s}.$$

Define the gap as the difference in the mean rewards of the best arm compared to the mean reward of the $i^{th}$ arm, that is, $\Delta_i := \mu^* - \mu_i$. The classical literature on multi-armed bandits [149] tells us that the best one can hope to do in this setting in the worst case is $\mathbb{E}[R_T^s] = \Omega(\sum_i \log(T)/\Delta_i)$. Several algorithms like UCB [150] and MOSS [151, 152] achieve this lower bound up to logarithmic (and constant) factors.

## The contextual bandit problem (Complex model)

The contextual bandit paradigm also studies limited-information feedback, but the rewards are now an unknown function of side, or *contextual*, information that is available to the learner. This model was first considered to model clinical trials [153]. Since then, it has been studied intensely both theoretically and empirically in many different application areas under many different pseudonyms. Applications of this paradigm include advertisement placement/web article recommendation [154, 155], clinical trials and mobile health-care [153, 156]. We point the reader to [156] for an extensive survey of the contextual bandits history and literature.

In this chapter, we will consider the simplest model for contextual bandits: the standard linear contextual bandits [157] paradigm. In this model, we assume there exists an underlying linear predictor $\theta^* \in \mathbb{R}^d$ shared across all arms[3]), $\alpha_{i,t} \in \mathbb{R}^d$ represents the contextual information and $\{\eta_{i,t}\}_{t=1}^n$ represents noise in the reward observations. We impose compactness constraints on the parameters: in particular, we have $\mu_i \in [-1, 1]$, $\theta^* \in \mathbb{B}_2^d(1)$. Further, the noise $\{\eta_{i,t}\}_{t=1}^T$ is assumed to be identical, independent, zero mean, and $\sigma$-sub-Gaussian.

[1] We touch more upon the debate between Bayesian and frequentist paradigms in a realm in which this has been relatively unexplored, game theory, in Chapter 4.

[2] Actually, Thompson sampling is a heuristic that can effectively utilize prior information in its algorithm design. However, it does not come with the Bayesian optimality certificate of the Gittins index policy — its guarantees are primarily frequentist in nature [144–146], although important Bayesian regret bounds have been proved as well [147].

[3] This is the model that was described in [157]. It is worth noting that more complex variants of this model with a separate $\theta_i^*$ for every $i \in [K]$ have also been empirically evaluated [154], and *biases* $[\mu_1, \ldots, \mu_K] \in \mathbb{R}^K$ of the $K$ arms, such that the mean rewards of the arms are affine functions of the contexts, that is, we have

$$g_{i,t} = \mu_i + \langle \theta^*, \alpha_{i,t} \rangle + \eta_{i,t}, \text{ for all } i \in [K].$$

At each round define $\kappa_t = \text{argmax}_{\kappa \in \{1,\ldots,K\}_{i=1}^{K}} \{\mu_\kappa + \langle \theta^*, \alpha_{\kappa,t} \rangle\}$ to be the best arm at round $t$. Here, we define pseudo-regret with respect to the optimal policy under the generative linear model:

$$R_T^c := \sum_{s=1}^{T} \left[ \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s,s} \rangle \right].$$

For the linear contextual bandit model *with finite number of actions*, it is well-known that variants of linear upper confidence bound algorithms like LinUCB [157] and OFUL [158] suffer at most $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{T})$ regret with respect to the optimal linear policy. Moreover, these algorithms are computationally efficient[4].

It is particularly clear that while employing the contextual bandits paradigm, the choice of policy class is critical to maximize the overall *reward* of the algorithm. As can be seen in applications of contextual bandits models for article recommendation [154], the choice is often made in hindsight, and more complex policy classes are used if the algorithm is run for more rounds. A quantitative understanding of how to do this is still lacking, and intuitively, we should expect the optimal choice of policy class to not be static. Ideally, we could design adaptive contextual bandit algorithms that would initially use simple policies, and switch over to more complex ones as more data is obtained.

Theoretically, what this means is that the regret bounds derived for a contextual bandit algorithm are only meaningful for rewards that are generated by a policy within the policy class to which the algorithm is tailored. If the rewards are derived from a "more complex" policy outside the policy class, even the optimal policy may neglect obvious patterns and obtain a very low reward. If the rewards are derived from a policy that is expressible by a much smaller class, the regret that is accumulated is unnecessary.

## Model selection, or, *does the contextual information matter?*

Let us view the model selection problem *vis-a-vis* the two model classes that we have described above: the standard multi-armed bandits paradigm (MAB) v.s. the standard linear contextual bandits [157] paradigm (CB). Observe that in CB, setting $\theta^* = 0$ yields the important case of the reward distribution being independent from the contextual information and thus the simple model (MAB) is *nested* within the complex model (CB). On one hand, if we knew the MAB structure beforehand, a simple upper confidence bound algorithm like UCB [150] would yield the optimal $\mathcal{O}(\log T)$ regret bound, *which does not depend on the*

[4]In the general contextual bandits paradigm, the question of computational efficiency has seen substantial research attention. Treating policies as experts (EXP4 [79]) with careful control on the exploration distribution led to the optimal regret bounds of $\mathcal{O}(\sqrt{KT \log |\Pi|})$ in a number of settings. From an *efficiency* point of view (where efficiency is defined with respect to an *arg-max-oracle* that is able to compute the best greedy policy in hindsight), the first approach conceived was the epoch-greedy approach [159], that suffers a sub-optimal dependence of $T^{2/3}$ in the regret. More recently, "randomized-UCB" style approaches [160] have been conceived that retain the optimal regret guarantee with $\widetilde{\mathcal{O}}(\sqrt{T})$ calls to the arg-max-oracle. This question of computational efficiency has generated a lot of subsequent, recent research interest [161–164].

*dimension of the contexts d.* But UCB ignores the contextual information and will not guarantee any control on the policy regret in the CB setting: it can even be linear. On the other hand, as we noted above, algorithms tailored to the CB setting like LinUCB [157] and OFUL [158] incur the minimax-optimal policy regret of $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{T})$ in the CB setting. But these algorithms also incur this sub-optimal dependence on the dimension even in simple instances when $\theta^* = 0$. Thus, we pay substantial extra regret by using the algorithm meant for CB on MAB instances, which have simpler structure.

The above discussion shows that neither of the algorithms tailored to the MAB or CB setting automatically adapt to the correct model class. This motivates the design of a *single* approach that adapts to the inherent complexity of the reward-generating model and obtains the optimal regret bound as if this complexity was known in hindsight. Specifically, we seek an answer to the following question:

*Does there exist a single algorithm that simultaneously achieves the $\mathcal{O}(\log T)$ regret rate on simple multi-armed bandit instances and the $\widetilde{\mathcal{O}}((\sqrt{d}+\sqrt{K})\sqrt{T})$ regret rate on linear contextual bandit instances?*

## 3.2 Related work: Challenges specific to limited-information model selection

The problem of policy class selection in contextual bandits has received some attention from an empirical perspective, although publicly available results are few and far between. A popular application of *linear* contextual bandits is to personalized article recommendation using hand-crafted features of users: in experiments conducted by Li et al [154], two classes of linear contextual bandit models with varying levels of complexity were compared to simple (multi-armed) bandit algorithms in terms of *overall reward* (which in this application represented the click-through rate of ads). A striking observation was that the more complex models won out when the algorithm was run for a longer period of time (e.g. 1 day as opposed to half a day). Surveys on contextual bandits as applied to mobile health-care [156] have expressed a desire for algorithms that adapt their choice of policy class according to the amount of information they have received (e.g. the number of rounds).

At a high level, we seek a theoretically principled way of doing this. A natural initial question is whether any of the full-information approaches that we described in Chapter 2 can be adapted to work in the bandit setting. Indeed, perhaps the most relevant work to online policy class selection involves significant attempts to *corral* a band of $M$ base bandit algorithms into a meta-bandit framework [165]. The idea is to bound the regret of the meta-algorithm in terms of the regret of the best base algorithm in hindsight. The CORRAL framework is a variation of the online validation framework proposed in Chapter 2, under limited-information feedback. Thus, it is very general and can be applied to any set of base algorithms, whether efficient or not. CORRAL, however, turns out not to be the optimal choice of *computationally efficient* algorithm for the multi-armed-vs-linear-contextual bandit

problem for a couple of reasons.

1. It is not clear what (if any) choice of base algorithms would lead to a computationally efficient algorithm that is also statistically optimal in a minimax sense simultaneously for both problems.

2. The meta-algorithm framework uses an experts algorithm (in particular, mirror descent with log-barrier regularizer and importance weighting on the base algorithms) to choose which base algorithm to play in each round. Thus, it is impossible to expect the instance-optimal regret rate of $\mathcal{O}(\log T)$ on the simple bandit instance. More generally, the CORRAL framework will not yield instance-optimal rates on any policy class[5].

As evidenced by the sub-optimalities of CORRAL in performing nested model selection, the principal difficulty in applying online validation approaches to contextual bandit model selection is the navigation of an even finer exploration-exploitation trade-off: algorithms (designed for particular model classes) that fall out of favor in initial rounds could be picked very rarely and the information required to truly perform model selection may be absent even after many rounds of play. CORRAL tackles this difficulty using the log-barrier regularizer for the meta-algorithm as a natural form of heightened exploration [166], together with clever learning rate schedules — but these turn out to not give optimal model selection guarantees in our setting. Similarly, directly extending SRM-based approaches to model selection in the contextual bandit problem without taking proper care of exploration schedules would lead to similar issues. In fact, in the *non-contextual* setup, impossibility results exist showing that model selection is not possible [167]. The question that we want to address is whether we can do better in a contextual environment.

## 3.3 Model selection: Multi-armed-bandit vs contextual bandit

Our algorithms use a more direct approach to model selection using cleverly designed statistical tests. We utilize a simple "best-of-both-worlds" principle: exploit the possible simple reward structure in the model until (unless) there is significant statistical evidence for the presence of complex reward structure *that would incur substantial complex policy regret if not exploited*. This algorithmic framework is inspired by the initial "best-of-both-worlds" results for stochastic and adversarial multi-armed bandits; in particular, the "Stochastic and Adversarial Optimal" (SAO) algorithm [168] (although the details of the phases of the algorithm and the statistical test are very different). In that framework, instances that are not stochastic (and could be thought of as "adversarial") are not always detected as such by the test. The test is designed in an elegant manner such that the regret is optimally bounded

[5]On our much simpler instance of bandit-vs-linear-bandit, we do obtain instance-optimal rates for at least the simple bandit model.

on instances that are not detected as adversarial, *even if an algorithm meant for stochastic rewards is used.* Our test to distinguish between simple and complex instances shares this flavor – in fact, all theoretically complex instances ($\theta^* \neq 0$) are not detected as such.

Another closely related algorithm, that also uses the sequential testing approach, is the concurrent work of [169] which tackles the problem of selecting among a hierarchy of linear classes with growing dimension. They work with stochasticity assumptions on the contexts that are *weaker* than the assumptions that we make in this chapter. However, they are only able to establish a sub-optimal bound on the regret of $\widetilde{\mathcal{O}}(d_*^{1/3}T^{2/3})$ (where $d_*$ is dimension of the optimal linear policy) as opposed to the minimax optimal regret rates (that scale with $T^{1/2}$) which we establish in this chapter. Our main observation is that commonly encountered sequences of contexts can help us carefully navigate the finer exploration-exploitation trade-off when the model classes are nested. We discuss comparisons between our algorithm OSOM and their algorithm, ModCB, in Section 3.6 at the end of this chapter.

## Construction of Confidence Sets

Underlying the design of *both* our algorithms is the design of appropriate upper confidence estimates corresponding to the bias of each arm, as well as the linear model parameter. We let $T_i(t) := \sum_{s=1}^{t} \mathbb{I}\left[A_s = i\right]$ be the number of times arm $i$ was pulled and $\bar{g}_{i,t} := \sum_{s=1}^{t} g_{i,s}\mathbb{I}\left[A_s = i\right]/T_i(t)$ be the average reward of that arm at the end of round $t$. For each arm we define the upper confidence estimate as follows,

$$\begin{aligned}\tilde{\mu}_{i,t} &:= \bar{g}_{i,t} \\ &+ \sigma\left[\frac{1+T_i(t)}{T_i^2(t)}\left(1+2\log\left(\frac{K(1+T_i(t))^{\frac{1}{2}}}{\delta}\right)\right)\right]^{\frac{1}{2}}.\end{aligned} \tag{3.1}$$

Lemma 6 in [158] (restated below as Lemma 3.3.1 here) uses a refined self-normalized martingale concentration inequality to bound $|\mu_i - \bar{g}_{i,t}|$ across all arms and all rounds.

**Lemma 3.3.1.** *Under the simple model, with probability at least* $(1-\delta)$ *we have,* $\forall i \in \{1,\ldots,K\}, \forall t \geq 0$*,*

$$\begin{aligned}&|\mu_i - \bar{g}_{i,t}| \\ &\leq \sigma\left[\frac{1+T_i(t)}{T_i^2(t)}\left(1+2\log\left(\frac{K(1+T_i(t))^{\frac{1}{2}}}{\delta}\right)\right)\right]^{\frac{1}{2}}.\end{aligned}$$

This controls the upper confidence bounds for the estimates of the bias terms $\{\mu_i\}_{i=1}^K$.

For any round $t > K$, let $\hat{\theta}_t$ be the $\ell^2$-regularized least-squares estimate of $\theta^*$ defined below.

$$\hat{\theta}_t = \left(\boldsymbol{\alpha}_{K+1:t}^\top \boldsymbol{\alpha}_{K+1:t} + I\right)^{-1} \boldsymbol{\alpha}_{K+1:t}^\top \mathbf{G}_{K+1:t}, \tag{3.2}$$

where $\boldsymbol{\alpha}_{K+1:t}$ is the matrix whose rows are the context vectors selected from round $K+1$ up until round $t$: $\alpha_{A_{K+1},K+1}^\top, \ldots, \alpha_{A_t,t}^\top$ and $\mathbf{G}_{K+1:t} = [g_{A_{K+1},K+1} - \tilde{\mu}_{A_{K+1},K}, \ldots, g_{A_t,t} - \tilde{\mu}_{A_t,t-1}]^\top$. Here we are regressing on the rewards seen to estimate $\theta^*$, while using the bias estimates $\tilde{\mu}_{i,t-1}$ obtained by our upper confidence estimates defined in Equation (3.1).

**Lemma 3.3.2.** *Let $\hat{\theta}_t$ be defined as in Equation* (3.2)*. Then, with probability at least* $(1-3\delta)$ *we have that for all $t > K$, $\theta^*$ lies in the set*

$$\mathcal{C}_t^c := \left\{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_2 \leq \mathcal{K}_\delta(t,T)\right\}, \tag{3.3}$$

*where $\mathcal{K}_\delta(t,T) = \widetilde{\mathcal{O}}(\sigma \cdot d \cdot \sqrt{T})$ is defined in Equation* (3.8d)*.*

We prove this lemma in Section 3.7.

## Optimal model selection under context diversity

It is of particular interest to identify whether *optimal* model selection is ever possible in contextual bandits. A "greedy" refinement of the sequential testing that uses no forced exploration is proposed below in Algorithm 1.

The intuition behind Algorithm 1 is straightforward. The algorithm starts off by using the simple model estimate of the recommended action, that is, $i_t$; until it has reason to believe that there is a benefit from switching to the complex model estimates. If the rewards are truly coming from the simple model, *or from a complex model that is well approximated by a simple multi-armed bandit model*, then Condition 3.6 *will not be violated* and the regret shall continue to be bounded under either model. However, if Condition 3.6 *is violated* then algorithm switches to the complex estimates – $j_t$ for the remaining rounds. The condition is designed using the function $\mathcal{W}_\delta(t,T)$ which is of the order $\widetilde{\mathcal{O}}(\sigma(d+\sqrt{K})\sqrt{t})$. This corresponds to the additional regret incurred when we attempt to estimate the extra parameter – $\tilde{\theta}_t \in \mathbb{R}^d$.

At each round Condition 3.6 compares the algorithm's *estimate* for the cumulative reward that could be obtained by playing according to the complex estimates – $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ – with the actual cumulative rewards seen so far $\sum_{s=K+1}^{t-1} g_{i_s,s}$ by sticking to the simple estimates.

Under the simple model, given our construction of the confidence sets the term $\sum_{s=K+1}^{t-1} \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ will be bounded by $\widetilde{\mathcal{O}}((d+\sqrt{K})\sqrt{t})$ as the true underlying vector $\theta^* = 0$. While the remaining terms $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} - g_{i_s,s}$ shall be at most $\widetilde{\mathcal{O}}(\sqrt{Kt})$; as the simple estimates $(i_s)$ shall be picking out the best arm quite often under the simple model. In fact under this model we show in Lemma 3.4.1 that Condition 3.6 is *not violated* with high probability and the algorithm shall continue using simple estimates throughout its entire run.

On the other hand, under the complex model, we switch to the complex estimates only if the difference between the algorithm's estimate for the cumulative reward that could be obtained by playing according to the complex estimates – $\sum_{s=K+1}^{t-1} \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle$ –

**Algorithm 1:** OSOM (Optimistic Selection Of Models)

**for** $t = 1, \ldots, K$ **do**
  Play arm $t$ and receive reward $g_{t,t}$, *(Play each arm at least once.)*
**for** $t = K+1, \ldots, n$ **do**
  `Current Model` $\leftarrow$ 'Simple'
  *Simple Model Estimate:*

$$i_t \in \underset{i \in \{1,\ldots,K\}}{\operatorname{argmax}} \{\tilde{\mu}_{i,t-1}\} \tag{3.4}$$

  *Complex Model Estimate:*

$$j_t, \tilde{\theta}_t \in \underset{i \in \{1,\ldots,K\}, \theta \in \mathcal{C}^c_{t-1}}{\operatorname{argmax}} \{\tilde{\mu}_{i,t-1} + \langle \alpha_{i,t}, \theta \rangle\}, \tag{3.5}$$

  where $\mathcal{C}^c_{t-1}$ defined in Equation (3.3).
  **if** `Current Model` = *'Simple' and* $t > K+1$ **then**
    Check the condition:

$$\sum_{s=K+1}^{t-1} \left\{ \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle - g_{i_s,s} \right\} \leq \mathcal{W}_\delta(t,T), \tag{3.6}$$

    where $\mathcal{W}_\delta(t,T)$ defined in Equation (3.8e).
    If violated then: `Current Model` $\leftarrow$ 'Complex'.
  If `Current Model` = 'Simple': Play arm $i_t$ and receive reward $g_{i_t,t}$.
  Else if `Current Model` = 'Complex': Play arm $j_t$ and receive $g_{j_t,t}$.
  Update $\{\tilde{\mu}_{i,t}\}_{i=1}^K$ and $\mathcal{C}^c_t$.

exceeds the rewards seen so far $\sum_{s=K+1}^{t-1} g_{i_s,s}$ by $\widetilde{\mathcal{O}}((d+\sqrt{K})\sqrt{t})$. That is, only when the algorithm starts to suffer a regret that is equal to the minimax rate of regret. While instead if this condition is not violated under the complex model, that is, our estimated cumulative reward for switching to the complex model is close to the rewards seen far. Then we show that the regret under the complex model is small even by using simple estimates. We do this in Lemma 3.4.2.

By combining the arguments outlined above our main theorem optimally bounds the regret of OSOM under either of the two reward-generating models. Underlying the success of our statistical test, and therefore *optimal* model selection guarantees, are the following

stochastic assumptions on the context vectors[6]. We assume that these contexts vectors $\alpha_{i,t} \in \mathbb{B}_2^d(1)$ and are drawn independent of the past from a distribution such that $\alpha_{i,t}$ is independent of $\{\alpha_{j,t}\}_{j\neq i}$ and, $\forall i \in [K]$ and $\forall t \in [T]$,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\alpha_{i,t}\right] &:= \mathbb{E}\left[\alpha_{i,t}\middle|\left\{\eta_{j,s}, \alpha_{j,s}\right\}_{j\in[K],s\in[t-1]}\right] = 0,\\ \mathbb{E}_{t-1}\left[\alpha_{i,t}\alpha_{i,t}^\top\right] &:= \mathbb{E}\left[\alpha_{i,t}\alpha_{i,t}^\top\middle|\left\{\eta_{j,s}, \alpha_{j,s}\right\}_{j\in[K],s\in[t-1]}\right]\\ &= \Sigma_c \succeq \rho_{\min}\cdot I, \end{aligned} \tag{3.7}$$

where we have $\rho_{\min} = c/d$ for some positive constant $c \in (0, 1]$ that does not depend on $d$. (Note that this scaling on $\rho_{\min}$ is because we have assumed $\|\alpha_{i,t}\|_2 \leq 1$, and so we trivially have $\rho_{\min} \leq 1/d$.)

Under this assumption, we are ready to state our model selection result.

**Theorem 3.3.3.** *With probability at least* $(1-9\delta)$*, we obtain the following upper bounds on regret for the algorithm* OSOM *(Algorithm 1):*

1. *Under the* Simple Model*:*

$$R_T^s \leq \sigma\cdot\sum_{i:\Delta_i>0}\left[3\Delta_i + \frac{16}{\Delta_i}\log\left(\frac{2K}{\Delta_i\delta}\right)\right].$$

2. *Under the* Complex Model*:*

$$R_T^c \leq 4(K+1) + 4\mathcal{W}_\delta(T,T) = \widetilde{\mathcal{O}}\left\{\sigma(d+\sqrt{K})\sqrt{T}\right\},$$

*where* $\mathcal{W}_\delta(T,T)$ *is defined in Equation* (3.8e)*.*

Notice that Theorem 3.3.3 establishes regret bounds on the algorithm OSOM which are near-minimax-optimal under both *simple model* and the *complex model* up to logarithmic factors. In fact, under the simple model we are able to obtain *problem-dependent* regret rates.

In the complex model we are close to the minimax rates obtained by OFUL (which holds for adversarial contexts as well), with a slight sub-optimality in the dimension dependence. A natural question for future work is if it is also possible to obtain problem dependent rates in the complex model simultaneously. For example under the complex model by using OFUL it is possible to show that regret grows poly-logarithmically with $T$: $R_T^c \leq \widetilde{\mathcal{O}}\left((d+K)^2/\Delta_\ell\right)$, where $\Delta_\ell$ is an appropriately defined *gap* in the linear model.

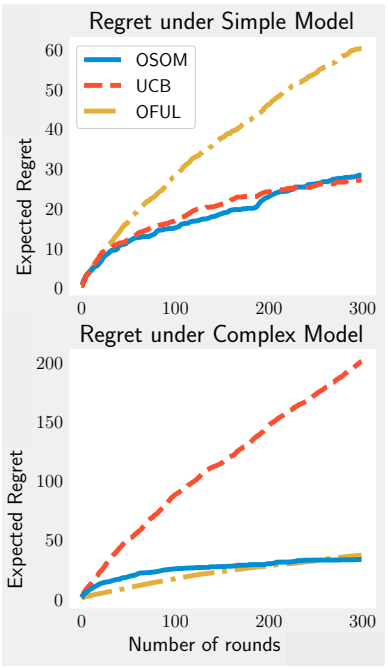



Figure 3.1: Experiments on synthetic data with $K = 5$, $d = 50$ and $T = 300$. The three algorithms plotted are OSOM, UCB and OFUL. Figure from [89].

## Empirical evaluation of algorithms

To experimentally corroborate our claims, we ran our model-selecting algorithm, OSOM, on both simple and complex instances. We compared its performance to that of UCB (which is optimal up to logarithmic factors under the simple model) and OFUL (which is minimax optimal under the complex model)[7].

When data is generated according to the simple model ($\theta^* = 0$), we see that OSOM and UCB suffer regret that is sub-linear, and is significantly lower than the regret suffered by

[6]Our assumption is essentially one of *context diversity* — the conditional mean of the context vectors are 0 and the co-variance matrix has its minimum eigenvalue *bounded below* by $\rho_{\min} := c/d$ for a positive constant $c \in (0, 1]$. The context diversity assumption has also been made to analyze the greedy algorithm in linear contextual bandits [170–172].

[7]Here is a more detailed description of the experiment: data was generated synthetically with the number of arms $K = 5$, and the dimension of $\theta^*$, $d = 50$. The mean rewards of the arms $\mu_i \sim \mathsf{Unif}(-1, 1)$, were drawn independently from a uniform distribution, and the context vectors $\alpha_{i,t}$ were drawn independently from the uniform distribution over the sphere. The noise $\eta_{i,t} \sim \mathcal{N}(0, 1)$ was drawn from a 1-dimensional Gaussian with unit variance. Under the simple model $\theta^* = 0$, while under the complex model $\theta^*$ was also drawn from the uniform distribution over the unit sphere in $d$-dimensions. In both the experiments we average over 50 runs over $T = 300$ rounds to estimate the expected regret incurred. The realizations of the problem were drawn independently for each run of each algorithm. For both OFUL and OSOM we used the empirical covariance matrix to build the upper confidence ellipsoid.

OFUL whose regret is also sub-linear but pays for the additional variance of estimating a more complex model. While when the data is generated from the complex model ($\|\theta^*\|_2 = 1$) the regret suffered by UCB is *linear* as it does not identify and estimate the linear structure of the mean rewards. Here, the regret suffered by both OFUL and OSOM is sub-linear and almost identical.

It is important to note here that algorithms designed solely for the linear contextual bandit problem, like OFUL, work for stochastic conditional rewards regardless of the sequence of contexts, which can be chosen *adversarially*. However, our goal here is to optimally adapt to simpler model structure while retaining the contextual bandit regret guarantee. Currently designed algorithms tailored to the linear contextual bandits problem, like OFUL, will fail at this objective even under the stochastic assumption. Our stochastic assumption essentially constitutes a *sufficient* condition for optimal model selection in linear contextual bandits. Whether it is *necessary*, that is, whether model selection is possible for the case of adversarial contexts, is an intriguing question left to future work.

## 3.4 Proofs

In this section, we collect the proofs of both algorithms.

### Proof of Theorem 3.3.3

In this section, present the key lemmas that underlie the proof of Theorem 3.3.3 below. (These lemmas are proved in Section 3.5.) We then use them to prove our main theorem.

To prove Theorem 3.3.3, we need to show that the regret of OSOM is appropriately bounded under either underlying model. In Lemma 3.4.1 we demonstrate that whenever the rewards are generated under the simple model, Condition 3.6 is *not violated* with high probability.

**Lemma 3.4.1.** *Assume that rewards are generated under the simple model. Then, with probability at least* $(1 - 5\delta)$*, we have for all* $t \in \{K + 2, \dots, T\}$*:*

$$\sum_{s=K+1}^{t-1} \left[\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle\right] - \sum_{s=K+1}^{t-1} g_{i_s,s} < \mathcal{W}_\delta(t-1, T).$$

This ensures that when the data is generated from the simple model, we have that the Boolean variable `Current Model` = 'Simple' throughout the run of the algorithm. Thus, the regret is equal to the regret incurred by the UCB algorithm, which is meant for simple model instances.

On the other hand, when the data is generated according to the complex model, we first demonstrate in Lemma 3.4.2 that the regret remains appropriately bounded if Condition 3.6 is *not violated*.

**Lemma 3.4.2.** *For all $t \in \{K+1, \ldots, T\}$. Let Condition 3.6 not be violated up until round $t+1$, that is,*

$$\sum_{s=K+1}^{t} \left\{ \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle - g_{i_s,s} \right\} \leq \mathcal{W}_\delta(t,T).$$

*Then, we have $R_t^c \leq 4K + 2\mathcal{W}_\delta(t,T)$, with probability at least $(1-5\delta)$.*

While when the data is generated according to the complex model and if the condition does get violated at a certain round, we switch to the estimates of the complex model, that is, $j_t$. This corresponds to a variant of the algorithm OFUL, which is meant for complex instances. Thus, the regret remains bounded in the subsequent rounds under this event as well (formally proved in Lemma 3.5.1 in Section 3.5). Combining the results of these three lemmas yields the regret bound, as described below.

*Proof of Theorem 3.3.3.* The proof is split into three cases.

**Simple model (MAB):** We have established in Lemma 3.4.1 that Condition 3.6 is *not violated* with probability at least $(1-5\delta)$ under the simple model. Conditioned on this event, OSOM plays according to the simple model estimate, $i_t$, for all rounds. Invoking Theorem 7 in [158] gives us that with probability at least $(1-\delta)$, $R_T^s \leq \sum_{i:\Delta_i>0} 3\Delta_i + (16/\Delta_i)\log(2K/\Delta_i\delta)$. Applying the union bound over these two events gives this regret bound with probability at least $(1-6\delta)$.

**Complex model (CB):** One out the two disjoint events are possible under the complex model.

**Case 1:** In this event Condition 3.6 is never violated throughout the run of the algorithm. Then by Lemma 3.4.2 we have

$$R_T^c \leq 4K + 2\mathcal{W}_\delta(T,T)$$

with probability at least $(1-5\delta)$.

**Case 2:** The other event is when Condition 3.6 is violated in round $\tau_* < T$. We know by Lemma 3.4.2:

$$R_{\tau_*-2}^c \leq 4K + 2\mathcal{W}_\delta(T,T)$$

with probability at least $(1-5\delta)$. Also, by Lemma 3.5.1:

$$\begin{aligned} R_{\tau_*:T}^c &:= \sum_{s=\tau_*}^{t} \left[ \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s,s} \rangle \right] \\ &\leq 2\mathcal{W}_\delta(T,T) \end{aligned}$$

with probability at least $(1 - 4\delta)$. We can decompose the cumulative regret up to round $T$ as follows:

$$R_T^c \leq R_{\tau_*-1}^c + R_{\tau_*:T}^c + 4,$$

where $R_{\tau_*:T}^c$ denotes the regret of the algorithm starting from round $\tau^*$ up to round $T$ and the 4 appears as it is the maximum regret that could be incurred in round $\tau_*$ by the algorithm under the complex model. By taking a union bound and using the decomposition of the regret above, we get $R_T^c \leq 4(K+1) + 4\mathcal{W}_\delta(T,T)$, with probability at least $(1 - 9\delta)$. □

It remains to prove Lemmas 3.4.1 and 3.4.2, and the remainder of this section is devoted to doing this.

## Useful functions

Before proving the lemmas, we formally define some useful functions that arise by applying the concentration inequalities on terms that appear while controlling the regret.

$$\tau_{\min}(\delta, T) := \left(\frac{16}{\rho_{\min}^2} + \frac{8}{3\rho_{\min}}\right) \log\left(\frac{2dT}{\delta}\right). \tag{3.8a}$$

$$\Upsilon_\delta(t,T) := \left(\frac{20}{3} + \frac{10\sigma}{3}\left[1 + 2\log\left(\frac{2KT}{\delta}\right)\right]^{\frac{1}{2}}\right) \tag{3.8b}$$
$$\left[\log\left(\frac{2dT}{\delta}\right) + \sqrt{t\log\left(\frac{2dT}{\delta}\right) + \log^2\left(\frac{2dT}{\delta}\right)}\right].$$

$$\mathcal{M}_\delta(t) := \sqrt{2\sigma^2\left(\frac{d}{2}\log\left(1 + \frac{t}{d}\right) + \log\left(\frac{1}{\delta}\right)\right)} + 1. \tag{3.8c}$$

$$\mathcal{K}_\delta(t,T) := \begin{cases} \mathcal{M}_\delta(t) + \Upsilon_\delta(t,T), & \text{if } K < t \leq K + \tau_{\min}(\delta,T), \\ \frac{\mathcal{M}_\delta(t)}{\sqrt{1+\rho_{\min}\cdot(t-K)/2}} + \frac{\Upsilon_\delta(t,T)}{1+\rho_{\min}\cdot(t-K)/2}, & \text{if } K + \tau_{\min}(\delta,T) < t. \end{cases} \tag{3.8d}$$

$$\mathcal{W}_\delta(t,T) := 2\sum_{s=K+1}^{t} \mathcal{K}_\delta(s-1,T) + \sigma\sqrt{\frac{1+t}{2}\log\left(\frac{1}{\delta}\right)} \tag{3.8e}$$
$$+ \left[2\sigma\sqrt{\left(1 + 2\log\left(\frac{Kt^{1/2}}{\delta}\right)\right)}\right]\sqrt{Kt}.$$

It is straightforward to verify that $\mathcal{W}_\delta(t,T) = \widetilde{\mathcal{O}}\left(\sigma(d+\sqrt{K})\sqrt{t}\right)$.

## 3.5 Proof of key lemmas

We define several statistical events that will be useful in proofs of the lemmas that follow in this section.

$$\mathcal{E}_1 := \left\{ \left| \sum_{s=K+1}^{t-1} \eta_{i_s,s} \right| \leq \sigma \sqrt{\frac{t}{2} \log\left(\frac{1}{\delta}\right)}, \forall t \in \{K+2, \ldots, T\} \right\}, \tag{3.9a}$$

$$\mathcal{E}_2 := \left\{ |\mu_i - \bar{g}_{i,t}| \leq \sigma \sqrt{\frac{1+T_i(s)}{T_i^2(s)} \left(1 + 2\log\left(\frac{K(1+T_i(s))^{1/2}}{\delta}\right)\right)}, \forall i \in [K] \text{ and } t \in [T] \right\}, \tag{3.9b}$$

$$\mathcal{E}_3 := \left\{ \|\hat{\theta}_t - \theta^*\|_2 \leq \mathcal{K}_\delta(t,T), \forall t \in \{K+1, \ldots, T\} \right\}. \tag{3.9c}$$

Event $\mathcal{E}_1$ represents control on the fluctuations due to noise: applying Theorem 3.7.3 in the one-dimensional case with $V = 1$ and $Y_s = 1$, we get $\mathbb{P}(\mathcal{E}_1^c) \leq \delta$ for all $t \geq 0$. Event $\mathcal{E}_2$ represents control on the fluctuations of the empirical estimate of the biases $[\mu_1, \ldots, \mu_K]$ around their true values: by Lemma 3.3.1 we have $\mathbb{P}(\mathcal{E}_2^c) \leq \delta$. Finally, event $\mathcal{E}_3$ represents control on the fluctuations of the empirical estimate of the parameter vector $\theta^*$ around its true value: by Lemma 3.3.2, we have $\mathbb{P}(\mathcal{E}_3^c) \leq 3\delta$. We define the desired event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ as the intersection of these three events. The union bound gives us $\mathbb{P}(\mathcal{E}^c) \leq 5\delta$. For the rest of the proof, we condition on the event $\mathcal{E}$.

### Regret under the simple model

We restate and prove the following lemma establishes that under the simple model, Condition 3.6 is not violated with high probability.

**Lemma 3.4.1.** *Assume that rewards are generated under the simple model. Then, with probability at least* $(1 - 5\delta)$*, we have for all* $t \in \{K+2, \ldots, T\}$*:*

$$\sum_{s=K+1}^{t-1} \left[ \tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle \right] - \sum_{s=K+1}^{t-1} g_{i_s,s} < \mathcal{W}_\delta(t-1, T).$$

*Proof.* Under the simple model, We have the model for the rewards is $g_{i,t} = \mu_i + \eta_{i,t}$.

Therefore, we have

$$\begin{aligned}
&\sum_{s=K+1}^{t-1}\left[\tilde{\mu}_{j_s,s-1}+\langle\alpha_{j_s,s},\tilde{\theta}_s\rangle\right]-\sum_{s=K+1}^{t-1}g_{i_s,s}\\
&\qquad=\sum_{s=K+1}^{t-1}\left[\tilde{\mu}_{j_s,s-1}+\langle\alpha_{j_s,s},\tilde{\theta}_s\rangle\right]-\sum_{s=K+1}^{t-1}\mu_{i_s}-\sum_{s=K+1}^{t-1}\eta_{i_s,s}\\
&=\sum_{s=K+1}^{t-1}-\eta_{i_s,s}+\sum_{s=K+1}^{t-1}[\tilde{\mu}_{i_s,s-1}-\mu_{i_s}]+\sum_{s=K+1}^{t-1}[\tilde{\mu}_{j_s,s-1}-\tilde{\mu}_{i_s,s-1}]+\sum_{s=K+1}^{t-1}\langle\alpha_{j_s,s},\tilde{\theta}_s\rangle\\
&=\underbrace{\sum_{s=K+1}^{t-1}-\eta_{i_s,s}}_{=:\Gamma_{no}}+\underbrace{\sum_{s=K+1}^{t-1}[\tilde{\mu}_{i_s,s-1}-\mu_{i_s}]}_{=:\Gamma_{sim1}}+\underbrace{\sum_{s=K+1}^{t-1}[\tilde{\mu}_{j_s,s-1}-\tilde{\mu}_{i_s,s-1}]}_{=:\Gamma_{sim2}}+\underbrace{\sum_{s=K+1}^{t-1}\langle\alpha_{j_s,s},\tilde{\theta}_s\rangle}_{=:\Gamma_{lin}}\\
&=\Gamma_{no}+\Gamma_{sim1}+\Gamma_{sim2}+\Gamma_{lin}.
\end{aligned}$$

Notice that the difference neatly decomposes into four terms, each of which we interpret below. The first term $\Gamma_{no}$ is purely a sum of the noise in the problem that concentrates under the event $\mathcal{E}_1$. The second term $\Gamma_{sim1}$ corresponds to the difference between the true mean reward $\mu_{i_s}$ and simple estimate of the mean reward $\tilde{\mu}_{i_s,s-1}$, which is controlled under the event $\mathcal{E}_2$. The third term $\Gamma_{sim2}$ is the difference between the mean rewards prescribed by the simple estimate and complex estimate $\tilde{\mu}_{i_s,s-1}$ and $\tilde{\mu}_{j_s,s-1}$ respectively. Finally, the last term $\Gamma_{lin}$ is only a function the estimated linear predictor (and since the true predictor is $\theta^*=0$, this term is controlled by even $\mathcal{E}_3$).

**Step (i)** *(Bound on $\Gamma_{no}$)*: Under the event $\mathcal{E}_1$, we have

$$\Gamma_{no}\leq\sigma\sqrt{\frac{t}{2}\log\left(\frac{1}{\delta}\right)}.$$

**Step (ii)** *(Bound on $\Gamma_{sim1}$)*: By the definition of $\tilde{\mu}_{i,s-1}$ we have,

$$\begin{aligned}
\Gamma_{sim1} &= \sum_{s=K+1}^{t-1} \tilde{\mu}_{i_s,s-1} - \mu_{i_s} \\
&\overset{(i)}{\leq} 2\sigma \sum_{s=K+1}^{t-1} \sqrt{\frac{1+T_{i_s}(s-1)}{T_{i_s}^2(s-1)} \left(1 + 2\log\left(\frac{K(1+T_i(s-1))^{1/2}}{\delta}\right)\right)} \\
&\leq 2\sigma \sum_{s=K+1}^{t-1} \sqrt{\frac{1+T_{i_s}(s-1)}{T_{i_s}^2(s-1)} \left(1 + 2\log\left(\frac{K(t-1)^{1/2}}{\delta}\right)\right)} \\
&= \left[2\sigma\sqrt{\left(1 + 2\log\left(\frac{K(t-1)^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K} \sum_{r=1}^{T_i(t-2)} \sqrt{\frac{1+r}{r^2}} \\
&\leq \left[2\sigma\sqrt{\left(1 + 2\log\left(\frac{K(t-1)^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K} \sum_{r=1}^{T_i(t-2)} 2\sqrt{\frac{1}{r}} \\
&\overset{(ii)}{\leq} \left[2\sigma\sqrt{\left(1 + 2\log\left(\frac{K(t-1)^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K} \sqrt{T_i(t-2)} \\
&\overset{(iii)}{\leq} \left[2\sigma\sqrt{\left(1 + 2\log\left(\frac{K(t-1)^{1/2}}{\delta}\right)\right)}\right] \sqrt{K(t-1)},
\end{aligned}$$

where $(i)$ follows under the event $\mathcal{E}_2$, $(ii)$ follows as

$$2 \sum_{r=1}^{T_i(t-2)} \sqrt{\frac{1}{r}} \leq 2 \int_0^{T_i(t-2)} \sqrt{\frac{1}{r}} \leq \sqrt{T_i(t-2)},$$

and $(iii)$ follows by Jensen's inequality and the fact that $\sum_{i=1}^{K} T_i(t-2) = t-2 < t-1$.

**Step (iii)** *(Bound on $\Gamma_{sim2}$)*: Equation (3.4), which shows the optimality of arm $i_s$, tells us that $\tilde{\mu}_{i_s,s-1} \geq \tilde{\mu}_{j_s,s-1}$ for all $s$. Therefore $\Gamma_{sim2} \leq 0$.

**Step (iv)** *(Bound on $\Gamma_{lin}$)*: By the Cauchy-Schwarz inequality, the constraint $\|\alpha_{i,t}\|_2 \leq 1$ and the triangle inequality, we get

$$\begin{aligned}
\Gamma_{lin} = \sum_{s=K+1}^{t-1} \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle &\leq \sum_{s=K+1}^{t-1} \|\alpha_{j_s,s}\|_2 \|\tilde{\theta}_s\|_2 \leq \sum_{s=K+1}^{t-1} \|\tilde{\theta}_s - \theta^*\|_2 \\
&\leq \sum_{s=K+1}^{t-1} \|\hat{\theta}_{s-1} - \tilde{\theta}_s\|_2 + \|\hat{\theta}_{s-1} - \theta^*\|_2 \leq 2 \sum_{s=K+1}^{t-1} \mathcal{K}_\delta(s-1, T),
\end{aligned}$$

where $\mathcal{K}_\delta(s-1, T)$ is defined in Equation (3.8d).

Combining the bounds on $\Gamma_{no}, \Gamma_{sim1}, \Gamma_{sim2}$ and $\Gamma_{lin}$ and by the definition of $\mathcal{W}_\delta(t-1,T)$, we have

$$\sum_{s=K+1}^{t-1} \left[\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle\right] - \sum_{s=K+1}^{t-1} g_{i_s,s} \leq \mathcal{W}_\delta(t-1,T),$$

which completes the proof. □

## Regret under the complex model

The bound on the regret under the complex model follows by establishing two facts. First, when Condition 3.6 is not violated, we demonstrate in Lemma 3.4.2 that the regret is appropriately bounded. Second, if the condition does get violated, say at round $\tau_*$, our algorithm OSOM chooses arms according to the complex model estimates '$j_t$' for $t \in [\tau_*, \ldots, T]$. In Lemma 3.5.1, we show that the regret remains bounded in this case as well.

We start with the first case by proving Lemma 3.4.2.

**Lemma 3.4.2.** *For all $t \in \{K+1, \ldots, T\}$. Let Condition 3.6 not be violated up until round $t+1$, that is,*

$$\sum_{s=K+1}^{t} \left\{\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s \rangle - g_{i_s,s}\right\} \leq \mathcal{W}_\delta(t,T).$$

*Then, we have $R_t^c \leq 4K + 2\mathcal{W}_\delta(t,T)$, with probability at least $(1-5\delta)$.*

*Proof.* Since we have already conditioned on the event $\mathcal{E}$, we can assume that events $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ hold. Note that if Condition 3.6 is not violated up to round t then we have that

$A_s = i_s$ for all $s \le t$. Using the definition of $R_t^c$, we get

$$\begin{aligned}
R_t^c &= \sum_{s=1}^{t} [\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \mu_{i_s} - \langle \theta^*, \alpha_{i_s,s} \rangle] \\
&\le 4K + \sum_{s=K+1}^{t} [\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \mu_{i_s} - \langle \theta^*, \alpha_{i_s,s} \rangle] \\
&= 4K + \sum_{s=K+1}^{t} (\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - g_{i_s,s}) + \sum_{s=K+1}^{t} (g_{i_s,s} - \mu_{i_s} - \langle \theta^*, \alpha_{i_s,s} \rangle) \\
&= 4K + \sum_{s=K+1}^{t} \Big( \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s} \rangle \Big) \\
&\qquad + \underbrace{\sum_{s=K+1}^{t} \Big( \tilde{\mu}_{j_s,s-1} + \langle \tilde{\theta}_s, \alpha_{j_s,s} \rangle - g_{i_s,s} \Big)}_{\le \mathcal{W}_\delta(t,T)} + \sum_{s=K+1}^{t} \eta_{i_s,s} \\
&\le 4K + \mathcal{W}_\delta(t,T) + \underbrace{\sum_{s=K+1}^{t} \Big( \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s} \rangle \Big)}_{=:\Gamma_{lin}} + \underbrace{\sum_{s=K+1}^{t} \eta_{i_s,s}}_{=:\Gamma_{no}},
\end{aligned}$$

where $4K$ is the maximum possible regret incurred in the first $K$ rounds under the complex model. By the definition of $\mathcal{E}_1$, we get $\Gamma_{no} \le \sigma\sqrt{((1+t)/2)\log(1/\delta)}$. Next, let us control $\Gamma_{lin}$. We have

$$\begin{aligned}
\Gamma_{lin} &= \sum_{s=K+1}^{t} \Big( \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s} \rangle \Big) \\
&= \sum_{s=K+1}^{t} \Big( \mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s} \rangle - \tilde{\mu}_{\kappa_s,s-1} - \langle \tilde{\theta}_s, \alpha_{\kappa_s,s} \rangle \Big) \\
&\qquad + \underbrace{\sum_{s=K+1}^{t} \Big( \tilde{\mu}_{\kappa_s,s-1} + \langle \tilde{\theta}_s, \alpha_{\kappa_s,s} \rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s} \rangle \Big)}_{\le 0},
\end{aligned}$$

where the non-positivity of the second term is because of the optimality of arm $j_s$ as expressed

in Equation (3.5). Hence, we have

$$\begin{aligned}\Gamma_{lin} &\leq \sum_{s=K+1}^{t} \Big(\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s}\rangle - \tilde{\mu}_{\kappa_s,s-1} - \langle \tilde{\theta}_s, \alpha_{\kappa_s,s}\rangle\Big) \\ &= \sum_{s=K+1}^{t} \mu_{\kappa_s} - \tilde{\mu}_{\kappa_s,s-1} + \sum_{s=K+1}^{t} \langle \alpha_{\kappa_s,s}, \theta^* - \tilde{\theta}_s\rangle \\ &\leq \sum_{s=K+1}^{t} \mu_{\kappa_s} - \tilde{\mu}_{\kappa_s,s-1} + \sum_{s=K+1}^{t} \|\alpha_{\kappa_s,s}\|_2 \|\theta^* - \tilde{\theta}_s\|_2 \\ &\leq \sum_{s=K+1}^{t} \mu_{\kappa_s} - \tilde{\mu}_{\kappa_s,s-1} + \sum_{s=K+1}^{t} \|\theta^* - \tilde{\theta}_s\|_2,\end{aligned}$$

where the last two inequalities follow from the Cauchy-Schwarz inequality and the constraint $\|\alpha_{i,t}\|_2 \leq 1$ respectively. Under the event $\mathcal{E}_2$, we have $\mu_{\kappa_s} - \tilde{\mu}_{\kappa_s,s-1} \leq 0$. Also, by the definition of $\tilde{\theta}_s$ and under event $\mathcal{E}_3$, we have

$$\Gamma_{lin} \leq 2 \sum_{s=K+1}^{t} \mathcal{K}_\delta(s-1, T).$$

Combining these bounds, we get

$$R_t^c \leq 4K + \mathcal{W}_\delta(t,T) + \sigma\sqrt{\frac{1+t}{2}\log\left(\frac{1}{\delta}\right)} + 2\sum_{s=K+1}^{t} \mathcal{K}_\delta(s-1,T) \leq 4K + 2\mathcal{W}_\delta(t,T)$$

under the assumption that event $\mathcal{E}$ holds. Since we already showed that $\mathbb{P}(\mathcal{E}) \geq 1 - 5\delta$, our proof is complete. □

Now, we move on to the second case. The next lemma shows that if Condition 3.6 was violated at round $\tau_*$ (which is, in general, a random variable), then playing the complex model estimates $j_s$ for all $s \geq \tau_*$ keeps the regret bounded in subsequent rounds.

**Lemma 3.5.1.** *If Condition 3.6 is violated at round $\tau_*$ that is,*

$$\sum_{s=K+1}^{\tau_*-1} \Big\{\tilde{\mu}_{j_s,s-1} + \langle \alpha_{j_s,s}, \tilde{\theta}_s\rangle - g_{i_s,s}\Big\} > \mathcal{W}_\delta(\tau_* - 1, T).$$

*Then with probability at least $(1-4\delta)$ we have,*

$$R^c_{\tau_*:T} := \sum_{s=\tau_*}^{t} [\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s}\rangle - \mu_{A_s} - \langle \theta^*, \alpha_{A_s,s}\rangle] \leq 2\mathcal{W}_\delta(T,T).$$

*Proof.* For this proof, we only need events $\mathcal{E}_2$ and $\mathcal{E}_3$ to simultaneously hold. We define the event $\mathcal{E}' := \mathcal{E}_2 \cap \mathcal{E}_3$. Again, by the union bound we have $\mathbb{P}(\mathcal{E}^c) \leq 4\delta$. For the rest of this proof we assume the event $\mathcal{E}'$.

If Condition 3.6 is violated at round $\tau^*$, then we have $A_s = j_s$ for all rounds $s \geq \tau_*$. Thus,

$$\begin{aligned}
R^c_{\tau_*:T} &= \sum_{s=\tau_*}^{T} \left[\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s}\rangle - \mu_{j_s} - \langle \theta^*, \alpha_{j_s,s}\rangle\right] \\
&= \sum_{s=\tau_*}^{T} \left[\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s}\rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s}\rangle\right] \\
&+ \sum_{s=\tau_*}^{T} \left[\tilde{\mu}_{j_s,s-1} + \langle \tilde{\theta}_s, \alpha_{j_s,s}\rangle - \mu_{j_s} - \langle \theta^*, \alpha_{j_s,s}\rangle\right] \\
&= \sum_{s=\tau_*}^{T} \left[\mu_{\kappa_s} + \langle \theta^*, \alpha_{\kappa_s,s}\rangle - \tilde{\mu}_{\kappa_s,s-1} - \langle \tilde{\theta}_s, \alpha_{\kappa_s,s}\rangle\right] \\
&+ \underbrace{\sum_{s=\tau_*}^{T} \left[\tilde{\mu}_{\kappa_s,s-1} + \langle \tilde{\theta}_s, \alpha_{\kappa_s,s}\rangle - \tilde{\mu}_{j_s,s-1} - \langle \tilde{\theta}_s, \alpha_{j_s,s}\rangle\right]}_{\leq 0} \\
&+ \sum_{s=\tau_*}^{T} \left[\tilde{\mu}_{j_s,s-1} + \langle \tilde{\theta}_s, \alpha_{j_s,s}\rangle - \mu_{j_s} - \langle \theta^*, \alpha_{j_s,s}\rangle\right],
\end{aligned}$$

where the second term is non-positive by the optimality of arm $j_s$ as expressed in Equation (3.5). Under the event $\mathcal{E}_2$, we have $\mu_i - \tilde{\mu}_{i,s-1} \leq 0$ for all $s > 0$ and $i \in [K]$. Therefore, we get

$$\begin{aligned}
R^c_{\tau_*:T} &\leq \sum_{s=\tau_*}^{T} \left[\langle \theta^* - \tilde{\theta}_s, \alpha_{\kappa_s,s}\rangle + \langle \tilde{\theta}_s - \theta^*, \alpha_{j_s,s}\rangle\right] + \sum_{s=\tau_*}^{T} [\tilde{\mu}_{j_s,s-1} - \mu_{j_s}] \\
&\leq \sum_{s=\tau_*}^{T} \|\tilde{\theta}_s - \theta^*\|_2 \left[\|\alpha_{\kappa_s,s}\|_2 + \|\alpha_{j_s,s}\|_2\right] + \sum_{s=\tau_*}^{T} [\tilde{\mu}_{j_s,s-1} - \mu_{j_s}] \\
&\leq \underbrace{2 \sum_{s=\tau_*}^{T} \|\tilde{\theta}_s - \theta^*\|_2}_{=:\Gamma_{lin}} + \underbrace{\sum_{s=\tau_*}^{T} [\tilde{\mu}_{j_s,s-1} - \mu_{j_s}]}_{\Gamma_{bias}},
\end{aligned}$$

where the inequalities follow by two applications of the Cauchy-Schwarz inequality and the

constraint $\|\alpha_{i,t}\|_2 \leq 1$. First we control $\Gamma_{lin}$. Under the event $\mathcal{E}_3$ we have

$$\Gamma_{lin} = 2\sum_{s=\tau_*}^{T} \|\tilde{\theta}_s - \theta^*\|_2 \leq 4\sum_{s=\tau_*}^{T} \mathcal{K}_\delta(s-1,T).$$

Next, we control the term $\Gamma_{bias}$. By the definition of $\tilde{\mu}_{j_s,s-1}$, we have

$$\begin{aligned}
\Gamma_{bias} = \sum_{s=\tau_*}^{T} [\tilde{\mu}_{j_s,s-1} - \mu_{j_s}] &\leq 2\sigma \sum_{s=\tau_*}^{T} \sqrt{\frac{1+T_{i_s}(s-1)}{T_{i_s}^2(s-1)}\left(1+2\log\left(\frac{K(1+T_i(s-1))^{1/2}}{\delta}\right)\right)} \\
&\leq 2\sigma \sum_{s=\tau_*}^{T} \sqrt{\frac{1+T_{i_s}(s-1)}{T_{i_s}^2(s-1)}\left(1+2\log\left(\frac{KT^{1/2}}{\delta}\right)\right)} \\
&\leq \left[2\sigma\sqrt{\left(1+2\log\left(\frac{K(T^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K}\sum_{r=1}^{T_i(T-1)} \sqrt{\frac{1+r}{r^2}} \\
&\leq \left[2\sigma\sqrt{\left(1+2\log\left(\frac{KT^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K}\sum_{r=1}^{T_i(T-1)} 2\sqrt{\frac{1}{r}} \\
&\leq \left[2\sigma\sqrt{\left(1+2\log\left(\frac{KT^{1/2}}{\delta}\right)\right)}\right] \sum_{i=1}^{K} \sqrt{T_i(T-1)} \\
&\overset{(i)}{\leq} \left[2\sigma\sqrt{\left(1+2\log\left(\frac{KT^{1/2}}{\delta}\right)\right)}\right] \sqrt{KT},
\end{aligned}$$

where $(i)$ follows by Jensen's inequality and the fact that $\sum_{i=1}^{K} T_i(T-1) = T-1 < T$. The rest of the inequalities can be verified by some simple algebra. Combining the bounds on the respective terms, we get

$$R^c_{\tau_*:T} \leq 4\sum_{s=\tau_*}^{T} \mathcal{K}_\delta(s-1,T) + \left[2\sigma\sqrt{\left(1+2\log\left(\frac{KT^{1/2}}{\delta}\right)\right)}\right]\sqrt{KT} \leq 2\mathcal{W}_\delta(T,T),$$

which completes the proof. □

## 3.6 Conclusions and future work

Our results should be thought of as initial progress on the goal of online model selection in contextual bandits, i.e. we identified *sufficient* conditions for contexts under which optimally preserving simple model performance with (slightly sub-optimal) complex model guarantees.‘ Under these conditions, we have shown that the algorithm tailored to the simple model automatically performs sufficient exploration to be able to perform model selection.

Whether these conditions are *necessary* is not clear, and more generally fundamental limits on contextual bandit model selection as well as instance-optimal algorithms remain open [173]. It is worth noting that the critical difficulty of the problem lies in reliably testing whether the maximal reward by any model in the complex model class is much greater than the maximal reward by any model in the simple model class. If this was given as side information, elegant schemes based on Corraling [165, 174] and regret balancing [175] show that optimal model selection is possible. However, knowing this information would correspond to actually knowing the model order, which is unrealistic when considering nested model classes in particular. Both our approach and that of [169] utilize being able to estimate this information in a much smaller number of samples than that required to estimate the optimal complex model, but in different ways. All of these observations suggest that the difficulty of contextual bandit model selection, at least in purely stochastic environments, could be intricately linked to the sample complexity of estimating this quantity. It would be interesting to turn these observations into concrete fundamental limits on model selection under limited-information feedback. Additionally, model selection beyond linear classes is of natural interest, and has already been considered in recent work [174, 176, 177].

Finally, we have not considered the possibility of an adversarial environment in this chapter. While we have "state-of-the-art" algorithms that adapt between stochasticity and adversity for a given model class [168, 178], it is likely that a tractable solution to the purely stochastic model selection problem will be needed before moving to *robust* model selection under limited-information feedback.

## 3.7 Omitted proof details

In this section, we fill in omitted proof details for completeness. We recall Lemma 3.3.2, which is an error bound on the ridge regression estimate $\hat{\theta}_t$, and present a proof below.

*Proof.* To unclutter notation, let $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{K+1:t}, \mathbf{G} = \mathbf{G}_{K+1:t}$. Further, define $\boldsymbol{\eta} = [\eta_{A_{K+1},K+1}, \dots, \eta_{A_t,t}]^\top$, $\boldsymbol{\mu} = [\mu_{A_{K+1}}, \dots, \mu_{A_t}]^\top$ and $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_{A_{K+1},K}, \dots, \tilde{\mu}_{A_t,t-1}]^\top$. By the definition of $\hat{\theta}_t$, we have

$$\begin{aligned}
\hat{\theta}_t &= \left(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I\right)^{-1}\boldsymbol{\alpha}^\top\mathbf{G} \\
&= \left(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I\right)^{-1}\boldsymbol{\alpha}^\top\left(\boldsymbol{\alpha}\theta^* + (\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} + \boldsymbol{\eta})\right) \\
&= \theta^* - \left(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I\right)^{-1}\theta^* + \left(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I\right)^{-1}\boldsymbol{\alpha}^\top(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + \left(\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I\right)^{-1}\boldsymbol{\alpha}^\top\boldsymbol{\eta}.
\end{aligned}$$

Now, let us define $V_t := \boldsymbol{\alpha}^\top\boldsymbol{\alpha} + I$. Then, for any vector $w \in \mathbb{R}^d$ (whose choice we will specify shortly), we get

$$\begin{aligned}
w^\top\left(\hat{\theta}_t - \theta^*\right) &= -w^\top V_t^{-1}\theta^* + w^\top V_t^{-1}\boldsymbol{\alpha}^\top(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + w^\top V_t^{-1}\boldsymbol{\alpha}^\top\boldsymbol{\eta} \\
&= -w^\top V_t^{-1/2}V_t^{-1/2}\theta^* + w^\top V_t^{-1/2}V_t^{-1/2}\boldsymbol{\alpha}^\top(\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}) + w^\top V_t^{-1/2}V_t^{-1/2}\boldsymbol{\alpha}^\top\boldsymbol{\eta}.
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left|w^\top\left(\hat{\theta}_t-\theta^*\right)\right| &\leq \|w\|_{V_t^{-1}}\left(\|\boldsymbol{\alpha}^\top\boldsymbol{\eta}\|_{V_t^{-1}}+\|\theta^*\|_{V_t^{-1}}+\|\boldsymbol{\alpha}^\top(\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}})\|_{V_t^{-1}}\right),\\ &\leq \|w\|_{V_t^{-1}}\left(\|\boldsymbol{\alpha}^\top\boldsymbol{\eta}\|_{V_t^{-1}}+\|\boldsymbol{\alpha}^\top(\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}})\|_{V_t^{-1}}+1\right), \qquad (3.10)\end{aligned}$$

where the second step follows as $\|\theta^*\|_{V_t^{-1}} \leq \sqrt{(1/\gamma_{\min}(V_t))}\cdot\|\theta^*\|_2 \leq 1$. We now define three events $\mathcal{E}_4, \mathcal{E}_5$ and $\mathcal{E}_6$ below:

$$\begin{aligned}\mathcal{E}_4 &:= \left\{\|\boldsymbol{\alpha}^\top\boldsymbol{\eta}\|_{V_t^{-1}} \leq \sqrt{2\sigma^2\log\left(\frac{\det(V_t)^{1/2}}{\delta}\right)}, \forall t\in\{K+1,\ldots,T\}\right\},\\ \mathcal{E}_5 &:= \left\{N_t := \left\|\sum_{s=K+1}^{t}\alpha_{A_s,s}\left(\mu_{A_s}-\tilde{\mu}_{A_s,t-1}\right)\right\|_2 \leq \Upsilon_\delta(t,T), \forall t\in\{K+1,\ldots,T\}\right\},\\ \mathcal{E}_6 &:= \left\{\gamma_{\min}(V_t)\geq 1+\rho_{\min}(t-K)/2, \forall t\in\{K+\tau_{\min}(\delta,T),\ldots,T\}\right\}.\end{aligned}$$

Define the event $\mathcal{E}'' := \mathcal{E}_4\cap\mathcal{E}_5\cap\mathcal{E}_6$. By Theorem 3.7.3 with $V=I$ we have, $\mathbb{P}(\mathcal{E}_4^c)\leq\delta$, by Lemma 3.7.2 we have $\mathbb{P}(\mathcal{E}_5^c)\leq\delta$ and Lemma 3.7.1 tells us that $(P)(\mathcal{E}_6^c)\leq\delta$. Therefore by a union bound $\mathbb{P}(\mathcal{E}^c)\leq 3\delta$. For the rest of the proof, we assume the event $\mathcal{E}''$. Hence, we get

$$\|\boldsymbol{\alpha}^\top\boldsymbol{\eta}\|_{V_t^{-1}} \leq \sqrt{2\sigma^2\log\left(\frac{\det(V_t)^{1/2}}{\delta}\right)} \overset{(i)}{\leq} \sqrt{2\sigma^2\left(\frac{d}{2}\log\left(1+\frac{t}{d}\right)+\log\left(\frac{1}{\delta}\right)\right)}, \qquad (3.11)$$

where $(i)$ follows by the technical Lemma 3.7.5. For the other term, we have

$$\left\|\boldsymbol{\alpha}^\top(\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}})\right\|_{V_t^{-1}} \leq \frac{N_t}{\sqrt{\gamma_{\min}(V_t)}} \leq \frac{\Upsilon_\delta(t,T)}{\sqrt{\gamma_{\min}(V_t)}}. \qquad (3.12)$$

Under $\mathcal{E}_6$, we have

$$\left\|\boldsymbol{\alpha}^\top(\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}})\right\|_{V_t^{-1}} \leq \begin{cases}\Upsilon_\delta(t,T), & \text{if } \tau_{\min}(\delta)\geq t-K>0,\\ \frac{\Upsilon_\delta(t,T)}{\sqrt{1+\rho_{\min}(t-K)/2}}, & \text{if } t-K>\tau_{\min}(\delta).\end{cases} \qquad (3.13)$$

Choosing $w = V_t(\hat{\theta}_t-\theta^*)$ and plugging in the upper bounds established in Equation (3.11) and Equation (3.13) into Equation (3.10), we get

$$\|\hat{\theta}_t-\theta^*\|_{V_t} \leq \begin{cases}\mathcal{M}_\delta(t)+\Upsilon_\delta(t,T), & \text{if } \tau_{\min}(\delta,T)\geq t-K>0,\\ \mathcal{M}_\delta(t)+\frac{\Upsilon_\delta(t,T)}{\sqrt{1+\rho_{\min}(t-K)/2}}, & \text{if } t-K>\tau_{\min}(\delta,T).\end{cases}$$

Recall the definition of $\mathcal{M}_\delta(t)$ in Equation (3.8c). Using the fact that $\|\hat{\theta}_t-\theta^*\|_2 \leq (1/\sqrt{\gamma_{\min}(V_t)})\|\hat{\theta}_t-\theta^*\|_{V_t}$ along with the event $\mathcal{E}_6$, we get

$$\begin{aligned}\|\hat{\theta}_t-\theta^*\|_2 &\leq \begin{cases}\mathcal{M}_\delta(t)+\Upsilon_\delta(t,T), & \text{if } \tau_{\min}(\delta,T)\geq t-K>0,\\ \frac{\mathcal{M}_\delta(t)}{\sqrt{1+\rho_{\min}(t-K)/2}}+\frac{\Upsilon_\delta(t,T)}{1+\rho_{\min}(t-K)/2}, & \text{if } t-K>\tau_{\min}(\delta,T),\end{cases}\\ &= \mathcal{K}_\delta(t,T),\end{aligned}$$

where the last equality is by the definition of $\mathcal{K}_\delta(t,T)$ in Equation (3.8d). □

Now, we establish a couple of concentration inequalities on quantities of interest in the proof of Lemma 3.3.2: these constitute Lemmas 3.7.1 and 3.7.2.

**Lemma 3.7.1.** *Define the matrix $M_t$ as*

$$M_t := I + \sum_{s=1}^{t} \alpha_{A_s,s}\alpha_{A_s,s}^\top.$$

*Then, with probability at least $(1-\delta)$, we have*

$$\gamma_{\min}(M_t) \geq 1 + \frac{\rho_{\min} t}{2}$$

*for all $\tau_{\min}(\delta, T) \leq t \leq T$.*

*Proof.* Note that by the definition of $M_t$, we have $\gamma_{\min}(M_t) = 1 + \gamma_{\min}\left(\sum_{s=1}^{t} \alpha_{A_s,s}\alpha_{A_s,s}^\top\right)$. By the assumption on the distribution of the contexts as specified in Equation (3.7), we have $\mathbb{E}_{s-1}\left[\alpha_{A_s,s}\alpha_{A_s,s}^\top\right] = \Sigma_c \succeq \rho_{\min} I$. Consider the matrix martingale defined by

$$Z_t := \sum_{s=1}^{t} \left[\alpha_{A_s,s}\alpha_{A_s,s}^\top - \Sigma_c\right] \text{ for } t = 1, 2, \ldots$$

with $Z_0 = 0$ and the corresponding martingale difference sequence $Y_s := Z_s - Z_{s-1}$ for $s = \{1, 2, \ldots\}$. As $\|\alpha_{A_s,s}\|_2 \leq 1$ and $\|\Sigma_c\|_{op} = \|\mathbb{E}_{s-1}\left[\alpha_{A_s,s}\alpha_{A_s,s}^\top\right]\|_{op} \leq 1$, we have

$$\|Y_s\|_{op} = \|\alpha_{A_s,s}\alpha_{A_s,s}^\top - \Sigma_c\|_{op} \leq 2.$$

We also have,

$$\begin{aligned}\left\|\mathbb{E}_{s-1}\left[Y_s Y_s^\top\right]\right\|_{op} = \left\|\mathbb{E}_{s-1}\left[Y_s^\top Y_s\right]\right\|_{op} &= \left\|\mathbb{E}_{s-1}\left[\left(\alpha_{A_s,s}\alpha_{A_s,s}^\top - \Sigma_c\right)\left(\alpha_{A_s,s}\alpha_{A_s,s}^\top - \Sigma_c\right)\right]\right\|_{op} \\ &\leq \left\|\mathbb{E}_{s-1}\left[(\alpha_{A_s,s}^\top \alpha_{A_s,s})\alpha_{A_s,s}\alpha_{A_s,s}^\top - \Sigma_c^2\right]\right\|_{op} \leq 2.\end{aligned}$$

By applying the Matrix Freedman inequality (Theorem 3.7.4 in Section 3.7) with $R = 2$, $\omega^2 = 2t$ and $u = \rho_{\min} t/2$, we get that if $t \geq (16/\rho_{\min}^2 + 8/(3\rho_{\min}))\log(2dT/\delta)$, then

$$\mathbb{P}\left\{\left\|\sum_{s=1}^{t} \alpha_{A_s,s}\alpha_{A_s,s}^\top - t \cdot \Sigma_c\right\|_{op} \geq \frac{\rho_{\min} t}{2}\right\} \leq \frac{\delta}{T}.$$

This implies that

$$\gamma_{\min}\left(\sum_{s=1}^{t} \alpha_{A_s,s}\alpha_{A_s,s}^\top\right) \geq \frac{\rho_{\min} t}{2}$$

for a given $t \in \{\tau_{\min}(\delta, T), \ldots, T\}$ with probability at least $(1 - \delta/T)$. Taking a union bound over all $t \in \{\tau_{\min}(\delta, T), \ldots, T\}$ yields the desired claim. □

**Lemma 3.7.2.** *Define the vector* $N_t := \sum_{s=K+1}^{t} \alpha_{i_s,s}(\mu_{i_s} - \tilde{\mu}_{i_s,s-1})$. *For all* $K < t \leq T$ *we have,*

$$\|N_t\|_2 \leq \Upsilon_\delta(t,T),$$

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

*Proof.* Consider $K < t \leq T$. Note that $\tilde{\mu}_{i_s,s-1}$ is a function of $g_{i_1,1},\ldots,g_{i_{s-1},s-1}$ and $i_1,\ldots,i_{s-1}$. Also, the simple model estimate $i_s$ is just a function of $g_{i_1,1},\ldots,g_{i_{s-1},s-1}$ and $A_1,\ldots,A_{s-1}$. Therefore, we have

$$\mathbb{E}_{s-1}\left[\alpha_{i_s,s}(\mu_{i_s} - \tilde{\mu}_{i_s,s-1})\right] = (\mu_{i_s} - \tilde{\mu}_{i_s,s-1})\mathbb{E}_{s-1}\left[\alpha_{i_s,s}\right] = 0$$

for all $s \in \{K+1,\ldots,t\}$, as $\alpha_{i_s,s}$ is assumed to drawn from a distribution with zero (conditional) mean. Recall that $\mu_{i_s} \in [-1,1]$. By the definition $\tilde{\mu}_{i_s,s-1}$, we have

$$\tilde{\mu}_{i_s,s-1} = \underbrace{\sum_{r=1}^{s-1} \frac{g_{i_s,r}\mathbb{I}\left[A_r = i_s\right]}{T_{i_s}(s-1)}}_{\in\{-1,1\}} + \sigma \sqrt{\underbrace{\frac{1+T_{i_s}(s-1)}{T_{i_s}^2(s-1)}}_{\leq 2} \underbrace{\left(1 + 2\log\left(\frac{K(1+T_{i_s}(s-1))^{1/2}}{\delta}\right)\right)}_{\leq 1+2\log(K(1+T)/\delta)}},$$

and therefore

$$|\mu_{i_s} - \tilde{\mu}_{i_s,s-1}| \leq 2 + \sigma\sqrt{2\left(1 + 2\log\left(\frac{K(1+T)}{\delta}\right)\right)} =: \mathcal{P}_T, \qquad \forall s \in \{1,\ldots,T\}.$$

Define a martingale $Z_{t-K} := N_t$ and the martingale difference sequence $Y_s := Z_s - Z_{s-1}$. Then we have, for any $s \in \{K+1,\ldots,t\}$,

$$\|Y_{s-K}\|_{op} = \|Y_{s-K}\|_2 \leq \left\|\alpha_{i_s,s}\left(\mu_{i_s} - \tilde{\mu}_{i_s,s-1}\right)\right\|_2 \leq \|\alpha_{i_s,s}\|_2 |\mu_{i_s} - \tilde{\mu}_{i_s,s-1}| \leq |\mu_{i_s} - \tilde{\mu}_{i_s,s-1}| \leq \mathcal{P}_T.$$

We also have

$$\left\|\mathbb{E}_{s-1}\left[\alpha_{i_s,s}\alpha_{i_s,s}^\top\left(\mu_{i_s} - \tilde{\mu}_{i_s,s-1}\right)^2\right]\right\|_{op} \leq \mathcal{P}_T^2\|\Sigma_c\|_{op} \leq \mathcal{P}_T^2,$$

and

$$\left\|\mathbb{E}_{s-1}\left[\alpha_{i_s,s}^\top\alpha_{i_s,s}\left(\mu_{i_s} - \tilde{\mu}_{i_s,s-1}\right)^2\right]\right\|_{op} \leq \mathcal{P}_T^2\|\alpha_{i_s,s}\|_2^2 \leq \mathcal{P}_T^2.$$

Invoking Theorem 3.7.4 with $R = \mathcal{P}_T$ and $\omega^2 = \mathcal{P}_T^2(t-K)$, we get

$$\mathbb{P}\left\{\|N_t\|_2 \geq \frac{\mathcal{P}_T}{3}\log\left(\frac{2dT}{\delta}\right) + \frac{\mathcal{P}_T}{3}\sqrt{18(t-K)\log\left(\frac{2dT}{\delta}\right) + \log^2\left(\frac{2dT}{\delta}\right)}\right\} \leq \frac{\delta}{T}.$$

From the definition of $\Upsilon_\delta(t,T)$ in Equation (3.8b) and applying the union bound over all $t \in \{K+1,\ldots,T\}$, we get

$$\mathbb{P}\left\{\exists t \in \{K+1,\ldots,T\} : \|N_t\|_2 \geq \Upsilon_\delta(t,T)\right\} \leq \delta.$$

This completes the proof. □

## Concentration inequalities and technical results

In this section we state technical concentration inequalities that are useful in our proofs. We start by defining notation specific to this section.

Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\xi_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $\xi_t$ is $\mathcal{F}_t$-measurable and $\xi_t$ is conditionally $\sigma$-sub-Gaussian. Let $\{Y_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $Y_t$ is $\mathcal{F}_{t-1}$-measurable. Assume that $V$ is a $d \times d$ positive definite matrix. For any $t > 0$ define

$$V_t := V + \sum_{s=1}^{t} Y_s Y_s^\top, \qquad S_t := \sum_{s=1}^{t} \xi_s Y_s.$$

With this setup in place, the following is a re-statement of Theorem 1 of [158], which is essentially a self-normalized concentration inequality.

**Theorem 3.7.3.** *For any $\delta > 0$, we have*

$$S_t^\top V_t^{-1} S_t = \|S_t\|_{V_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right)$$

*with probability at least $(1-\delta)$ for all $t \geq 0$.*

Next we state a version of the Matrix Freedman Inequality [179, Corollary 1.3] that we use multiple times in our arguments. Define a matrix martingale as a sequence $\{Z_s : s = 0, 1, \ldots\}$ such that $Z_0 = 0$ and

$$\mathbb{E}\left[Z_s | \mathcal{F}_{s-1}\right] = Z_{s-1} \qquad \text{and} \qquad \mathbb{E}\left[\|Z_s\|_{op}\right] \leq \infty, \qquad \text{for } s = 1, \ldots.$$

Also define the martingale difference sequence $X_s := Z_s - Z_{s-1}$.

**Theorem 3.7.4.** *Consider a matrix martingale $\{Z_s : s = 0, 1, \ldots\}$ whose values are matrices with dimension $d_1 \times d_2$, and let $\{X_s : s = 0, 1, \ldots\}$ be the martingale difference sequence. Assume that the difference sequence is almost surely uniformly bounded, that is,*

$$\|X_s\|_{op} \leq R \qquad a.s. \qquad \text{for } s = 1, 2 \ldots$$

*Define two predictable quadratic variation processes of the martingale:*

$$W_{col,t} := \sum_{s=1}^{t} \mathbb{E}\left[X_s X_s^\top | \mathcal{F}_{s-1}\right] \qquad \text{and}$$

$$W_{row,t} := \sum_{s=1}^{t} \mathbb{E}\left[X_s^\top X_s | \mathcal{F}_{s-1}\right] \qquad \text{for } t = 1, 2, \ldots$$

*Then for all $u \geq 0$ and $\omega^2 > 0$, we have*

$$\mathbb{P}\left\{\exists t \geq 0 : \|Z_t\|_{op} \geq u \text{ and } \max\left\{\|W_{col,t}\|_{op}, \|W_{row,t}\|_{op}\right\} \leq \omega^2\right\}$$
$$\leq (d_1 + d_2) \exp\left(-\frac{u^2/2}{\omega^2 + Ru/3}\right).$$

The final technical result we recap characterizes the growth of the determinant of the matrix $V_T$, and is useful in constructing our confidence sets for the estimate of $\theta^*$. This result is a restatement of Lemma 19.1 in the pre-print [180].

**Lemma 3.7.5.** *Let $V_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix and $z_1, \ldots, z_T \in \mathbb{R}^d$ be a sequence of vectors with $\|z_t\|_2 \leq L < \infty$ for all $t \in [T]$. Further, let $v_0 := tr(V_0)$ and $V_T := V_0 + \sum_{s=1}^{T} z_s z_s^\top$. Then, we have*

$$\log\left(\frac{\det(V_T)}{\det(V_0)}\right) \leq d\log\left(\frac{v_0 + TL^2}{d\det^{1/d}(V_0)}\right).$$

# Part II

# Game Theory In The Presence Of Learning

# Chapter 4

# Learning from strategic, non-adversarial data

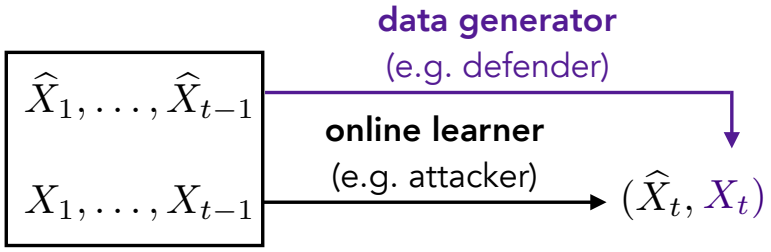



Figure 4.1: Online learning from competitively generated data. The learner's actions are denoted by $\{\widehat{X}_t\}_{t\geq 1}$ and the competitive agent's actions are denoted by $\{X_t\}_{t\geq 1}$. Note that both $(X_t, \widehat{X}_t)$ can depend on all of the past information $\{X_s, \widehat{X}_s\}_{s=1}^{t-1}$. Figure made using Keynote.

In Chapter 1, we motivated our study of online learning by setting up the fundamental goal of being able to design a learning algorithm that distinguishes between three kinds of data: stochastic, strategic and cooperative. In Part I of this thesis, we discussed methodology for adapting between stochastic and *adversarially strategic* data; however, in most interesting scenarios, the data will be generated with a strategic but non-adversarial incentive in mind. In both the examples of online marketplaces (Section 1.1) and dynamic spectrum sharing (Section 1.2), the most commonly encountered agents are selfish and rational; therefore, learning in the presence of strategic agents necessitates an examination of how to learn from *competitively generated data*. Figure 4.1 provides a depiction of the ensuing repeated game between learner and competitive "data generator".

In this chapter, we study this question continuing to work in the online learning framework. As we examined in Section 1.3, the problem of learning from competitively generated data is unavoidably tied up with understanding explicit mechanisms for how this data will be generated in response to learning. Thus, we cannot study optimality for either learner or data generator in isolation—we need to study *repeated game equilibria*, or approximate notions of it. To simplify the problem at hand, we assume a specialized setting of *one-sided learning*, in which *the data generator is assumed to know the learner's utility function.* As we will see, this helps us connect explicitly to the traditional understanding, developed in Bayesian repeated game theory, of the eventual outcome of interaction between learner and data generator. Moreover, the one-sided learning setup, also commonly called *information asymmetry*, is well-motivated in several applications such as security[1] games [181], where the learner is an attacker who wishes to identify a target to attempt to compromise, and the "data generator" is a defender who decides which target to defend with her protection resources from possible attack. Information asymmetry is sometimes applicable to model the interaction in online marketplaces: for e.g. in auctions between a single seller and bidder, while (as we saw in Section 1.1) the seller may not know even the bidder's prior over utility functions, it is conceivable that bidder knows the seller's utility function owing to its generic nature, i.e. to maximize monetary revenue or social welfare.

Thus, the focus of this chapter is to postulate *natural and explicit rules* that both the learner and the data-generator will follow in repeated interaction. We do so by taking a novel *frequentist* perspective on the ensuing repeated game with (one-sided) incomplete information. Using these rules, we will address a central question: *does the generator of data want to reveal, or obfuscate, her (private) utility function information?*

## 4.1 Review: One-sided learning and Stackelberg equilibrium

Classical work on reputation-building in Bayesian repeated game theory [182–186] has shown that the eventual outcome of *repeated* interaction between learner and "data generator", in terms of time-averaged payoffs expected by both agents, corresponds to the outcome of a special kind of *one-shot* interaction between a designated *leader* (here, the data generator) and a designated *follower*. The nature of this one-shot interaction is called a *Stackelberg game*, and is depicted in Figure 4.2. The critical difference from a traditional one-shot, simultaneous, non-cooperative game is that in the Stackelberg setup, the leader has the ability to commit to, and reveal her *mixed* strategy in advance – and the follower has the ability to observe this commitment and respond to it. Under this framework, we can easily determine an *optimal* commitment for the leader, which is commonly called the Stackelberg

[1]It is natural to ask why security games typically manifest as non-zero-sum. The reason for this is that the defender often has a different utility function from the attacker, e.g. she may prefer defending certain targets over others, while the attacker simply wants to attack which-ever target is most likely to be left open. We will shortly see concrete examples of these types of games.

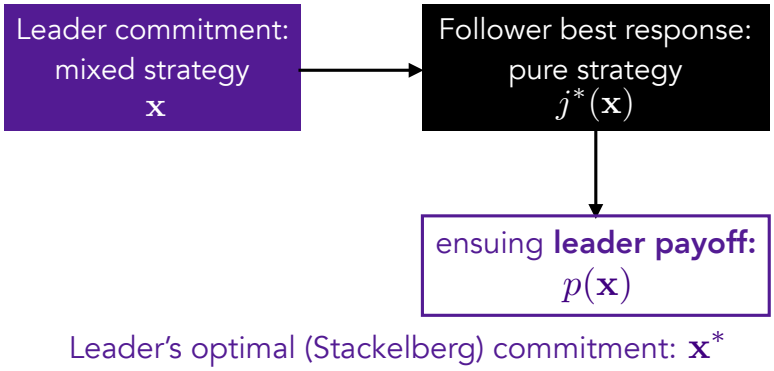



Figure 4.2: Extensive form of the one-shot Stackelberg game between leader and follower. $\mathbf{x}$ denotes the leader commitment, and $j^*(\mathbf{x})$ denotes the follower best-response to the leader commitment. Leader Stackelberg commitment is given by $\mathbf{x}^*$, and her Stackelberg payoff is denoted by $p^*$. Figure made using Keynote.

commitment, and the corresponding payoff obtained by the leader, which is commonly called the *Stackelberg payoff*. Together, the set of the strategies followed by leader and follower are called the *Stackelberg equilibrium* of the one-shot game.

Thus, an informal summary of the classical work on reputation-building is as follows: The eventual outcome of *repeated* interaction with one-sided learning from competitively generated data is the Stackelberg equilibrium of the *one-shot* non-zero-sum game. Here, the data generator is designated as the leader and the learner is designated as the follower. (Accordingly, for the rest of this chapter, we will use the terms leader and follower to refer to data generator and learner respectively.) Crucially, the leader benefits significantly from this commitment power, i.e. her ensuing Stackelberg equilibrium payoff is always at least as much as her simultaneous equilibrium payoff [187]. This can be interpreted as the implicit advantage that the leader enjoys as a result of the information asymmetry: she is utilizing the information that she, alone, possesses of her utility function to maximal advantage. Since it is an outcome of infinitely repeated interaction, the Stackelberg solution concept assumes a very idealized setting in which the mixed strategy commitment is *exactly revealed* to the follower. We could ask what might happen when these assumptions are relaxed, e.g. what if the leader could only demonstrate her commitment in a finite number of interactions? The questions arise of how she would modify her strategy to maximize payoff, and how much commitment power she would continue to enjoy.

At a deeper level, the insightful connection between one-sided learning and Stackelberg equilibrium explains *why* a patient leader can eventually achieve the full power of a reputation/commitment – but does not provide an explicit mechanism for *how* she can achieve

it[2]. There are potentially many ways in which the leader could realize mixed-strategy commitment power. In a paradigmatic example of Bayesian persuasion (Section 1 of [188]), a prosecutor, who privately knows a defendant's guilt, convinces a judge to convict some proportion of innocent defendants by committing to a *randomized* signaling mechanism. In this mechanism, she provides a guilty signal with non-zero probability conditioned on the defendant being actually innocent. Let's say that the prosecutor's optimal signaling mechanism has her do this with probability 1/2. The prosecutor's reputation for signaling in this manner is actually established through her track record with multiple judges, a track record that she herself can generate. We know that she will eventually develop a reputation for signalling optimally – but how she chooses to get there is unclear, and an intriguing question. Should she signal guilty for every innocent defendant independently, with probability 1/2? Should she signal guilty less often for initial defendants, and more frequently for later ones? What partial reputation power, if any, does she hold after interacting in this way with, say, a 100 judges? What if she instead signaled in a deterministic fashion – in such a way that it looks as though she signals guilty for 1/2 of innocent defendants most of the time? Would such a strategy be optimal, or brittle?

These are difficult questions to concretely answer, because reasoning about constructive ways to build a leader reputation critically involves reasoning about how followers will *learn* and make inferences from the leader's history. Since the game is non-cooperative, additional complexity is introduced by the potential for leader incentive to *deceive* this learning process, and potentially go even beyond the power of Stackelberg commitment. In what follows, we discuss the Bayesian perspectives on repeated games with one-sided learning, and motivate a novel frequentist perspective to attempt to answer the twin questions of:

1. *How a follower should learn from data generated by a leader*, and
2. *How a leader should shape her data in response to follower learning.*

## 4.2 Classical Bayesian setting: Reputation building and Stackelberg equilibrium

The *one-shot* Stackelberg solution concept dates back to von Stackelberg [189] who demonstrated the power of commitment in a quantity-setting duopoly. Schelling [190] provides a succinct description for the power of commitment: *"In independent decision situations, weakness can confer strength – power may result from the power to bind oneself."* He also notes that commitment can be beneficial only if the communication channel is sufficiently

[2]This is because the analysis of the Bayesian repeated game model, while extremely insightful for the case of pure-strategy reputation, is intractable for the case of a general mixed-strategy reputation. We chose frequentist learning models for followers for our model instead. We compare these modeling assumptions in detail in Section 4.3.

reliable. In traditional economic market models, such as firm competition [191] and seller-buyer participation games [192], such commitment is usually to a *pure strategy*; and the communication channel involves the relay of a public signal which may or may not be noisy.

The establishment of *leader credibility* in even a pure-strategy commitment is not trivial. Credibility is traditionally explained by connecting the pure strategy Stackelberg solution concept to the asymptotic limit of reputation building through repeated interaction[3]. Reputation effects were first observed in the *chain-store paradox*, a firm-competition game where an incumbent firm would often deviate from Nash equilibrium behavior and play its aggressive Stackelberg (pure, in this case) strategy [191]. In seminal work, Kreps, Wilson, Milgrom and Roberts provide theoretical justification for this particular game [182, 183] by modeling a $(N+1)$-player interaction between a monopolist and multiple entrants, and studying the limiting payoff of the sub-game-perfect-Nash equilibrium (SPNE) of an ensuing game as $N \to \infty$. They show that the monopolist would eventually play her *pure Stackelberg strategy* in the SPNE of this game endowed with a common (finitely supported) prior either on the leader's payoff structure, or on leader behavior either being "rational", or being constrained to play with a single pure strategy. The latter case models the possibility of a leader satisfying a "commitment type"[4]. Fudenberg and Levine subsequently show positive reputation attainment in two-player games endowed with a common prior on pure "commitment types" [184], and generalize their results to a continuum of mixed commitment types with the possibility of imperfectly observed leader actions [185]. The most general theoretical result for positive reputation is that any leader of "rational type"[5] can eventually realize the *payoff* of any mixed commitment (including the Stackelberg commitment) in all sub-game perfect Nash equilibria (SPNE) of the game as the discount factor goes to 1, i.e. an infinitely patient leader. Gossner [186] provides a recent, insightful analysis through a characterization of follower responses to a particular commitment type using relative entropy techniques from information theory [195]. In the general case, the SPNE is not unique and there are multiple possible ways in which a *rational* leader could achieve a mixed reputation. While the results are asymptotic (i.e. for the infinitely repeated game), it has been suggested that to realize her ideal Stackelberg payoff, the leader needs to play the game for the longest. Example games such as the product-choice game (Example 4.1 in the survey [194]) hint at the potential instability of the mixed commitment under finite observability being at the heart of the reason for this. Our work formalizes this idea, which turns out to be central to the issue of how to *explicitly construct* leader strategies that can plausibly build mixed reputation.

Related in spirit to the study of mixed commitments under observational uncertainty are intriguing results for *pure strategy* commitments under a different noise model. For pure strategy commitments, the noise model is essentially Selten's traditional trembling hand perturbation [196]: there is a positive probability that the intended pure strategy is

[3]For two excellent surveys, see Sorin [193] and Mailath, Samuelson [194].

[4]This explicitly models the spirit of commitment power resulting from the power to bind oneself.

[5]Also called "payoff types" [194].

distorted to a different one, which is then signaled to the follower[6]. Bagwell [198] shows that if one restricts to pure strategies, even a minimal amount of observation noise in the pure Stackelberg commitment (which in this case constitutes a discrete shift to another pure strategy) can lead to a complete loss of commitment power. Subsequently, Van Damme and Hurkens [197] show that allowing the leader to use a mixed strategy, while only revealing the pure strategy realization to the follower[7], admits the existence of "robust" commitments that approach the pure-strategy-Stackelberg commitment as the noise vanishes. In the repeated setting, followers update their posterior based on leader observations as well as the common prior for the game – and how effective reputation can be depends both on observability and on the nature of the common prior. In particular, the prior has to have sufficient mass on the "Stackelberg commitment" type. In an extreme situation where the commitment type is instead a "bad commitment"[8] and the leader actions are imperfectly observed, Ely and Välimäki [199] show that followers can respond in a way that is sub-optimal for the leader, and thus reputation can be undesired. The intermediate situation where commitment types can be either good (e.g. Stackelberg) or bad is more nuanced – here, Ely, Fudenberg and Levine [192] show that a leader may still choose not to realize her reputation through repeated interaction with followers.

## 4.3 Towards a frequentist paradigm for repeated-game interaction

As we motivated in the introduction to this chapter, our principal aim is to move beyond an understanding of *why* agents achieve Stackelberg equilibrium, to *how* agents achieve Stackelberg equilibrium through repeated interaction. We take a frequentist perspective on modeling repeated-game interaction, and here we discuss our reasons for adopting this perspective.

In the repeated game setting, a frequentist modeling choice is a significant departure from the traditional frameworks that we have discussed above for understanding reputation building [182–186]. These are fundamentally *Bayesian*. In these Bayesian frameworks (as in our discussion in the statistical learning setup in Section 1.3), there is a prior that is commonly used by leader and all followers, over either *leader behavior*, or the *leader payoff matrix* as we have seen. This follows the "common prior" framework that was pioneered by Harsanyi [200] for tractability in studying repeated games with incomplete information. In contrast, we (along with [201]) assume no common prior and consider followers that use *frequentist* learning rules to infer about leader behavior.

[6]A subtle difference is that the trembling hand only affects the follower's response – the leader still realizes her payoff according to her original strategy [197]. Our one-shot model is similar.

[7]This is completely different from a mixed commitment, where the entire mixture is revealed.

[8]The class of games they define are "participation games", in which follower incentives are structured such that they do not wish to participate with leaders holding a "bad reputation".

The primary benefit of using a Bayesian framework with a common prior is that the Bayes-Nash equilibrium of the stage game in every round is then well-defined, and thus the *sub-game perfect Nash equilibria (SPNE)* of the repeated game are well-defined. The equilibrium payoff of the game for the leader can be analyzed *in the asymptotic limit* when the game is repeated infinitely with a discount factor approaching 1.

By eschewing a Bayesian micro-foundation with a common prior, we compromise on the ability to define and analyze a SPNE. We will see in Section 4.8 that the modeling choices in a frequentist framework for leader and follower strategies are conceptually highly non-trivial; consequently, our results make only partial headway towards defining a frequentist notion of SPNE. However, we elucidate below that a frequentist paradigm for repeated-game interaction introduces new promise that make it worth this difficulty in modeling. This promise is encapsulated in the form of three key desiderata central to modern statistics and optimal decision-making:

1. *Simplicity.* While the classic works [185, 186] recover the eventual attainment of leader power-of-commitment *to any mixed strategy*, extracting natural leader and follower strategies that form a SPNE is challenging to say the least. The proofs do invoke Bayesian confidence sets that capture the convergence of follower estimates of leader commitment to a neighborhood that appropriately shrinks as more rounds are played. However, this only tells us that mixed-strategy SPNEs *exist* that converge to the payoff afforded by a particular leader commitment in the asymptotic limit. This does not directly portray explicit strategies that leader and follower use in SPNE, or formalize the SPNE in a form that is easily computable. The frequentist framework will allow us to formalize remarkably simple rules for leader and follower that approximate the SPNE properties. These simple rules constitute a clear-cut distillation[9] of the ideas that manifest in the Bayesian setting with far more complex machinery as in [185].

2. *Understanding trade-offs.* The Bayesian framework does not easily answer certain key questions, such as whether it is *easier* for the leader to obtain Stackelberg payoff or some other commitment payoff through repeated interaction. This is because the paradigm is predominantly asymptotic, and it turns out that in asymptopia all commitments are equivalent in their learn-ability. However, this is a critical question when the leader and follower interact only for a finite number of rounds. Examples such as the product-choice game (Example 4.1 in the survey [194]) hint that the Stackelberg commitment might be "tougher" to achieve, in a certain sense, owing to possible instability in its perception; however, this intuition had not been formalized in prior work. Using the frequentist paradigm, we effectively uncover a fundamental *trade-off* that the leader needs to navigate to achieve commitment power. That is, we show that she chooses the commitment that she wishes to aspire to by balancing two key criteria: *closeness to the optimal Stackelberg commitment* and *preservation of follower learn-ability.* The

[9]Much in the same way that the maximum-likelihood-estimation principle is a distillation of the key ideas in maximum-a-posteriori estimation in classical statistics.

kernel of this result is derived in a much simpler one-shot setting with uncertainty in leader commitment perception (Theorem 4.7.4), but turns out to be pivotal to our understanding of the more complex repeated game.

3. *Robustness.* Central to formalizing results in the Bayesian framework is the *common prior* assumption as pioneered by Harsanyi [200]. It is not clear whether this assumption is reasonable in modern practice, nor if any of these guarantees are robust in the absence of this assumption (for a sample of philosophical discussions around these issues, see [202]). We should expect, as in classical statistics, that a sufficient number of samples of data will "drown out" the effect of a prior; the question is how to formalize this idea in game theory with the additional difficulties posed by multiple agents. While the study of prior mismatch is of natural interest, a successful frequentist approach would also make substantial process towards this goal by providing *prior-free* guarantees. Moreover, this approach provides tractable guarantees for a finite number of rounds, helping us formalize the idea of *partial reputation.*

The desiderata above have justified frequentist modeling choices in statistics theory[10]. They take on heightened relevance in the more challenging setting of repeated game theory with incomplete information. Taken together, they tell us that a frequentist paradigm for repeated game theory, while nascent in its conceptual development, forms an important and needed *complement* to the traditional Bayesian game theory. This is especially true for deployment of game-theoretic interaction in settings with automated agents, where critical understanding of explicit strategies followed by agents in equilibrium interaction is necessary.

While frequentist modeling of game-theoretic interaction remains highly challenging, it is worth noting that frequentist learning rules, mostly at the *heuristic level*, have seen classical precedent in game theory. In simultaneously repeated games with two-sided incomplete

[10]Another modern example at the heart of this debate is the multi-armed bandit (MAB) problem, which we discussed at length in Chapter 3. As a reminder, the MAB problem incorporates aspects of sequential decision-making with limited information feedback but in a purely statistical paradigm. The classical Bayesian paradigm formalizes the MAB problem as a partially observed Markov decision process (POMDP), and allows the formal definition of an exactly Bayes-optimal *index policy* [142, 143]. The flip-side is that these policies are complicated and not immediately interpretable. On the other hand, the frequentist paradigm formalizes optimality in terms of an asymptotic notion of sequential consistency [149, 203, 204] and/or non-asymptotic pseudo-regret minimization [205, Chapter 2]. While these notions of optimality are more flexible, they allow for the design of *simple* and approximately optimal heuristics like upper-confidence bounds [150] and Thompson sampling [148]. These heuristics highlight an important *exploration-exploitation trade-off* that is present in the index policies as well, but easier to see in the simple updates used by UCB and Thompson sampling. Accordingly, this exploration-exploitation trade-off has been applied in the far more challenging setup of reinforcement learning (not necessarily with optimality guarantees), for which the corresponding Bayesian POMDP is less tractable, see e.g. [206, 207]. Finally, the frequentist paradigm allows for learning algorithms for the MAB problem that are prior-free, thus *robust* to prior mis-specification. Much like in the repeated game setting, it is not immediately clear what the impact of prior mismatch would be on Gittins index policies. These considerations suggest that the frequentist and Bayesian paradigms have complemented each other in the MAB problem, and fundamental theoretical advances in both paradigms continue to be important.

| Result | One-shot model | Repeated model |
|---|---|---|
| Instability of Stackelberg/determinism | Proposition 4.7.1 | Theorem 4.9.5 |
| Robust commitments/leader rules | Theorem 4.7.4 | Proposition 4.9.2 |
| (Dis)incentive for unpredictability | Theorem 4.7.5 | Proposition 4.9.6 |

Table 4.1: Table of results.

information (which we will discuss at length in Chapter 5), these learning rules have yielded considerable success as plausible mechanisms for time-averaged convergence to relevant solution concepts [85, 208, 209]; particularly when they are designed for repeated strategic interaction with a potential adversary (formally, when they satisfy notions of Blackwell approachability/Hannan consistency/"no-regret") [80, 210, 211]. Remarkably, these learning rules are "uncoupled" [87] in the sense that players who use them do not utilize any information about their opponents' utility functions (which can be in a continuum) except through observed history/payoff. Clearly, we need to utilize the same "uncoupled" principle for followers in our learning model. Moreover, we allow followers essential *qualitative* flexibility in their choice of frequentist inference – in particular, they can:

1. Attempt exact, *pure-action* forecasts based on deterministic prediction from memory;
2. Forecast according to a statistical learning rule, *or*
3. Use the hedging principle in forecasting to avoid worst-case errors[11].

We know that each of these inference rules is best-suited for a leader rule that is:

1. Deterministically predictable,
2. Statistically learn-able *and*
3. Maximally unpredictable.

We will show in the next few sections how a frequentist paradigm with the above (statistically speaking, composite) skeleton for leader and follower rules helps us move from understanding *why* reputation is built to understanding *how* it is built. As a consequence of well-known, clean non-asymptotic characterizations of frequentist learning rules, we are also able to obtain a precise, quantitative understanding of the power of *partial reputation*.

## Our contributions

Now that we have motivated a frequentist paradigm, we summarize the aggregate of our results[12] *constructs* a robust leader mechanism for building a mixed strategy reputation, and characterizes the amount of partial reputation power that a leader thus enjoys after interacting with a finite number of myopic followers. Our mechanism is both strategy-proof to follower manipulation, and *approximately* optimal for the leader subject to sensible follower rules.

We build to this conclusion by first providing a comprehensive analysis of the one-shot setting, i.e. a Stackelberg leader-follower game in which a follower obtains a limited number of observations of the leader commitment. We note that for almost all non-zero-sum games (Proposition 4.7.1 in Section 4.7) the payoff of the classical, mixed Stackelberg commitment is not robust to even an *infinitesimal amount of* observational uncertainty[13]. Next, we propose robust commitment rules (Theorem 4.7.4 in Section 4.7) for leaders and show that we can approach the Stackelberg payoff as the number of observations increases. The robust commitment construction involves optimizing a trade-off between robustly preserving the follower best response and staying close to the ideal Stackelberg commitment, by moving the commitment a little bit into the interior of an appropriate convex polytope. Finally, we show that any possible advantage for the leader from limited observability is only related to follower response mismatch, and show that this advantage is limited (Theorem 4.7.5 in Section 4.7). This implies that a leader cannot gain much by misrepresenting her commitment and eliciting a sub-optimal response from the follower.

Next, we provide analogs of these results in the more challenging setting of repeated interaction between one leader and several myopic followers, where both observational uncertainty and uncertainty in belief are present. In a similar spirit to the conclusion about the Stackelberg commitment being unstable to observational uncertainty, we show that leader strategies that maximize payoff against followers that naively use a statistical learning rule are extremely unstable against more intelligent followers (Theorem 4.9.5 in Section 4.9). We show this result for two broad ensembles[14] of Stackelberg games. We generalize the idea in the robust commitment rules from Theorem 4.7.4 to provide an explicit *randomized* leader mechanism that builds reputation through repeated interaction (Proposition 4.9.2 in Section 4.9). We note that under this mechanism the follower is not incentivized to de-

[11] Contrary to intuition in zero-sum-game settings, the no-regret principle by itself may not be the best follower learning rule under all circumstances. In fact, in a recent study [212] the optimal leader payoff against a follower who follows certain kinds of no-regret strategies was shown to sometimes be greater than Stackelberg, suggesting that no-regret by itself is not necessarily a desirable learning rule for the follower.

[12] Very early preliminary versions of this work were presented at NeurIPS '17 in the workshop on "Learning in the presence of strategic behavior" and this chapter constitutes the expanded and extended form of work that was presented at ACM Economics and Computation '19.

[13] A similar phenomenon has been observed for trembling hand noise in pure strategy commitments [197]. A succinct statement of our result is that *all mixed strategies tremble by their very nature when finitely observed.*

[14] These ensembles reflect the reality of security games and persuasion respectively.

viate from a natural statistical learning rule. Finally, in the spirit of Theorem 4.7.5 from the one-shot model, we show that, no matter how sophisticated her mechanism, the leader has minimal incentive, and possible disincentive, to deceive followers who use a *universally calibrated* learning rule[15] (Proposition 4.9.6 in Section 4.9). Correspondences between our results in the one-shot and repeated models are presented in Table 4.1.

As we motivated in Section 4.3, the results for repeated interaction are presented in the framework of a new model that departs from Bayesian tradition, and instead consider followers who use frequentist principles for their inference. This modeling choice gave us tractability into the nature of leader and follower strategies that lead to reputation building, including at finite stages of interaction.

## 4.4 Warm-up: One-shot game with partially revealed commitment

There are several steps involved to get to the eventual frequentist understanding of the repeated game that we recapped above. We will heavily build on our clear and complete understanding for the much simpler one-shot setting, which we now describe.

### Preliminaries

We represent a two-player leader-follower game in normal form by the pair of $d \times n$ matrices $(A, B)$, where $A \in \mathbb{R}^{d \times n}$ denotes the leader payoff matrix and $B \in \mathbb{R}^{d \times n}$ denotes the follower payoff matrix. We denote the leader mixed strategy space by $\Delta_d$ (where $\Delta_k$ for any $k$ represents the $k$-dimensional probability simplex) and the follower mixed strategy space by $\Delta_n$. From now on, we define an *effective dimension* of a game as a number $m < d$ for which the effective payoff matrices of leader and follower respectively are $A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$, $B = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$, and the effective set[16] of leader strategies is given by a convex polytope $K \subseteq \Delta_m$.

We consider a setting of *asymmetric private information* in which the leader knows about the follower preferences (i.e. she knows the matrix $B$) while the follower does not know about the leader preferences[17] (i.e. he possesses no knowledge of the matrix $A$).

In the traditional setting, the leader has committed to some mixed strategy $\mathbf{x} \in \Delta_m$. The follower believes in leader commitment and can perfectly observe the commitment $\mathbf{x}$.

[15] Such a learning rule also satisfies the property of no-internal-regret, but the guarantee of calibration is a strictly stronger one.

[16] This definition is important for Stackelberg security games, in which the defender strategy space looks exponential in the number of targets $m$ - but the actual manifestation of all leader strategies is in fact $m$-dimensional. In particular, a defender strategy manifests as a distribution over different targets being covered.

[17] This is an important assumption for this chapter, and is critical to the traditional reputation building framework [182, 183].

We denote the follower's set of theoretically best *pure-strategy* responses to a mixed strategy commitment $\mathbf{x}$ by $\mathcal{K}^*(\mathbf{x}) \subseteq [m]$. We have

$$\mathcal{K}^*(\mathbf{x}) := \arg\max_{j \in [n]} \langle \mathbf{x}, \mathbf{b}_j \rangle.$$

We make an important assumption that has been used in classical literature to define the existence of a Stackelberg commitment [213, 214]. This assumption is as follows: when the set $\mathcal{K}^*(\mathbf{x})$ has multiple pure strategies, the follower responds with the pure strategy in the set $\mathcal{K}^*(\mathbf{x})$ that is most favorable to the leader[18]. That is, the follower responds with pure strategy

$$j^*(\mathbf{x}) := \arg\max_{j \in \mathcal{K}^*(\mathbf{x})} \langle \mathbf{x}, \mathbf{a}_j \rangle.$$

We define *best-response regions* as the set of leader commitments that would elicit the pure strategy response $j$ from the follower, i.e. $\mathcal{R}_j := \{\mathbf{x} \in K : j^*(\mathbf{x}) = j\}$.

With these definitions, we can define the leader's ideal payoff when her commitment is fully revealed:

**Definition 4.4.1.** *A leader who commits and credibly reveals mixed strategy $\mathbf{x} \in \Delta_m$ should expect payoff*

$$f_\infty(\mathbf{x}) := \langle \mathbf{x}, \mathbf{a}_{k^*} \rangle.$$

*Therefore, the leader's **Stackelberg payoff** is the solution to the program*

$$f^*_\infty := \max_{\mathbf{x} \in \Delta_m} f_\infty(\mathbf{x}).$$

*The argmax of this program (well-defined because of our tie-breaking assumption) is denoted as the **Stackelberg commitment** $\mathbf{x}^*_\infty$. Further, we denote the best response faced in Stackelberg equilibrium by $j^* := j^*(\mathbf{x}^*_\infty)$.*

The Stackelberg commitment is optimal for the leader under two conditions: *the follower* 100% *believes the leader is committed to a fixed strategy*, and *the follower knows exactly the leader's committed-to strategy.* For a finite number of interactions between leader and followers, neither is true.

[18]The technical reason for this tie-breaking rule is to be able to define the Stackelberg commitment as an explicit *maximizer* (which in itself gives a subtle clue to its fragility). Interestingly, positive results in the Bayesian repeated-game formulation of reputation [185, 186] assume that the follower breaks ties *against* the favor of the leader (thus, they do not explicitly consider Stackelberg commitments, only the ideal Stackelberg payoff which can be defined as a sup instead of a max). We will present our results throughout with the assumption of follower tie-breaking in the favor of the leader, but will comment on how they might vary if the follower instead responded with a mixture among the strategies in $\mathcal{K}^*(\mathbf{x})$.

## 4.5 One-shot game: Finite observability of commitment

We start with the simpler one-shot Stackelberg game played between one leader and *one* follower. In this game, there is a shared belief in commitment, but there is uncertainty in how the commitment is revealed. In particular, the follower does not know the exact (mixed) strategy that the leader has committed to – he can only see a finite number of its realizations.

### Commitment uncertainty model

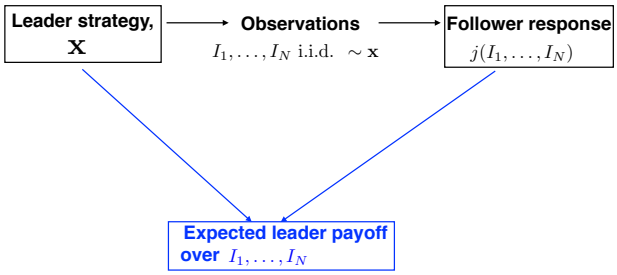



Figure 4.3: Illustration of one-shot Stackelberg game between leader and *one* follower who observes $N$ noisy samples of leader commitment, instead of the commitment itself. Figure from [90].

We adopt a model in which the leader plays first, but can only reveal her commitment $\mathbf{x}$ through $N$ *pure strategy plays* $I_1, I_2, \ldots, I_N$ i.i.d $\sim \mathbf{x}$. The commitment is known (to both leader and follower) to come from a set of mixed strategies $\mathcal{X} \subseteq K$. This model was first studied empirically for Stackelberg security games [215], but with a common prior on the leader payoff matrix $A$. This model was more recently studied for zero-sum games [201] without a prior, exactly like our setting.

The follower can respond with a pure strategy $j_N(I_1, I_2, \ldots, I_N)$, whose choice we will specify shortly. Importantly, the follower response function $j_N(\cdot)$ can only depend on the leader realizations: not the true commitment $\mathbf{x}$ or the true leader payoff matrix $A$. After the follower chooses his response, the leader and follower payoffs are realized in expectation. The sequential nature of moves by leader and follower, after which payoffs are realized, is depicted in Figure 4.3.

We can express the expected[19] leader payoff as a function of her chosen commitment for any follower best response function.

[19]Expectations over utility have been implicitly taken. All expectations thereof in mathematical notation are over the additional randomness in the realizations $I_1, \ldots, I_N$ unless specified otherwise.

**Definition 4.5.1.** *A leader that can express $N$ realizations of her hidden commitment $\mathbf{x}$, faced with a follower using response function $j_N(\cdot)$, expects payoff*

$$f_N(\mathbf{x}) := \mathbb{E}[\langle \mathbf{x}, \mathbf{a}_{j_N(I_1,\ldots,I_N)} \rangle]. \tag{4.1}$$

## Follower response model

A reasonable follower response will constitute some sort of statistical inference about the mostly likely identity of the leader commitment from the $N$ *samples* provided, and an "optimal" response based on this inference. We make this precise through frequentist and Bayesian models for inference.

**Frequentist model:** In the frequentist model, which was first studied in [201] for zero-sum security games, we do not assume a prior on the leader commitment. We denote the maximum likelihood estimate (MLE) of the leader's mixed strategy, as seen by the follower, by $\widehat{\mathbf{X}}_N$. It is reasonable to expect, under certainty of leader commitment, that a "rational"[20] follower would best-respond to $\widehat{\mathbf{X}}_N$, i.e. play the pure strategy

$$j^*_{\mathsf{F}}(I_1, \ldots, I_N) := j^*(\widehat{\mathbf{X}}_N). \tag{4.2}$$

Thus, the leader should expect payoff

$$f_N(\mathbf{x}) := \mathbb{E}[\langle \mathbf{x}, \mathbf{a}_{j^*_{\mathsf{F}}(\widehat{\mathbf{x}}_N)} \rangle]$$

against a follower that response according to Equation (4.2). The maximal payoff a leader can expect is

$$f^*_N := \max_{\mathbf{x} \in \Delta_m} f_N(\mathbf{x}) \tag{4.3}$$

and it acquires this payoff by playing the argmax strategy $\mathbf{x}^*_N$. Observe that the objective function $f_N(\mathbf{x})$ is not convex in its argument $\mathbf{x}$, and is thus NP-hard to exactly maximize.

**Bayesian model:** We provide an expected-utility-maximization interpretation for the follower best response in Equation (4.2) by considering an equivalent Bayesian model for follower inference. The latter model was empirically studied in [215]. We assume that the follower starts with a prior $P_0(\cdot)$ over all $m$-dimensional multinomial distributions for the identity of the leader commitment (and the leader is aware of this). We denote
$P_N(\mathbf{x}; (I_1, I_2, \ldots, I_N))$ as the posterior probability that the leader chose commitment $\mathbf{x}$ after $N$ observations. We assume that the follower starts with a prior $P_0(\cdot)$ on the identity of the leader commitment (and the leader knows the prior). Then, the follower will respond with

[20] Rational is in quotes because expected utility theory does not have a concrete meaning in frequentist inference of an unknown leader commitment. However, the best estimate of this unknown commitment, given no prior information, is clearly the MLE.

the pure strategy that maximizes his expected utility under the posterior:

$$j^*_{\mathsf{B}}(I_1, \ldots, I_N) := \arg\max_{j \in [n]} \int_{\mathbf{x}' \in \Delta_d} P_N(\mathbf{x}'; (I_1, \ldots, I_N)) \langle \mathbf{x}', \mathbf{b}_j \rangle$$
$$= \arg\max_{j \in [n]} \langle \overline{\mathbf{x}}_N, \mathbf{b}_j \rangle \text{ where}$$
$$\overline{\mathbf{x}}_N(I_1, \ldots, I_N) := \mathbb{E}_{\mathbf{x}' \sim P_N(\cdot; (I_1, \ldots, I_N))} \mathbf{x}'.$$

Notice that the special case of $P_0(\cdot)$ being the Dirichlet prior with hyper-parameter $\boldsymbol{\alpha} \to \mathbf{0}$ corresponds to $\overline{\mathbf{x}}_N(I_1, \ldots, I_N) = \widehat{\mathbf{X}}_N$, and in this case $j^*_{\mathsf{B}}(I_1, \ldots, I_N) = j^*_{\mathsf{F}}(I_1, \ldots, I_N)$, i.e. the follower best response under the frequentist and Bayesian models is the same[21]. This gives a *post-hoc* justification for using the frequentist model through expected utility theory for a specific prior on the commitments[22]. Having established this equivalence, we henceforth use the frequentist model, which will prove to be much more tractable in our forthcoming model of repeated interaction.

Under the model of observational uncertainty, we are interested in characterizing the optimal leader commitment as well as optimal leader performance. That is, we want to understand how close $f^*_N$ is to $f^*_\infty$, and also how close $\mathbf{x}^*_N$ is to $\mathbf{x}^*_\infty$. An answer to the former question would tell us how observational uncertainty impacts the first-player advantage. An answer to the latter question would shed light on whether the best course of action deviates significantly from Stackelberg commitment. We are also interested in algorithmic techniques for *approximately computing* the quantity $f^*_N$, as doing so exactly would involve solving a non-convex optimization problem.

## 4.6 Related work in the one-shot setting

The *mixed-strategy* Stackelberg solution concept has seen contemporary application in engineering [181, 216] and persuasion mechanisms [188]. Here, the entire mixture has to be credibly revealed to the follower for the Stackelberg solution concept to be realized. Von Stengel and Zamir [187] show that in all general sum games, mixed strategy Stackelberg equilibrium is beneficial over simultaneous (Nash) equilibrium when the commitment can be fully revealed, and often strictly so. We explore two examples where the mixed nature of the Stackelberg commitment is critical. First, the *Stackelberg security game* [181] is played between a defender, who deploys a randomized protection strategy on her targets; and an attacker, who observes the defender *mixed strategy* and responds (at a high level) by attacking the target(s) that are most likely to be left open. Second, Kamenica and Gentzkow's

[21] More precisely, the leader payoff in the Bayes-Nash equilibrium of the Bayesian game is exactly equal to $f^*_N := \max_{\mathbf{x} \in \Delta_m} f_N(\mathbf{x})$.

[22] While all our results are derived for the Dirichlet prior which is maximally uninformative, we expect the same scaling to hold under any informative prior, although the nature of the constants will change to reflect the informativity of the prior. This will be the case for any prior that is positively supported on *all* leader mixed strategies (also assumed by Fudenberg and Levine [185]).

*Bayesian persuasion game* [188], in which the sender has the ability to persuade a receiver to act in a manner that is beneficial to the sender. She does this by committing to a randomized signaling mechanism conditioned on information that is private to her[23]. In general, this persuasion power is effective only when the commitment is mixed.

For both these cases, a mixed commitment will not be fully revealed or believed. In security games, the attacker will usually observe a finite number of deployments of the defender's resource, as opposed to the allocation strategy itself (which is often mixed). Persuasion power could be credibly built up through a sender's repeated interaction with multiple receivers – what will actually be observed is her history, and thus *realizations* of the signal, not the distribution itself. In addition, the receivers have no a-priori reason to believe that the sender is indeed committed to a fixed signaling mechanism. Thus, we should expect that the sender realizes her persuasion power *only partially.*

It is useful to review algorithmic perspectives on mixed-strategy Stackelberg computation[24]. Conitzer and Sandholm [214] show that computing the optimal mixed commitment under perfect observability corresponds to a linear program. In contrast, the problem of computing the optimal commitment under *finitely* limited observability corresponds to a robust optimization problem that is, in fact, NP-hard [215, 220]; so reasoning about the optimal commitment in the presence of noise is algorithmically non-trivial. An et al and Shieh et al [215, 220] consider a model of observational uncertainty with a Bayesian prior and posterior update based on samples of behavior, and propose heuristic algorithmic techniques to compute the optimum. They show empirically that there could be a positive return over and above the Stackelberg payoff. We will show that such positive return is indeed possible, but the amount of gain is limited and dissipates rapidly as more observations are available. Blum, Haghtalab and Procaccia [201] prove that the Stackelberg commitment itself approximates the optimal payoff in the special case of zero-sum games. We show that this reasoning does not extend to the general sum case and that the Stackelberg commitment is generically unstable.

The problem of *communication constraints* in the commitment has also received a lot of interest in the recent algorithmic persuasion literature with different models for the uncertainty. Such noise models include compression in the mechanism [221, 222], and binary uncertainty in the mixed commitment either being fully observed or fully hidden [223]. Uncertainty from the point of view of bounded follower rationality [224] has also been considered. The primary distinction in our work (as well as the settings of An et al, Shieh et al and Blum et al), is that *the manifestation of the uncertainty is itself random.* Thus, a unique component of our results involves directly reasoning about the stochasticity of the follower response.

[23] Kamenica and Gentzkow study persuasion in the most general form to date. The precursor to these persuasion mechanisms was signaling games [217], and in fact, one of the examples of mixed strategy reputation as studied by Fudenberg and Levine [185] bears similarity to the prosecutor-judge example. Recently, Ely [218] has considered an extension to dynamic persuasion mechanisms, where the private information can evolve in a stochastic manner.

[24] For an excellent survey, see [219].

Finally, limited observability in leader commitment necessitates follower inference. The flipped problem, in which a leader does not know the follower's utility function and needs to learn her optimal commitment, is also interesting and has seen a lot of recent activity. Approaches on how a leader should do this range from learning adversary models from historical data in security games [225, 226] to no-regret online learning approaches [227]. The former approach can get closer to optimal commitment power if the samples from followers are stochastic [225], but could be sub-optimal against followers who attempt to exploit the learning process. No-regret learning approaches are robust to such follower exploitation, but cannot use historical data. Other paradigms with an incumbent and myopic agents that involve a learning problem for all parties have also been explored. One prominent example is the setting of Bayesian exploration [228], in which there is private information unknown to all players. In this setting, an incumbent commits to a mechanism that incentivizes myopic agents to carry out partial exploration to learn about the private information, as opposed to only exploit their current knowledge.

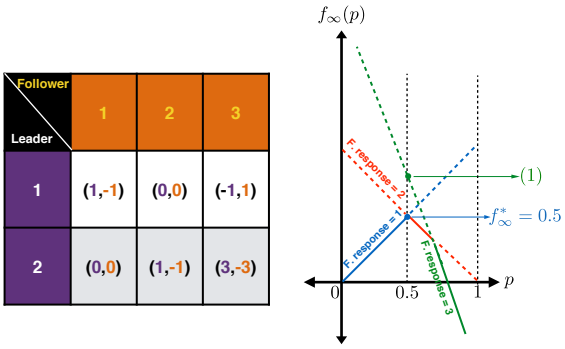
## 4.7 Results for the one-shot model with limited observability

We now analyze the one-shot model that was defined in Section 4.5. The salient features of this model are as follows: the leader fixes a commitment $\mathbf{x}$ in advance, reveals $N$ iid realizations of this commitment, and the follower best responds according to a maximum-likelihood inference rule. We wish to understand the leader's optimal payoff in the limited-observation model as well as how she can approximately achieve it.

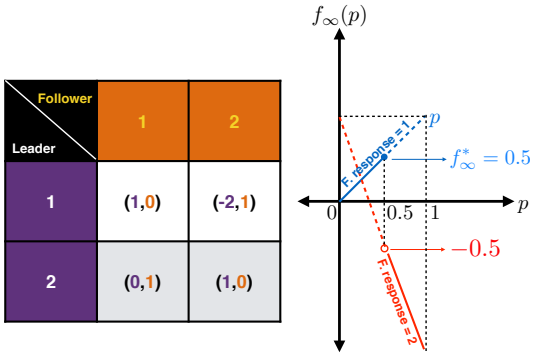
### Instability of traditional Stackelberg commitment to observational uncertainty

A natural first question is whether the traditional Stackelberg commitment, which is clearly optimal if the game were being played infinitely (or equivalently, if the leader had infinite commitment power and perfect public commitment), is also suitable for finite play. We show through a few paradigmatic examples that the answer can vary. Figure 4.4 depicts the standard normal form representation of the leader-follower Stackelberg game [214] for these illustrations, as well as illustrates the structure of the *ideal leader payoff function* $f_\infty(\cdot)$. We will see that practically all insight into the limited-observation payoff function $f_N(\cdot)$ can be characterized by the structure of the ideal leader payoff function $f_\infty(\cdot)$, and we made these illustrations to convey intuition about the nature of the results.

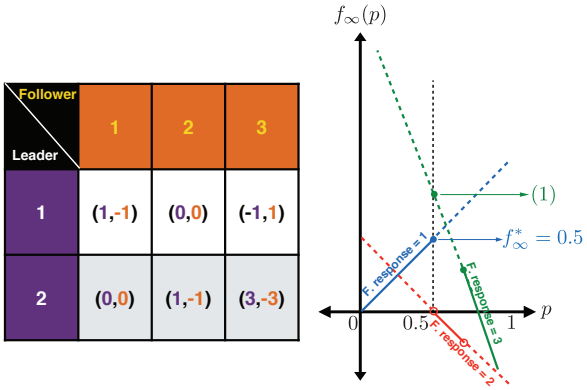
**Example 2.** *We consider a* $2 \times 3$ *zero-sum game, represented in normal form in Figure 4.4a. We can express the leader strategy according to the probability* $p$ *with which she will pick*




(a) $2 \times 3$ zero-sum game.




(b) $2 \times 2$ non-zero-sum game.




(c) $2 \times 3$ non-zero-sum game.

Figure 4.4: Illustration of examples of zero-sum game and non-zero-sum games in the form of normal form tables and ideal leader payoff function $f_\infty(\cdot)$. We denote the probability that the leader will play strategy 1 by $p \in [0, 1]$, and fully describes leader mixed commitment for $2 \times n$ games. Observe that the ideal leader payoff function $f_\infty(p)$ is piece-wise-affine in $p$, and for the non-zero-sum games *discontinuous* at the Stackelberg commitment $p^*_\infty$. Figures from [90].

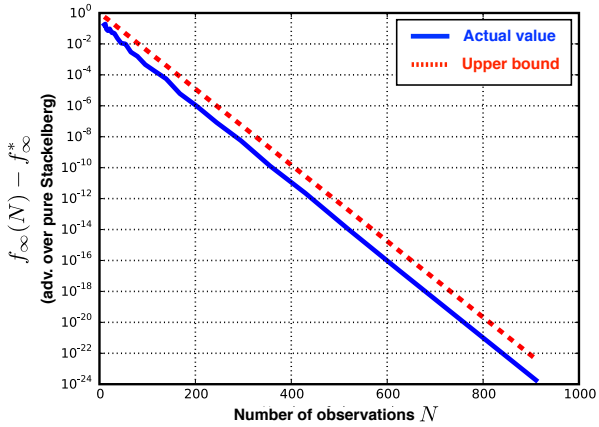



Figure 4.5: Semi-log plot of extent of advantage over Stackelberg payoff as a function of $N$ in the $2 \times 3$ *zero-sum* game depicted in Figure 4.4a. Figure from [90].

*strategy* 1, *and the leader payoff* $f_\infty(p)$ *is as follows:*

$$\begin{cases} f(p;1) := p \text{ if follower best responds with strategy 1} \\ f(p;2) := 1-p \text{ if follower best responds with strategy 2} \\ f(p;3) := 3-4p \text{ if follower best responds with strategy 3} \end{cases}$$

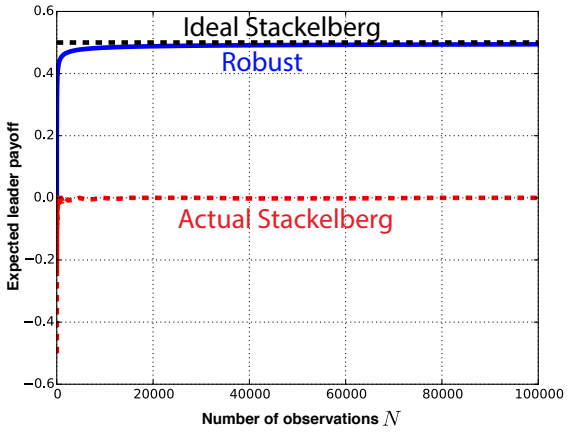
*Since the game is zero-sum, the follower responds in a way that is worst-case for the leader. This means that we can express the leader payoff as*

$$f_\infty(p) = \min\{f(p;1), f(p;2), f(p;3)\}. \tag{4.4}$$

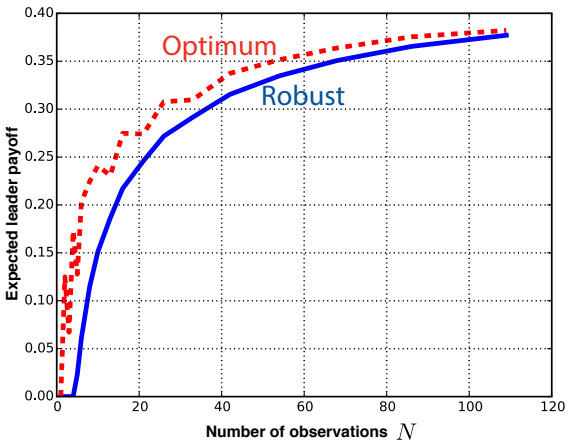
*This leader payoff structure is depicted in Figure 4.4a. Notice that for this example, the ideal leader payoff function* $f_\infty(p)$ *is continuous in* $p$ *– this is because it is the minimum of three affine functions in* $p$ *given by* $f(p;1)$, $f(p;2)$ *and* $f(p;3)$. *We can express the Stackelberg payoff as*

$$f^*_\infty = \max_{p\in[0,1]} f_\infty(p) = 1/2,$$

*attained at* $p^*_\infty = 1/2$. *We wish to evaluate* $f_N(p^*_\infty)$. *It was noted [201] that* $f_N(p^*_\infty) \geq f^*_\infty(p^*_\infty)$ *by von Neumann's minimax theorem, but not always clear whether strict inequality would hold (that is, if observational uncertainty gives a strict advantage). The semi-log plot in Figure 4.5 shows that the extent of improvement decreases exponentially with* $N$. *A simple calculation*




(a) Performance of the sequence of robust commitments $\{\mathbf{x}_N\}_{N\geq 1}$ and the Stackelberg commitment $\mathbf{x}^*_\infty$ as a function of $N$.




(b) Performance of sequence of robust commitments $\{\mathbf{x}_N\}_{N\geq 1}$ as compared to the optimum performance $f^*_N$ (brute-forced).

Figure 4.6: Example of the $2 \times 2$ non-zero-sum game depicted in Figure 4.4b, for which observational uncertainty is always undesirable. Figures from [90].

*yields*

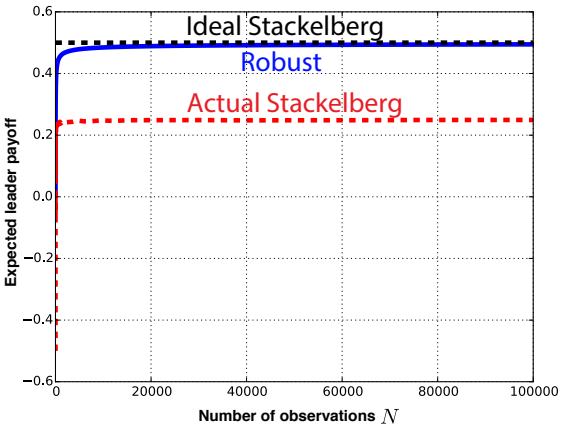
$$\begin{aligned} f_N(1/2) &= \Pr[\widehat{P}_N \leq 2/3] f^*_\infty + \Pr[\widehat{P}_N > 2/3] f(1/2;3) \\ \implies f_N(1/2) - f^*_\infty &= \Pr[\widehat{P}_N > 2/3]\,(f(1/2;3) - f^*_\infty) \\ &= \frac{1}{2}\Pr[\widehat{P}_N > 2/3] \\ &= \frac{1}{2}\Pr[\widehat{P}_N - 1/2 > 1/6] \\ &\leq \exp\{-N D(2/3 \parallel 1/2)\} \end{aligned}$$

*where $D(. \parallel .)$ denotes the Kullback-Leibler divergence, and the last inequality is due to Sanov's theorem [229]. The reason that the advantage decreases exponentially with $N$ is because as $N$ increases, we see a decrease in the effective stochasticity that sometimes elicits the more favorable follower response, i.e. action 3.*

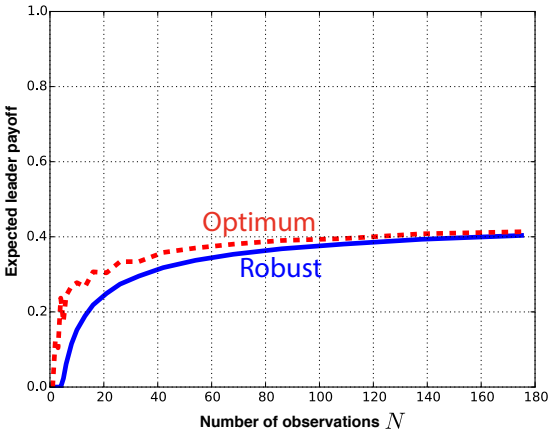
Example 2 showed us how the traditional Stackelberg commitment power could be increased, albeit by a small amount, by occasionally eliciting more favorable responses. We now provide an example illustrating that the commitment power can disappear completely.

**Example 3.** *We consider a $2\times 2$ non-zero-sum game, represented in normal form and leader payoff structure in Figure 4.4b. Here, the ideal leader payoff function is*

$$f_\infty(p) = \begin{cases} p \text{ if } p \leq 1/2 \\ 1 - 3p \text{ if } p > 1/2. \end{cases}$$




(a) Performance of the sequence of robust commitments $\{\mathbf{x}_N\}_{N\geq 1}$ and the Stackelberg commitment $\mathbf{x}^*_\infty$ as a function of $N$.




(b) Performance of sequence of robust commitments $\{\mathbf{x}_N\}_{N\geq 1}$ as compared to the optimum performance $f^*_N$ (brute-forced).

Figure 4.7: Example of the $2 \times 3$ non-zero-sum game depicted in Figure 4.4c, in which observational uncertainty could either help or hurt the leader. Figures from [90].

*This is very close to the example provided by [201], which we repeat for storytelling value. Notice that $f^*_\infty = 1/2, p^*_\infty = 1/2$, but the advantage evaporates with observational uncertainty. For any finite (odd) $N$, we have*

$$\begin{aligned} f_N(1/2) &= \Pr[\widehat{P}_N \leq 1/2](1/2) + \Pr[\widehat{P}_N > 1/2](-1/2) \\ &= 1/2 \times 1/2 - 1/2 \times 1/2 = 0. \end{aligned}$$

*This implies that $f^*_\infty - f_N(p^*_\infty) = 1/2$ and so $\lim_{N\to\infty} f^*_\infty - f_N(p^*_\infty) \neq 0$. This is clearly a very negative result for the robustness of Stackelberg commitment, and tells us that the traditional mixed Stackelberg commitment $p^*_\infty$ is far from ideal under limited observability. In this example, stochasticity in follower response is not desired, principally because of the discontinuity in the leader payoff function at $p^*_\infty$, which can be clearly seen in Figure 4.4b.*

Example 3 displayed the possibility of a significant disadvantage of observational uncertainty – this disadvantage arose from the sizable probability of a mismatched response (follower response 2 instead of follower response 1). The game considered was special in that there was no potential for *gain* from a mismatched response, while in a zero-sum game like Example 2, a mismatched response is always favorable to the leader. Our next example generalizes these cases and provides insight into what could happen for a general non-zero-sum game.

**Example 4.** *Our final example considers a* $2 \times 3$ *non-zero-sum game, whose normal form and leader payoff structure are depicted in Figure 4.4c. The ideal leader payoff function is*

$$f_\infty(p) = \begin{cases} p \text{ if } p \leq 1/2 \\ 1/2 - p \text{ if } 1/2 < p \leq 5/7 \\ 3 - 4p \text{ if } p > 5/7. \end{cases}$$

*As in the other examples,* $f^*_\infty = 1/2, p^*_\infty = 1/2$*. Notice that this example captures both positive and negative effects of stochasticity in response. On one hand, follower response* 2 *is highly undesirable (a la Example 3) but follower response* 3 *is highly desirable (a la Example 2). The net effect is*

$$\begin{aligned} f_N(\frac{1}{2}) &= \frac{1}{2} \Pr\left[\widehat{P}_N \leq \frac{1}{2}\right] + \Pr\left[\frac{1}{2} < \widehat{P}_N < \frac{5}{7}\right](0) + \Pr\left[\widehat{P}_N \geq \frac{5}{7}\right](1) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \Pr\left[\widehat{P}_N \geq \frac{5}{7}\right](1) \\ &\leq \frac{1}{4} + \frac{1}{2} \exp\left\{-ND(\frac{5}{7} \parallel \frac{1}{2})\right\}. \end{aligned}$$

*A quick calculation thus tells us that* $f_N(p^*_\infty) <= f^*_\infty$ *if* $N \geq 8$*, showing that Stackelberg in fact has poor robustness for this example. Intuitively, the probability of the "bad" stochastic event remains constant while the probability of the "good" stochastic event decreases exponentially with* $N$*. Even more damningly, we see that* $\lim_{N\to\infty} f^*_\infty - f_N(p^*_\infty) \geq \lim_{N\to\infty} \frac{1}{4} - \frac{1}{2}\exp\{-ND(5/7 \parallel 1/2) = 1/4$*, again showing that the traditional Stackelberg commitment is far from ideal.*

While the three examples detailed above provide differing conclusions, there are some common threads. For one, in all the examples, committing to the Stackelberg mixture $\mathbf{x}^*_\infty$ can result in the follower being agnostic between more than one response. For both the non-zero-sum game examples, a very slight mis-perception in the estimation of the true strategy $\mathbf{x}^*_\infty$ led to a different, worse-than-expected response and this mis-perception happened with a sizeable, non-vanishing probability[25]. On the flip-side, a different response could also lead to better-than-expected payoff, raising the potential for a gain over and above $f^*$. However, these *better-than-expected responses* cannot share a boundary with the Stackelberg commitment, and we will see that the probability of eliciting them decreases exponentially

[25]Note that this mis-perception is *not* detrimental to the follower when the leader is in fact playing traditional Stackelberg commitment, because the follower is actually agnostic between all of these responses. The leader, on the other hand, very much cares how the follower chooses his response – which manifests in the discontinuity of $f_\infty(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^*_\infty$. This discontinuity is fundamental to the fact that the leader and follower's utilities are neither exactly aligned or anti-aligned (as they would be in a zero-sum game [201]). Since there are a finite number of neighboring follower responses, the gap between the leader's utility from each of them will be a non-zero number — leading to a jump discontinuity.

with $N$. The net effect is that the Stackelberg commitment is, most often, not robust – and this is even the case for small amounts of uncertainty.

Our first result is a formal statement of the instability of traditional, *mixed* Stackelberg commitments for a general $2 \times n$ game. We denote the leader probability of playing strategy 1 by $p \in [0, 1]$, and the Stackelberg commitment's probability of playing strategy 1 by $p^*_\infty$. Furthermore, let $\phi(t)$ denote the CDF of the standard normal distribution $\mathcal{N}(0, 1)$. We are now ready to state the result.

**Proposition 4.7.1.** *For any $2 \times n$ leader-follower game in which $p^*_\infty \in (0, 1)$ and $f_\infty(p)$ discontinuous at $p = p^*_\infty$, we have*

$$f_N(p^*_\infty) \leq f^*_\infty - C\left(\phi(\sqrt{N}C') - \frac{1}{2} - \frac{C'}{\sqrt{N}}\right) + C'' \exp\{-NC'''^2\} \tag{4.5}$$

*where $C, C', C'', C'''$ are strictly positive constants depending on the parameters of the game. This directly implies the following:*

1. *For some $N_0 > 0$, we have $f_N(p^*_\infty) < f^*_\infty$ for all $N > N_0$.*
2. *We have $\lim_{N\to\infty} f_N(p^*_\infty) < f^*_\infty$.*

The proof of Proposition 4.7.1 is contained in Appendix 4.11. The technical ingredients in the proof are the Berry-Esseen theorem [230, 231], used to show that the detrimental alternate responses on the Stackelberg boundary are non-vanishingly likely – and the Hoeffding bound, used to tail bound the probability of potentially beneficial alternate responses not on the boundary[26].

As we noted earlier, Proposition 4.7.1 represents a *mixed-commitment analog* of the known instability of pure strategy Stackelberg equilibria to "trembling hand" noise [197, 198]. A succinct description of this result is that *"all strictly mixed commitments tremble by their nature"*. The necessary and sufficient conditions that we characterize for this property to hold are remarkably general: one, that there is a discontinuity at the Stackelberg boundary, and two, that the Stackelberg commitment is mixed. Games for which one of these conditions *does not hold* fall into one of two special classes, each of which we remark on below.

**Remark 4.7.2.** *For games in which the mixed-strategy Stackelberg equilibrium coincides with a pure strategy, the follower's best response is always as expected regardless of the number of observations. There is no trade-off and it is simply optimal to play Stackelberg even under observational uncertainty.*

[26]It is worth noting that a similar argument as presented here could be extended to a general $m \times n$ game, using iid random vectors instead of random variables and considering a demarcation into best-response regions as illustrated in Figure 4.15. We only restrict attention to the $2 \times n$ case for ease of exposition. Note that our explicit constructions in Theorem 4.7.4 are defined for a general $m \times n$ game.

**Remark 4.7.3.** *For the zero-sum case (as in Example 2), it was observed [201] that a Stackelberg commitment is made assuming that the follower will respond in the worst case. If there is observational uncertainty, the follower can only respond in a way that yields payoff for the leader that is better than expected. This results in an expected payoff greater than or equal to the Stackelberg payoff $f^*_\infty$, and it simply makes sense to stick with the Stackelberg commitment $\mathbf{x}^*_\infty$. As we have seen, this logic does not hold up for non-zero-sum games because different responses can lead to worse-than-expected payoff. One way of thinking of this is that the function $f_\infty(.)$ can generally be discontinuous in $\mathbf{x}$ for a non-zero-sum game, but is always continuous for the special case of zero-sum.*

For a zero-sum game, the first condition does not hold (as can clearly be seen in Figure 4.4a); and for a game where the Stackelberg commitment happens to be pure, the second condition does not hold.

Proposition 4.7.1 directly implies that the ideal Stackelberg payoff is only obtained for the exact case of $N = \infty$ (when the commitment is perfectly observed), and that for any value of $N < \infty$ there is a *non-vanishing reduction* in payoff. In the simulations in Section 4.7, we will see that this gap is empirically significant.

## Robust commitments achieving close-to-Stackelberg performance

The message of Proposition 4.7.1 is that, in general, the traditional Stackelberg commitment $\mathbf{x}^*_\infty$ is undesirable. The mixed commitment $\mathbf{x}^*_\infty$ is pushed to an *extreme point* of the best-response-region $\mathcal{R}_{j^*}$ to ensure optimality under idealized conditions; and this is precisely what makes it sub-optimal under uncertainty. What if we could move our commitment a little bit into the interior of the region $\mathcal{R}_{j^*}$ instead, such that we can get a *high-probability-guarantee* on eliciting the expected response, while staying sufficiently close to the idealized optimum? Our next result quantifies the ensuing trade-off and shows that we can explicitly construct the commitment to approximate Stackelberg performance. The approximation gets better and better as $N$ increases.

**Theorem 4.7.4.** *Let the best-response polytope $\mathcal{R}_{j^*}$ be non-empty in $\mathbb{R}^{m-1}$. Then, provided that the number of samples $N \geq \widetilde{\mathcal{O}}(m)$, we can construct commitment $\mathbf{x}_{N,\eta}$ for every $0 < \eta < 1/2$ such that*

$$f^*_\infty - f_N(\mathbf{x}_N) = \widetilde{\mathcal{O}}\Big( \Big(\frac{m}{N}\Big)^\eta + e^{-C\cdot N^{1-2\eta}} \Big) \tag{4.6}$$

*for some constant $C > 0$. Furthermore, these constructions are computable in polynomial in $(m, n)$ time from the Stackelberg commitment $\mathbf{x}^*_\infty$. (The $\widetilde{\mathcal{O}}(\cdot)$ contains constant factors that depend on both the local and global geometry of the best-response-region $\mathcal{R}_{j^*}$. For a formal statement that includes these factors, see Lemma 4.11.10.)*

The full proof of Theorem 4.7.4, deferred to Appendix 4.11, involves some technical steps to achieve as good as possible a scaling in $N$. The caveat of Theorem 4.7.4 is that

commitment power can be robustly exploited in this way only if there are enough observations of the commitment. One obvious requirement is that the best-response-region $\mathcal{R}_{j^*}$ needs to be non-empty in $\mathbb{R}^{m-1}$. Second, the number of observations $N$ needs to be greater than the *effective dimension* of the game for the leader, $m$. This is a natural requirement to ensure that the follower has learned at least a meaningful estimate of the commitment. Third, the "constant" factors in Theorem 4.7.4 actually reflect properties about both the local and global geometry of the polytope; see the proof in Appendix 4.11 for more details. Geometric properties that intuitively lead to undesirable scaling in the constant factors of the robustness guarantee are listed below:

1. The Stackelberg commitment being a "pointy" vertex: this can lead to a commitment being far away from the boundary in certain directions, but closer in others, making it more likely for a different response to be elicited.

2. Local constraints being very different from global constraints, which implies that commitments too far in the interior of the local feasibility set will no longer satisfy all the constraints of the best-response-region.

Even with these caveats, Theorem 4.7.4 provides a general analytical framework for constructing robust commitments by making a natural connection to interior-point methods in optimization[27].

The extent to which these robust commitments approximate the ideal Stackelberg payoff can be observed through simulation on the non-zero-sum games in Examples 3 and 4 (remember that the Stackelberg commitment was non-robust for both games). Figures 4.6a and 4.7a compare the expected payoff obtained by our robust commitment constructions $\{\mathbf{x}_N\}_{N\geq 1}$ for different numbers of samples $N$, and for the games described in Examples 3 and 4 respectively. The benchmark with respect to which we measure this expected payoff is the Stackelberg payoff $f^*_\infty$ (obtained by Stackelberg commitment under *infinite observability* and *tie-break-ability in favor of the leader*). We also observe a significant gap between the payoffs obtained by these robust commitment constructions and the payoff obtained if we used the Stackelberg commitment $\mathbf{x}^*_\infty$. Section 4.7 contains more extensive empirical evaluation on random ensembles of security games.

## Approximating the maximum possible payoff

Theorem 4.7.4 shows that not only can we compute robust commitments in polynomial time, but also that these commitments have an intuitive analytical interpretation. They are designed to simultaneously stay close to the Stackelberg commitment as well as maximize the probability of preserving the expected follower response, i.e. *preserve follower learn-ability of*

[27]Noting that interior point methods are provably polynomial-time algorithms to solve LPs, it is plausible to think that in fact, stopping the interior point method appropriately early would also give us a robustness guarantee - which would imply that finding optimal *robust* commitments is even easier than finding optimal commitments!

*commitment*. Thus, the performance of commitments approximates the idealized Stackelberg payoff $f^*_\infty$ as a function of the number of observations. There is, however, no reason why the leader should always be incentivized to preserve follower learn-ability of her commitment – after all, as we saw in Example 2, not all response mismatches are sub-optimal. It is thus possible that she could realize a payoff over and above the ideal Stackelberg payoff. We now investigate how much this additional payoff can be *for any commitment choice*, by upper bounding the value of $f^*_N$ which is the leader's optimal payoff under observational uncertainty. Since the leader payoff function in Equation (4.1) is in general non-convex in $\mathbf{x}$, it is NP-hard to exactly compute; but empirical evaluations [215, 220] have noted that the leader could yield a payoff greater than traditional Stackelberg.

Rather than the complexity-theoretic approach of constructing a polynomial-time approximation algorithm, our approach is approximation-theoretic[28]. We show that in the large-sample case, we cannot do much better than the actual Stackelberg payoff $f^*_\infty$; informally speaking, our ability to deceive the follower into responding *strictly-better-than-expected* is limited. Combining this with the robust commitment construction of Theorem 4.7.4, we obtain an approximation to the optimum payoff.

The main result of this section is stated below.

**Theorem 4.7.5.** *For any $m \times n$ leader-follower game, we have*

$$f^*_N \leq f^*_\infty + Cn\sqrt{\frac{m}{N}}.$$

*for some constant $C > 0$ depending on the parameters of the game $(A, B)$.*

As a corollary the commitment construction defined in Theorem 4.7.4 provides a $\widetilde{\mathcal{O}}\left(\sqrt{\frac{1}{N}}\right)$-additive-approximation algorithm to $f^*_N$. The proof of Theorem 4.7.5 is provided in Appendix 4.11.

Theorem 4.7.5 tells us that the robust commitments are essentially optimal in that a leader *could* engineer her commitment to elicit favorable response mismatches – but any additional gain in payoff over ideal Stackelberg payoff would be minimal. The practical benefit that Theorem 4.7.5 affords us is that we now have an approximation to the optimum payoff the leader could possibly obtain, which can be computed in polynomial time after computing the Stackelberg equilibrium, which itself is polynomial time [214]. This is because the robust commitment is obtained by first computing Stackelberg equilibrium $\mathbf{x}^*_\infty$, and then deviating away from $\mathbf{x}^*_\infty$ in the magnitude and direction specified, the latter of which is a linear-time operation.

We can see the approximate optimality of our robust commitment constructions for the non-zero-sum games in Examples 3 and 4. For the case of 2 leader actions we compute

[28]In other words, the extent of approximation is measured by the *number of samples* as opposed to the run-time of an algorithm. This is very much the flavor of previously-obtained results on Stackelberg zero-sum security games [201].

the maximum possible obtainable payoff $f_N^*$ through brute force, and compare the value to the robust commitment payoff. As shown in Figures 4.6b and 4.7b, this comparison is particularly valuable for smaller values of $N$. We notice that the values are much closer even than our theory would have predicted, even for small values of $N$. For these examples, we also do not see a gain over traditional Stackelberg payoff.

## Additional simulations on random ensembles of games

| Reward \ Target | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Defender (if protected) | Unif[0,1] | Unif[0,1] | Unif[0,1] | Unif[0,1] | Unif[0,1] |
| Defender (if unprotected) | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] |
| Attacker (if protected) | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] | Unif[-1,0] |
| Attacker (if unprotected) | Unif[0,1] | Unif[0,1] | Unif[0,1] | Unif[0,1] | Unif[0,1] |

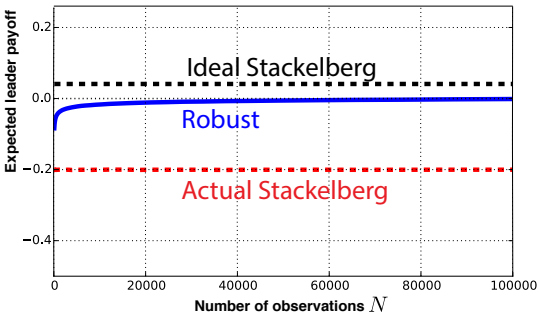
Figure 4.8: Illustration of random ensemble of $5 \times 5$ security game. Figure from [90].

We close this section with additional simulation on random ensembles of games to broadly evaluate our robust commitment constructions. We create a random ensemble of $5 \times 5$ games, inspired by the security framework, in which the defender can defend one of 5 targets, and the attacker can attack one of these 5 targets. The defender and attacker rewards are chosen to be uniformly at random between $[0, 1]$, and their penalties are uniform at random between $[-1, 0]$. This is essentially the random ensemble that was used in previous empirical work on security games [215]. Figure 4.8 shows the construction of this ensemble.

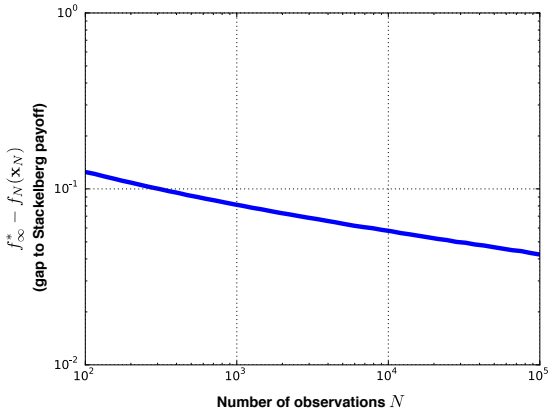
The purpose of a random ensemble is to show that the properties we observed above – unstable traditional Stackelberg commitments, robust commitment payoffs approximating the optimum – are the norm rather than the exception. Figure 4.9 illustrates the results. The average performance of the sequence of robust commitments $\{\mathbf{x}_N\}_{N \geq 1}$ on the ensemble, as well as the traditional Stackelberg commitment $\mathbf{x}_\infty^*$ is plotted in Figure 4.9a against the benchmark of ideal Stackelberg payoff $f_\infty^*$. Figure 4.9b depicts the rate of convergence of the gap in robust commitment performance to the idealized Stackelberg payoff – we can clearly see the $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ rate of convergence in this log-log plot. Finally, Figure 4.9c plots the *percentage gap* between robust commitment payoff and idealized Stackelberg payoff as a function of $N$.

We can make the following conclusions from these plots:

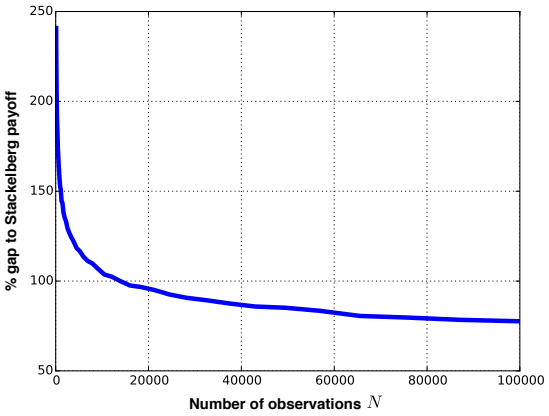
1. The Stackelberg commitment is extremely non-robust *on average*. In fact, we noticed that this was the case with high probability.




(a) Plot of expected defender payoff when defender uses robust commitments – compared to Stackelberg commitment as well as idealized Stackelberg payoff.




(b) Log-log plot of the gap between robust commitment payoff and idealized Stackelberg payoff.




(c) Percentage plot of the gap between robust commitment payoff and idealized Stackelberg payoff.

Figure 4.9: Performance of robust commitments and traditional Stackelberg commitments in random $5 \times 5$ Stackelberg security games for a finite number of observations of defender commitment. Figures from [90].

2. The robust commitments are doing much better *on average* than the original Stackelberg commitment even for very large values of $N$. The stark difference in payoff between the two motivates the construction of the robust commitment, which is essentially as easy to compute as the Stackelberg commitment.

## 4.8 Model for repeated interaction

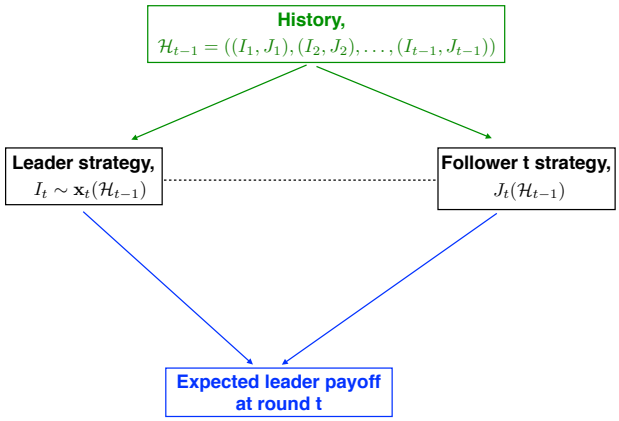



Figure 4.10: Illustration of the stage game at round $t$ between leader and follower $t$, both of whom observe history of play $\mathcal{H}_{t-1}$. The dotted line indicates that leader and follower $t$ play *simultaneously*. Figure from [90].

The limited observation model with a shared belief in leader commitment is realistic for engineering applications of the Stackelberg solution concept, like Stackelberg security games (and indeed, empirical solutions to the limited-observability objective in Equation (4.3) have been proposed and evaluated [215, 220]). However, it is unrealistic for modeling the manifestation of commitment in paradigms like Bayesian persuasion where the reputation of the persuader is key. Consider Kamenica and Genztkow's introductory example of the power of Bayesian persuasion, in which a prosecutor persuades a judge to convict defendants even when they may be innocent. The prosecutor does this through a particular *randomized* signalling scheme conditioned on the outcome of the prosecutor's private investigation (for e.g. she knows the true identity of the defendant). This signalling scheme, and especially its randomization, can only be *credibly* revealed by creating a history of persuasion. Creating this history involves signalling publicly to *other* judges in *other* cases.

Informally, we are interested in approximately optimal signalling schemes for leader over a finite number of rounds. There are two important distinctions from the one-shot model:

1. The leader is repeatedly interacting: she is still revealing $N$ signals, but sequentially. Every time she reveals a signal, she plays against a new follower (identical to the previous followers in his payoff function) and realizes a payoff herself.

2. The leader is not obliged to commit to a *fixed* signalling scheme, e.g. iid realizations of an a-priori fixed mixed strategy. She can update her signalling strategy *sequentially* based on the current history of follower responses. Thus, follower(s) *will not* play the game with established belief in a leader commitment.

## Leader model

Like the seminal Bayesian models for asymptotic reputation building, we consider a repeated game played over multiple rounds between the leader and multiple followers. In round $t$, the leader faces follower $t$, and executes move $I_t \in [m]$ (which may be randomized). In response, the follower executes move $J_t \in [n]$ (which may also be randomized). We define history

$$\mathcal{H}_{t-1} := \{(I_s, J_s)\}_{s=1}^{t-1} \tag{4.7}$$

and require the leader and follower moves to be functions of this history, parameterized by their respective payoff matrices. In other words, we consider functions (that we allow to be randomized)

$$I_t := I_t(\mathcal{H}_{t-1}; A)$$
$$J_t := J_t(\mathcal{H}_{t-1}; B)$$

for the leader and ($t^{th}$) follower moves at round $t$, and define leader and follower *rules* by $\{I_t(\cdot)\}_{t=1}^T$ and $\{J_t(\cdot)\}_{t=1}^T$ respectively. Critically, the followers are *myopic*, and do not know the value of $T$, i.e. how long the leader will be playing the game for[29]. Thus, our leader and follower are using *anytime* rules and we can analyze $T \to \infty$ as well as what happens at finite $T$. We define payoffs for the repeated game as equal to the *time-averaged reward*, as in stochastic dynamic programming [232]. Here, stochasticity comes from the random realizations of the leader and follower rules.

**Definition 4.8.1.** *For a fixed follower rule $J_t(\mathcal{H}_{t-1}; B)$, we define the time-averaged leader payoff function by:*

$$f_T(\{I_t\}_{t=1}^T) := \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} A_{I_t(\mathcal{H}_{t-1};A),J_t(\mathcal{H}_{t-1};B)}\right] \tag{4.8}$$

*where the expectation is taken over any randomization used in the leader and follower rules $I_t(\cdot), J_t(\cdot)$.*

We have not defined a common prior for the repeated game, which means that we cannot define a SPNE[30]. Nevertheless, it is both sensible and interesting to study the evolution of

[29] If followers knew this, any repeated analysis would unravel.

[30] At a high level, this choice enables us to construct explicit leader and follower mechanisms that can be thought of as an approximate equilibrium, or at least approximately achieving the equilibrium payoff. We provide a detailed comparison between Bayesian and frequentist models in Section 4.3.

the leader objective in Equation (4.8) against follower rules that utilize principled *frequentist* inference principles in their decision-making. We now describe the building blocks of these principled follower rules.

## Follower(s) response model(s)

Note that follower $t$ is participating in the game only at round $t$, and it is thus reasonable to think of him as *myopic*. Since Follower $t$ is also rational, he will play to maximize *what he believes to be his expected payoff* in round $t$. To formalize this, we appeal to the notion of a forecast.

**Definition 4.8.2.** *We define an optimal response model for Follower $t$* ***conditioned on a forecast*** *as one that uses, as a sufficient statistic, a* ***forecast function***

$$\mathbf{r}_t : ([m] \times [n])^{t-1} \to \Delta_m \tag{4.9}$$

*for how Follower $t$ believes the leader is going to play at round $t$. In particular, for this forecast function and history $\mathcal{H}_{t-1}$, the follower responds with* ***pure strategy***

$$J_t(\mathbf{r}_t) := j^*(\mathbf{r}_t(\mathcal{H}_{t-1})), \tag{4.10}$$

*where we continue to assume the follower to break ties in favor of the leader.*

In this way, we can define any response model for the *aggregate of followers* 1 to $\infty$ through the sequence of forecast functions $\{\mathbf{r}_t(\cdot)\}_{t=1}^{\infty}$. In the absence of a common prior, it remains to be seen what a sensible forecasting rule for the aggregate of followers is, and in fact the choice of forecasting rule critically determines how well the leader can do (or conversely, how poorly she can do) for her own choice of rule. We specify some explicit choices, and briefly describe them, below.

**Definition 4.8.3** (Empirical averages forecast)**.** *Follower $t$ uses an empirical averages forecast if she uses forecasting rule*

$$\mathbf{r}_{t,\mathsf{avg}}(\mathcal{H}_{t-1}) := \widehat{\mathbf{X}}_{t-1}. \tag{4.11}$$

Through the Bayesian interpretation as discussed in Section 4.5, using the empirical averages forecast would be optimal by expected utility theory *if the leader had committed to an iid rule $I_t$ i.i.d. $\sim \mathbf{x}$*. There is, of course, no reason why the leader should do this a-priori *in the repeated setting*, and in fact the leader might be incentivized to deviate considerably from such a rule if she observes followers naively responding in this way[31]. This would make

[31]Interestingly, empirical averages forecasts together with myopic best responses have been widely used in experimentation for learning in simultaneous games played repeatedly [78], even though players may have incentive to deviate from such protocols. As Fudenberg and Kreps say in their work on experimentation for learning in simultaneous repeated games [233], *"We do not imagine that players would maintain their belief in an asymptotic environment of stationary and independently chosen behavior strategies if the evidence that players accumulate manifestedly discomfirms this hypothesis."*

the empirical averages forecasting rule *by itself* an unsatisfactory choice for the aggregate of followers. Later, we will affirm this concretely – but in the meantime, we suggest two natural alternatives.

**Definition 4.8.4** (Predictive modelling forecast)**.** *Consider an unknown parameter $\theta^* \in \Omega$, where $\Omega$ can be a discrete or continuous space. Consider a* ***predictive model for the sequence***

$$\mathbf{Y}_t := P_t((Y_1, \ldots, Y_{t-1}), \mathbf{W}_t; \theta^*) \tag{4.12}$$

*where $\mathbf{W}_t$ is stochastic noise independent of the sequence $(Y_1, \ldots, Y_{t-1})$. We denote the collection of (vector-valued) prediction functions $\mathscr{P} := \{P_t(\cdot; \theta^*)\}_{t=1}^T$. Then, we define the* ***predictive modelling forecast*** *for follower $t$ corresponding to model $\mathscr{P}$:*

$$\mathbf{r}_{t,\mathcal{P}}(\mathcal{H}_{t-1}) := \widehat{P}_t((I_1, \ldots, I_{t-1}), \mathbf{0}), \tag{4.13}$$

*where $\widehat{P}_t(\mathcal{H}_{t-1})$ constitutes the maximum likelihood estimate of the predictive model from history $\mathcal{H}_{t-1}$ (and could involve the plug-in estimate of $\theta^*$ or something more sophisticated).*

Equation (4.12) appears very complicated – but already encapsulates the empirical averages forecast as a specific example: the iid case. (Here, we would have $\theta^* = \mathbf{x} \in \Delta_m = \Omega$ and predictive model $P_t((I_1, \ldots, I_{t-1}), \mathbf{W}_t; \mathbf{x}) = \mathbf{x} + \mathbf{W}_t$.) What Definition 4.8.4 additionally affords us is a significantly broader framework for temporally predictable sequences, parameterized by prediction functions $\mathscr{P}$ and parameter space $\Omega$, *beyond the iid case*: some simple examples are Markov processes, periodic sequences, and even sequences generated by pseudo-random number generators. It is even more starkly clear here that the followers using such a forecast only makes sense if a) the leader is indeed generating her sequence from such a predictive model, b) the a-priori unknown parameters of the model are learn-able from much fewer than $t$ samples. Such predictive forecasts can be extremely poor in the absence of these assumptions. Nevertheless, as we will see, considering such forecasts will become extremely valuable if the leader indeed chooses such a predictive model, *especially if the generated sequence is deterministic or close to deterministic in its realization.*

Both the above forecasting rules assumed predictability of the leader rule. We consider, as our final forecasting rule, a pessimistic point of view on leader predictability round to round. We do this through the concept of calibration. Before defining a universally calibrated forecast, we define some additional notation. For an arbitrary forecasting rule $\mathbf{r}_t(\cdot)$, executed leader rule $(I_1, \ldots, I_T)$ and fixed forecast $\mathbf{r}$, define

$$N_T(\mathbf{r}) := \sum_{t=1}^{T} \mathbb{I}\left[\mathbf{r}_t(\mathcal{H}_{t-1}) = \mathbf{r}\right]$$

as the number of times forecast $\mathbf{r}$ was used (this can itself be a random quantity as it depends on the history and potential randomization in the forecast). Also define

$$\widehat{\mathbf{X}}_T(\mathbf{r}) := \frac{\sum_{t=1}^{T} \mathbb{I}\left[\mathbf{r}_t(\mathcal{H}_{t-1}) = \mathbf{r}\right] \widehat{e}_{I_t}}{N_T(\mathbf{r})}$$

as the empirical mean of the leader generated sequence whenever forecast $\mathbf{r}$ was used. Here $\widehat{e}_{I_t}$ is the standard basis vector corresponding to action $I_t$. Now we can define a universally calibrated forecast.

**Definition 4.8.5** (Universally calibrated forecast [208])**.** *The sequence of followers follows a universally calibrated forecast, which we denote by*

$$\mathbf{r}_{t,\mathsf{univ}}(\mathcal{H}_{t-1}) \tag{4.14}$$

*if for any leader rule $\{I_t(\cdot)\}_{t=1}^T$ (potentially randomized) and sufficiently large $T$, we have*

$$\frac{1}{T}\sum_{\mathbf{r}\in\Delta_m} N_T(\mathbf{r})\|\widehat{\mathbf{X}}_T(\mathbf{r}) - \mathbf{r}\|_1 = \frac{o(T)}{T} \text{ almost surely,} \tag{4.15}$$

*where almost surely is over the randomness in a) execution of the leader sequence, b) execution of the follower forecasting rule.*

The now-classical link between calibrated learning rules and convergence to correlated equilibrium has been established in repeated *simultaneous* games with two long-term players who do not know each other's utility functions [208]. The idea of calibration is intricately connected to the notion of "no-worst-case-regret" in repeated, zero-sum games and sequential prediction in the worst case. The way we will consider calibration is slightly different here: we are only using a universally calibrated rule for the followers, and not for the leader. This is because under our model of asymmetric private information, the leader actually knows the follower utility function and, more importantly, *is not trying to forecast follower responses.* That is, the leader cannot, in general, infer anything about the actions of new followers from the observation of previous followers – as we have assumed that they are acting independently. Under this assumption of independent action, one can conversely ask whether it is plausible for the aggregate of followers to implement a universally calibrated forecasting rule. While acknowledging the implausibility, we believe this is still an interesting and important case to study for the following reasons:

1. A certificate of universal calibration ensures that the aggregate of followers is responding optimally in the asymptotic sense *regardless of the leader's chosen rule.*

2. We may not actually expect the aggregate of followers to follow a universally calibrated sequence of forecasts; but this represents the worst case for a leader who is attempting to systemically mislead followers for additional personal gain.

## 4.9 Results for repeated interaction

A qualitative summary of results that we obtained in the one-shot model of Section 4.5 is as follows:

1. *Mixed* Stackelberg commitments are generically unstable to even an infinitesimal amount of observational uncertainty (Proposition 4.7.1).

2. *Robust, mixed* commitments can be explicitly constructed by trading off the power of mixture in commitment with preservation of follower learn-ability (Theorem 4.7.4).

3. The leader has minimal incentive (Theorem 4.7.5), and can have *dis-incentive* (Proposition 4.7.1), to deceive the follower into responding sub-optimally in a feasible way.

While the repeated interaction model of Section 4.8 is conceptually more complex, we will recover similar guiding principles to the above.

## Two illustrative examples for leader behavior

In the repeated interaction model, we want to understand how leaders should optimally behave and build their reputation (the optimization problem in Definition 4.8.1) when playing against followers who are *responding optimally as well* (i.e. playing one of the rules from Definition 4.8.3, 4.8.4, and 4.8.5 depending on which would maximize their aggregate payoff). Before stating our main results formally, let us view the main ideas through two simple examples. As in Section 4.7, we use Figure 4.11 to depict the stage games in normal form as well as ideal leader payoff function. The first example is a reproduction of Example 3 as a repeated security game.

### Repeated security game

**Example 4:** In this example, the leader is a defender facing a sequence of attackers (followers). The pure strategies for defender and attackers are indexed by targets $\{1, 2\}$ – the defender can choose to defend target 1 or 2, and the attacker can choose to attack target 1 or 2. Payoff matrices[32] $\{\{A_{ij}\}, \{B_{ij}\}\}$ for both players as well as the ideal defender payoff function are represented in Figure 4.11b.

We denote the realized strategies at round $t$ for defender and attacker respectively by $I_t \in \{1, 2\}$ and $J_t \in \{1, 2\}$. At round $t$, defender and attacker observe common history $\mathcal{H}_{t-1} = \{(I_s, J_s)_{s=1}^{t-1}\}$, and can play according to rules $I_t := i_t(\mathcal{H}_{t-1})$ and $J_t := j_t(\mathcal{H}_{t-1})$ respectively.

Recall the definition of the ideal defender payoff function as depicted in Figure 4.11b, and that the optimal mixed commitment for the defender, expressed in terms of probability of protecting target 1, is $p^*_\infty = 0.5$. Her ideal Stackelberg payoff is $f^*_\infty = 0.5$. We have seen how she can achieve this payoff through robust commitments when there is shared belief in commitment but limited observability. We will now use this example to understand how she

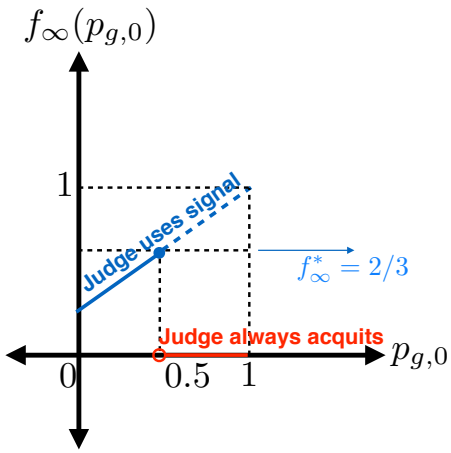
[32]The intuition for the payoffs is: the defender gets a unit payoff for neither target being compromised, but more negative payoff from target 2 being compromised than target 1. On the other hand, the attacker gets unit payoff from compromising either of the targets.

$\Pi = 0$
**(innocent,** probability 2/3**)**

| Prosecutor \ Judge | Convict | Acquit |
|---|---|---|
| Signal 'g' | (1,0) | (0,1) |
| Signal 'i' | (1,0) | (0,1) |

$\Pi = 1$
**(guilty,** probability 1/3**)**

| Prosecutor \ Judge | Convict | Acquit |
|---|---|---|
| Signal 'g' | (1,1) | (0,0) |
| Signal 'i' | (1,1) | (0,0) |

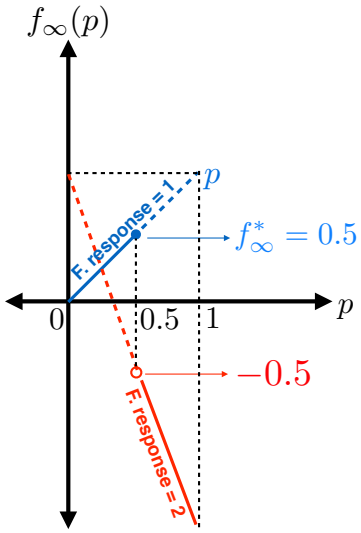



**Probability that prosecutor sends 'g' for an innocent defendant =** $p_{g,0}$

**Overall fraction of convictions =** $f_\infty(p_{g,0})$

(a) Persuasion game.

| Defender \ Attacker | Target 1 | Target 2 |
|---|---|---|
| Target 1 | (1,0) | (-2,1) |
| Target 2 | (0,1) | (1,0) |




(b) Security game.

Figure 4.11: Examples of non-zero-sum repeated security and persuasion games. In the security game, $p$ denotes the probability with which the defender will defend target 1. In the persuasion game, $p_{g,0}$ denotes the probability with which the prosecutor will signal guilty

could achieve this payoff by additionally *establishing belief through repeated interaction with attackers*.

### Persuasion through repeated interaction

Our next example is a persuasion game where the signalling mechanism is *not* a-priori known or believed, with the additional complication of there being private information that is known to the leader but unknown to the follower.

**Example 5 (Example from [188], one-shot):** We reproduce the example from Kamenica and Gentzkow's seminal work on Bayesian persuasion in which a prosecutor is trying to convince a judge that a defendant is guilty. We denote the true culpability of the defendant by the $\{0,1\}$-valued random variable, $\Pi = \mathbb{I}[\text{defendant is guilty}] \sim \text{Ber}(\pi)$ for some $\pi \in (0, 0.5)$. The game proceeds in a sequence of steps:

1. The prosecutor *commits* to a signalling mechanism, uniquely determined by

$$p_{g,1} := \Pr[Y = \text{g} | \Pi = 1]$$
$$p_{g,0} := \Pr[Y = \text{g} | \Pi = 0].$$

   This mechanism is exactly revealed to the judge.

2. The true state of the defendant, $\Pi$, is exactly revealed to the prosecutor.

3. The prosecutor draws a signal $Y \sim p_{g,\Pi}$ and reveals it to the judge.

4. The judge decides to make a conviction or acquit based on expected utility theory under her posterior belief about the identity of the defendant.

Viewed as a one-shot, *Bayesian* Stackelberg game (with the identity of the defendant being asymmetric private information), the leader is the prosecutor and the follower is the judge. The leader and follower payoff matrices are depicted in normal form in Figure 4.11a, where the private information $\Pi$ is only visible to the leader (prosecutor). The leader (prosecutor) has 4 pure strategies to choose from, corresponding to signaling either 'g' or 'i' conditioned on the private information $\Pi$. Any mixed strategy over these 4 pure strategies can be expressed as a randomized signalling mechanism $\{p_{g,1}, p_{g,0}\}$. The follower (judge) has 2 pure strategies to choose from, convict ($Z = \text{c}$) or acquit ($Z = \text{a}$).

We constrain the prosecutor strategy space to $p_{g,1} = 1$, i.e. the prosecutor will truthfully signal guilty ('g') if the defendant is truly guilty[33] and consider the choice of $p_{g,0} \in [0, 1]$. Thus, if the judge sees signal 'i', he knows with certainty that the defendant was innocent

[33] A routine calculation, which we omit, shows that this is optimal for the prosecutor – the intuition is that the prosecutor has no incentive to lie about the identity of a truly guilty defendant, because she always wants to maximize her number of convictions.

and will respond with *pure strategy* 'a', yielding payoff 0 for the prosecutor. On the other hand, if the judge sees signal 'g', he will infer that the defendant was guilty with probability

$$\Pr[\Pi = 1|Y = \text{g}] = \frac{\pi}{\pi + (1-\pi)p_{g,0}}, \tag{4.16}$$

and based on the expressions above for his payoff function, he will respond with *pure strategy* 'c' *if and only if* we have $\Pr[\Pi = 1|Y = \text{g}] \geq \frac{1}{2}$. Thus, the leader (prosecutor) payoff function is given by:

$$\begin{aligned} f_\infty(p_{g,0}) &= \Pr[Y = \text{g}]\mathbb{I}\left[\Pr[\Pi = 1|Y = \text{g}] \geq \frac{1}{2}\right] \\ &= (\pi + (1-\pi)p_{g,0}) \cdot \mathbb{I}\left[\frac{\pi}{\pi + (1-\pi)p_{g,0}} \geq \frac{1}{2}\right]. \end{aligned} \tag{4.17}$$

A simple calculation yields that the optimal mechanism, i.e. the value of $p_{g,0}$ that maximizes $f_\infty(\cdot)$, is given by

$$p^*_{g,0} = \frac{\pi}{1-\pi}$$

and the optimal payoff is given by $f^*_\infty = 2\pi$.

Having described the one-shot setting, we describe a natural repeated-game formulation that could explain how the prosecutor is able to reveal her mixed signalling mechanism.

**Example 5 (repeated version)** Consider a prosecutor interacting with $T$ judges over $T$ rounds. In each round, the prosecutor works with *a different judge on the case of a different defendant*. That is, at round $t$, prosecutor receives private information $\Pi_t$ i.i.d $\sim \text{Ber}(\pi)$ about the culpability of defendant $t$. Then, she reveals signal $I_t \in \{\text{'g','i'}\}$ to the judge. The judge has access to *history* constituting the prosecutor's previous signals, judges' prior decisions $\{J_s \in \{\text{'c', 'a'}\}\}_{s=1}^{t-1}$ and the corresponding true culpability of previous defendants, i.e. we have

$$\mathcal{H}_{t-1} = \{(I_s, J_s, \Pi_s)_{s=1}^{t-1}\}. \tag{4.18}$$

Based on these, he makes a decision $J_t := j_t(\mathcal{H}_{t-1}; \pi)$ according to a learning rule[34] that only has access to history $\mathcal{H}_{t-1}$, knows the common prior on defendants $\pi$ *but not the actual culpability of the current defendant.* We are interested in identifying prosecutor rules $I_t := i_t(\mathcal{H}_{t-1}; \pi, \Pi_t)$ that would maximize her expected payoff for sensible judge response rules, based on forecasts as defined in Definition 4.8.1.

The persuasion model introduces additional bells and whistles in the information structure, primarily due to the presence of external private information. Strictly speaking, it is a

[34]"Incomplete information" for the judge represents both the unknown identity of the defendant, and the utility function of the prosecutor.

slightly more complex framework than the 2-player leader followers repeated game described in Section 4.8; and there are several possible variants one could consider to the information structure[35]. Nevertheless, persuasion models provide one of the most compelling contemporary examples of a non-trivially established reputation, and so we chose a simple persuasion model as a motivating example.

We will consider the special case where $\pi = 1/3$ and $p^*_{g,0} = 1/2$ for ease of exposition. In this case, the optimal payoff is given by $f^*_\infty = 2/3$, which means that the prosecutor can successfully get $2/3$ of the defendants convicted *even though only* $1/3$ *of them are truly guilty.* In reality, there is no reason for judges to believe that a prosecutor would commit to a fixed signalling mechanism, nor would they know the mechanism exactly. We wish to understand how, then, a prosecutor can realize this persuasive power through repeated interaction with judges.

### Robust, randomized leader schedules

We start by describing *a* mechanism by which the leader can establish belief in mixed commitment while eventually approaching her mixed Stackelberg payoff. We first describe this mechanism for the security game example.

**Example 4:** The mechanism involves a defender rule and an attacker response, both of which we describe below.

**Defender rule:** For $\eta < 1/2$, we fix

$$I_t \sim \begin{cases} 1 \text{ w.p. } \max\left\{\frac{1}{2} - \frac{1}{t^\eta}, 0\right\} \\ 2 \text{ w.p. } \min\left\{\frac{1}{2} + \frac{1}{t^\eta}, 1\right\} \end{cases} . \tag{4.19}$$

We denote as shorthand $p_t := \max\left\{\frac{1}{2} - \frac{1}{t^\eta}, 0\right\}$, the probability that defender defends target 1 in round $t$; and $P_t \sim \text{Ber}(p_t)$ as the indicator that she actually defended target 1 in round $t$. *Observe that we are simply using the robust commitment construction corresponding to $t$ unknown samples at round $t$, constructed according to Theorem 4.7.4.* This constitutes a *randomized* rule with independent, but not identically distributed deployments across rounds.

**Attacker rule:** We consider attackers who respond using the *empirical averages forecast* as in Definition 4.8.3; that is, attacker $t$ uses empirical estimate $\widehat{P}_{t-1} := \frac{1}{t-1}\sum_{s=1}^{t-1} P_t$ for his forecast in round $t$, and responds with:

$$J_t = \begin{cases} 1 \text{ if } \widehat{P}_{t-1} \leq 1/2 \\ 2 \text{ otherwise.} \end{cases} \tag{4.20}$$

[35]In particular, we do not mean to suggest that the setting in Example 5 is the *only* way persuasion could be realized. For example, the assumption that past realizations of private information are revealed in history seems particularly strong for practical use. The nature of theoretical results will be heavily dependent on the information structure used in modeling.

We will analyze the time-averaged payoff that the defender should expect from using the rule in Equation (4.19) against attackers who respond according to Equation (4.20). Recall that we use $\mathbb{E}[\cdot]$ to denote the expectation over the randomization in defender rule (and thus randomization in attacker responses). For a general randomized defender rule $(p_1, \ldots, p_T)$, the expected time-averaged payoff against such attackers is denoted by

$$\begin{aligned}
&f_{T,\mathsf{avg}}((p_1, \ldots, p_T)) \\
&:= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t A_{1,J_t} + (1-P_t)A_{2,J_t}\right] \\
&= \mathbb{E}\Big[\frac{1}{T}\sum_{t=1}^{T} P_t \mathbb{I}\left[\widehat{P}_{t-1} \le 1/2\right] A_{1,1} \\
&+ P_t \mathbb{I}\left[\widehat{P}_{t-1} > 1/2\right] A_{1,2} + (1-P_t)\mathbb{I}\left[\widehat{P}_{t-1} \le 1/2\right] A_{2,1} \\
&+ \Pr\left[\widehat{P}_{t-1} > 1/2\right] A_{2,2}\Big],
\end{aligned}$$

where we have simply substituted attacker $t$'s response $J_t$ according to the learning rule in Equation (4.20). The outcomes marked in red ($(I_t = 1, J_t = 2)$ and $(I_t = 2, J_t = 1)$) are undesired by the defender. Outcome $(I_t = 1, J_t = 2)$ is particularly poor, yielding a payoff of $-2$ for the defender.

Denote $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\mathcal{H}_{t-1}]$ as shorthand for the conditional expectation of a quantity given history $\mathcal{H}_{t-1}$. Substituting $A_{1,1} = 1, A_{2,2} = 1, A_{1,2} = -2$ and $A_{2,1} = 0$, and using the tower property of conditional expectation, we evaluate the above expression for the randomized defender rule to get

$$\begin{aligned}
&f_{T,\mathsf{avg}}((p_1, \ldots, p_T)) \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} -2P_t \cdot \mathbb{I}[\widehat{P}_{t-1} > 1/2] + (1-P_t)\cdot \mathbb{I}[\widehat{P}_{t-1} > 1/2] + P_t \cdot \mathbb{I}[\widehat{P}_{t-1} \le 1/2]\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (1-4P_t)\cdot \mathbb{I}[\widehat{P}_{t-1} > 1/2] + P_t\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_t\left[(1-4P_t)\cdot \mathbb{I}[\widehat{P}_{t-1} > 1/2] + P_t\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[(1-4p_t)\mathbb{I}[\widehat{P}_{t-1} > 1/2] + p_t\right] \text{ (using independence across rounds)} \\
&\ge \underbrace{\frac{1}{T}\sum_{t=1}^{T} p_t}_{A} - \underbrace{\frac{1}{T}\sum_{t=1}^{T} 3\cdot \Pr\left[\widehat{P}_{t-1} > 1/2\right]}_{B} \text{ (using linearity of expectation, and } 1-4p_t \ge -3).
\end{aligned}$$

Here, term $A$ represents the remaining time-averaged power of mixed commitment (in particular, we want to know how much less this term is than ideal Stackelberg), and term $B$ represents the probability that the undesired response $J_t = 2$ is elicited, which happens whenever $\widehat{P}_{t-1} > 1/2$. It turns out that we can show[36] that

$$A \geq \frac{1}{2} - \frac{2}{T^\eta}$$
$$B \leq \frac{C}{T} \text{ for some } C > 0,$$

and thus $\mathbb{E}\left[f_{T,\mathsf{avg}}((P_1, \ldots, P_T))\right] \geq \frac{1}{2} - \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ (taking $\eta \sim 1/2$), showing approximate optimality of this leader rule and eventual convergence to the defender's ideal Stackelberg payoff $1/2$. Recall that the Stackelberg commitment is $p^*_\infty = 1/2$, and the randomized defender rule is in fact converging to this as more rounds of the game are played.

Figure 4.12 provides intuition for how the randomized rule works. In initial rounds, the defender plays very conservatively and is much more likely to defend target 2 than target 1 to ensure a very low probability of undesired attacker response (i.e. attack of target 2). Later, the defender can afford to mix up her defense more and move closer to the boundary of eliciting different attacker responses, which is $p = 1/2$. Figure 4.12a shows that the *effective commitment power* realized at round $t$, denoted by $\overline{p}_t := \frac{1}{t}\sum_{s=1}^t p_s$, is slightly less than the effective commitment power in the one-shot model with $t$ limited observations, which would be simply $p_t$. This represents a small additional price of establishment of belief in commitment, and manifests in a slight difference in the overall payoffs under the one-shot and repeated models as seen in Figure 4.12b. Nevertheless, the difference is small, on the order of $\frac{1}{\sqrt{t}}$ for a round $t$, and is exactly characterized in Proposition 4.9.2. Moreover, observe that under this defender rule, the attackers are not only asymptotically best responding according to Equation (4.20) (this is because we can show that $\widehat{P}_t \to^{\text{a.s.}} p^*_\infty = 1/2$), but also they are doing so at the best possible rate, in accordance with information-theoretic lower bounds on estimation [195]. This lends additional robustness to this defender rule as a constructive way to achieve mixed-strategy Stackelberg equilibrium through repeated interaction.

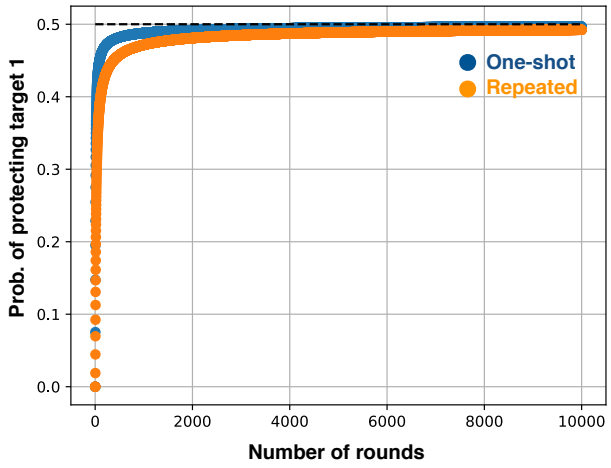
This intuition also guides us to a randomized prosecutor rule to achieve persuasion power.

**Example 5. Prosecutor rule:** At round $t$, let $N_{t-1} := \sum_{s=1}^{t-1} \mathbb{I}[\Pi_s = 0]$ represent the number of innocent defendants seen so far. Furthermore, we denote $s_1, s_2, \ldots, s_j, \ldots$ to be the epochs of arrivals of innocent defendant numbers $1, 2, \ldots, j, \ldots$ (For convention, we denote $s_0 = 0$ and $s_{N_T} = T$.)

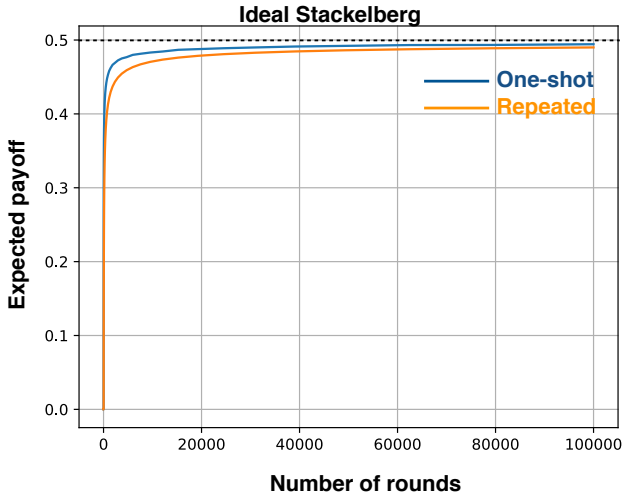
Then, we will consider the prosecutor to use rule

$$p_{g,0}(t) = \max\left\{\frac{1}{2} - \frac{1}{N_{t-1}^\eta}, 0\right\} \tag{4.21}$$
$$p_{g,1}(t) = 1, \tag{4.22}$$

[36] This is a special case of Proposition 4.9.2, which is proved in Section 4.12.




(a) Evolution of randomized leader rule, i.e. effective commitment power $\frac{1}{t}\sum_{s=1}^{t} p_s$ as a function of $t$.




(b) Gap between expected payoff of randomized leader rule and ideal Stackelberg payoff.

Figure 4.12: Performance of the robust randomized leader rule in a repeated security game between defender and attackers. Figures from [90].

and denote as shorthand $\mathbb{I}[I_t = \text{'g'}] = P_t \sim \text{Ber}(p_{g,\Pi_t}(t))$, i.e. the unconditional probability that the prosecutor signals 'g' on any round.

**Judge rule:** We consider judges that use an empirical averages forecast of the probability with which the prosecutor signals 'g' for an innocent defendant. We denote this forecast at round $(t-1)$ by $\widehat{P}_{g,0}(t-1) := \frac{\sum_{s=1}^{t-1} \mathbb{I}[\Pi_s = 0] P_t}{N_{t-1}}$. Based on this forecast, we define the judge response model to be

$$J_t = \begin{cases} \text{'c' if } \frac{\pi}{\pi + (1-\pi)\widehat{P}_{g,0}(t-1)} \geq \frac{1}{2} \\ \text{'a' otherwise.} \end{cases} \tag{4.23}$$

For any sequence of defendants $\{\Pi_1, \Pi_2, \ldots, \Pi_T\}$, we define the expected prosecutor payoff, averaged over time, against a sequence of judges that respond according to Equation (4.23):

$$\begin{aligned} f_{T,\mathsf{avg}}((p_1, \ldots, p_T)) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t \cdot \mathbb{I}[J_t = \text{'c'}]\right] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t \cdot \mathbb{I}\left[\frac{\pi}{\pi + (1-\pi)\widehat{P}_{g,0}(t-1)} \geq \frac{1}{2}\right]\right], \end{aligned}$$

where $\mathbb{E}[\cdot]$ denotes expectation *only over* the realizations of the prosecutor rule. Recalling that we specified $\pi = 1/3$, we have (in an argument similar to the preceding example),

$$\begin{aligned} &\mathbb{E}\left[f_{T,\mathsf{avg}}((P_1, \ldots, P_T))\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t \cdot \mathbb{I}\left[\frac{1}{1 + 2\widehat{P}_{g,0}(t-1)} \geq \frac{1}{2}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t \cdot \mathbb{I}\left[\widehat{P}_{g,0}(t-1) \leq \frac{1}{2}\right]\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_{t-1}\left[P_t - P_t \cdot \mathbb{I}\left[\widehat{P}_{g,0}(t-1) > \frac{1}{2}\right]\right]\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0] p_{g,0}(t) - p_{g,0}(t) \cdot \mathbb{I}\left[\widehat{P}_{g,0}(t-1) > \frac{1}{2}\right] \cdot\right] \\ &\geq \underbrace{\frac{1}{T}\sum_{t=1}^{T} (\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0] p_{g,0}(t))}_{A} - \underbrace{\frac{1}{T}\sum_{t=1}^{T} \Pr\left[\widehat{P}_{g,0}(t-1) > \frac{1}{2}\right]}_{B} \end{aligned}$$

Here, $A$ represents the time-averaged power of mixed-strategy persuasion, and $B$ represents the cumulative negative effect that can arise from a mis-perceived persuasion, which would result in judges ignoring the signal and acquitting. Because these terms also depend

on the realizations of the external private information $(\Pi_1, \ldots, \Pi_T)$, bounding them is a slightly more intricate procedure – but the same intuition holds, and we can still do it. We refer an interested reader to the full calculation in Appendix 4.13, and simply present the final result here: taking a further expectation over the realizations of the private information $(\Pi_1, \ldots, \Pi_T)$, one gets

$$\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left[\mathbb{E}\left[f_{T,\mathsf{avg}}((P_1,\ldots,P_T))\right]\right] \geq \frac{2}{3} - \frac{2}{(0.5T)^\eta} - e^{-\frac{0.01T^2}{8}} - \frac{C}{T}$$

for some constant $C > 0$, and thus the prosecutor approaches her ideal Stackelberg payoff, equal to 2/3, at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ (taking $\eta \sim 1/2$).

The natural indexing for time in build-up of persuasion power actually involves the number of innocent defendants seen so far (which is close to $\frac{2}{3}T$ when $T$ is large). The prosecutor starts off very conservative and is much more likely to recommend an acquittal for innocent defendants to ensure credibility. After many more innocent defendants have been encountered, she can afford to mix up her signals more and more, and move closer to the boundary of maximal persuasion power.

These examples outline *one* way in which leaders can build up reputation credibly. Is it essentially the *only way*?

### Deception and accelerated establishment of commitment against naive followers

To answer the above question, it is instructive to consider the optimal leader rule relative to *naive* followers that always use the empirical averages forecast from Definition 4.8.3 in our two examples. While we do not expect *intelligent* followers to blindly behave in this way, the results are surprisingly insightful.

**Example 4.** Recall, from Section 4.9, that the defender objective function against attackers who always respond according to Equation (4.20) (i.e. *naive* attackers who use the empirical averages forecast), is given by

$$f_{T,\mathsf{avg}}(p_1,\ldots,p_T) := \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(1-4P_t)\cdot\mathbb{I}[\widehat{P}_{t-1} > 1/2] + P_t\right].$$

For convenience, we slightly rewrite this expression to get

$$f_{T,\mathsf{avg}}(p_1,\ldots,p_T) := \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(1-3P_t)\cdot\mathbb{I}[\widehat{P}_{t-1} > 1/2] + P_t\cdot\mathbb{I}[\widehat{P}_{t-1} \leq 1/2]\right]. \tag{4.24}$$

We already saw that the randomized defender rule from Equation (4.19) approximately achieves the defender's ideal Stackelberg payoff, equal to 1/2. However, if faced with such naive attackers, the defender can do much better. It is easy to verify that $f_{T,\mathsf{avg}}(P_1,\ldots,P_t) \leq$

1 for any *deterministic* leader rule $(P_1, \ldots, P_T) \in \{0,1\}^T$ (because the maximal expected payoff the defender can get at any round is 1, when all targets are successfully defended). This payoff can be *achieved* if the defender can incentivize the naive attacker to attack *precisely the target that she is planning to defend* on every round. Consider the following deterministic defender rule, which does precisely this:

$$P_t = \begin{cases} 1 \text{ if } t \text{ is odd} \\ 0 \text{ if } t \text{ is even} \end{cases}. \tag{4.25}$$

| Round | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Defender strategy | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |
| $\widehat{P}_t$ | 1 | 0.5 | 0.667 | 0.5 | 0.6 | 0.5 | 0.571 | 0.5 | 0.555 | 0.5 |
| Attacker response | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |

Table 4.2: Table to show the evolution of $\widehat{P}_t$ with $t$. Notice that the defender strategy engineers her strategy to ensure that $\widehat{P}_{t-1} = 1/2$ on odd rounds, eliciting an attacker response of 1; and $\widehat{P}_{t-1} > 1/2$ on even rounds, eliciting an attacker response of 2.

Table 4.2 shows the evolution of the empirical averages $\widehat{P}_t$ as a function of $t$ for the first 10 rounds (up-to 3 decimal points). Observe that when $t$ is even, $\widehat{P}_{t-1} > 1/2$, so the attacker attacks target 2. Similarly, when $t$ is odd, $\widehat{P}_{t-1} = 1/2$, so the attacker attacks[37] target 1. This is a highly desired outcome for the defender – note from Equation (4.25) that the defender is in fact defending target 2 on even rounds, and target 1 on odd rounds – so this rule results in successful defense on every round. Explicitly, we substitute these properties into Equation (4.24) to get the following defender payoff for the deterministic rule:

$$f_{T,\mathsf{avg}}(P_1, \ldots, P_T) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}[t \text{ even}] \cdot (1 - 3 \cdot 0) + \mathbb{I}[t \text{ odd}] \cdot 1 = 1.$$

By using the deterministic rule in Equation (4.25), the defender is doing something very simple: she is baiting an attacker into responding with attacking alternate targets in odd and even rounds. Since she is able to predict which target will be attacked next, she can defend that target in that round and achieve the maximal payoff of 1 in every round. Figure 4.13 shows that the defender effectively straddles the boundary between eliciting attacker responses 1 and 2, in clear contrast to the randomized defender rule from Section 4.9, which aimed to (almost) always elicit attacker response 1.

We have shown that the optimal payoff against naive attackers (equal to 1) is strictly better than even the ideal Stackelberg payoff (equal to 0.5). This additional payoff is arising,

[37] As with the Stackelberg solution concept, this follows critically from the tie-breaking assumption in favor of the defender. Even if the attacker broke ties randomly, the defender would be able to defend her target on half of the even rounds and achieve a time-averaged payoff of 3/4.

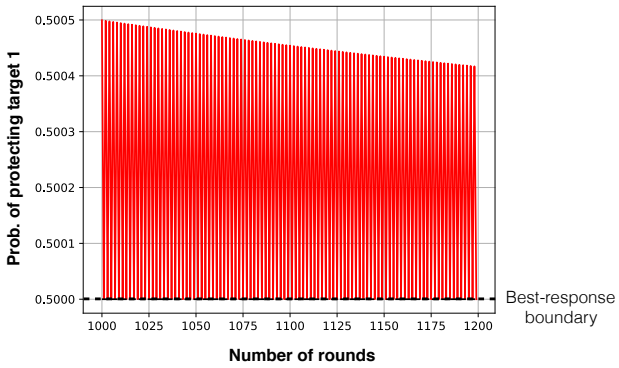



Figure 4.13: Evolution of naive leader rule, i.e. effective commitment power $\frac{1}{t}\sum_{s=1}^{t} p_s$ as a function of $t$. Figure from [90].

fundamentally, from *deception*. This is a general property for games in which deceiving followers is a dominant strategy over achieving ideal Stackelberg payoff. We characterize a general ensemble of games that satisfies this property, which we denote a *strictly-deception-dominant* ensemble, in Section 4.9.

We will see that the leader is not always incentivized to deceive followers – the case of persuasion tells quite a different story.

**Example 5.** Consider the time-averaged payoff function that we defined earlier for the prosecutor against judges who always use an empirical averages forecast, i.e. respond according to Equation (4.23):

$$f_{T,\mathsf{avg}}((p_1,\ldots,p_T)) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} P_t \mathbb{I}\left[\widehat{P}_{g,0}(t-1) \leq \frac{1}{2}\right]\right].$$

We considered randomized rules for the prosecutor that approximated her ideal Stackelberg payoff. We now ask whether the prosecutor can do better if she assumes judges who naively respond according to Equation (4.23). For any rule $(p_1,\ldots,p_T)$, we will see that

$$f_{T,\mathsf{avg}}((p_1,\ldots,p_T)) \leq \frac{T-N_T}{T} + \frac{N_T}{2}, \tag{4.26}$$

implying that

$$\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left[f_{T,\mathsf{avg}}((p_1,\ldots,p_T))\right] \leq \frac{1}{3} + \frac{1}{2}\cdot\frac{2}{3} = \frac{2}{3},$$

and thus the optimal payoff corresponds to the ideal Stackelberg payoff of the prosecutor. It turns out that this exact payoff can only achieved by *deterministic rules*. A full proof of the necessity of determinism is carried out for a more general ensemble of games that

includes persuasion games in Section 4.12, and we do not reproduce the argument here. We do, however, show here that the ideal Stackelberg payoff is achievable.

We consider prosecutor strategies that truthfully signal a guilty defendant, i.e. $P_t = 1$ whenever $\Pi_t = 1$. Recalling our notation for the epochs of innocent defendant arrivals $s_1, \ldots, s_j, \ldots, s_{N_T}$, we have for any deterministic strategy $(P_1, \ldots, P_T)$:

$$f_{T,\mathsf{avg}}((P_1, \ldots, P_T)) = \frac{1}{T}\sum_{t=1}^{T} (\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0]P_t) \cdot \mathbb{I}\left[\widehat{P}_{g,0}(t-1) \leq \frac{1}{2}\right]$$

The prosecutor rule $P_t$ has a non-trivial specification only on innocent defendant epochs $s_1, \ldots, s_{N_T}$. Here is an example of a prosecutor rule that maximizes the above objective:

$$P_{s_j} = \begin{cases} 0 \text{ if } j \text{ odd} \\ 1 \text{ if } j \text{ even.} \end{cases} \tag{4.27}$$

One can verify that this rule has the attractive property that $\widehat{P}_{g,0}(s_j - 1) \leq \frac{1}{2}$ for all $j \in [N_T]$, thus $\widehat{P}_{g,0}(t) \leq \frac{1}{2}$ for all $t \in [T]$. Thus, judges respond by always convicting when they see a guilty signal; resulting in overall payoff

$$\begin{aligned} f_{T,\mathsf{avg}}((P_1, \ldots, P_T)) &= \frac{1}{T}\sum_{t=1}^{T} (\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0]P_t) \cdot \mathbb{I}\left[\widehat{P}_{g,0}(t-1) \leq \frac{1}{2}\right] \\ &= \frac{T - N_T}{T} + \frac{1}{T}\sum_{j=1}^{N_T} P_{s_j} \\ &= \frac{T - N_T}{T} + \frac{N_T}{2T}, \end{aligned}$$

which, in expectation over the defendant sequence becomes exactly equal to $\frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} = \frac{2}{3}$.

There are also other deterministic rules that can satisfy this condition and help the prosecutor achieve exactly her ideal Stackelberg payoff. We have not shown it explicitly here (a formal argument is in Theorem 4.9.5) – but it turns out that *ideal Stackelberg payoff is the best she can do*, and moreover it is achievable only by a deterministic rule. At a high level, this is because eliciting convictions from the judges when they see a guilty signal is an *obviously dominant strategy* for the prosecutor – she always strongly prefers to do it, however few guilty signals she actually sends[38]. In the proof of Theorem 4.9.5, we characterize a general ensemble of $2 \times 2$ games for which this property holds and show that in such games, the leader cannot realize more than her ideal Stackelberg payoff even against naive followers. Moreover, this exact payoff can only be realized through deterministic rules.

[38] As long as the number is non-zero, which it will be since the prosecutor always signals guilty for truly guilty defendants.

### Vanishing credibility against intelligent followers

One should be naturally suspicious of the leader rules we just outlined. First off, *we do not* expect the followers (attackers and judges respectively) to continue to forecast using empirical averages (as in Definition 4.8.3) when they are aware that the leader (defender and prosecutor respectively) has deviated from a randomized rule with independent deployments. Such a forecast would be extremely sub-optimal[39] for them (and indeed, in the security game example, leads to the attackers always getting payoff 0).

Furthermore, the prevalence of *determinism* in the "optimal" rules means that not only will the followers deviate from the empirical averages forecasting rule, they can use a much, much better forecasting rule that completely compromises the leader's reputation. Let us see how this happens through our two examples.

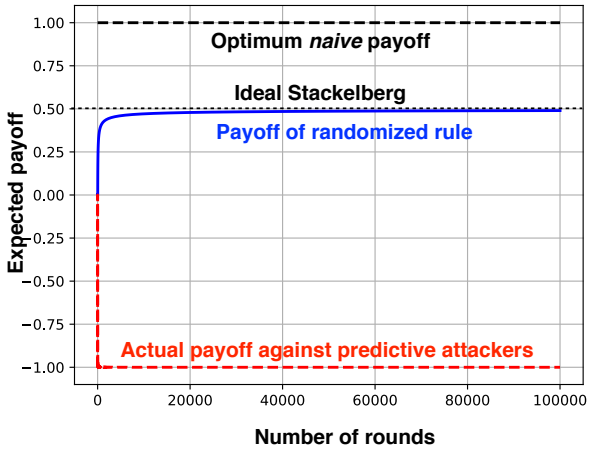



Figure 4.14: A comparison of the ensuing defender payoff from a) the randomized rule, b) the deterministic rule against naive attackers, c) the deterministic rule against intelligent attackers. Figure from [90].

**Example 4.** Recall that for the security game the uniquely optimal defender rule against naive attackers was in Equation (4.25), reproduced below:

$$P_t = \begin{cases} 1 \text{ if } t \text{ is odd} \\ 0 \text{ if } t \text{ is even} \end{cases}.$$

Consider attackers who, instead of using the empirical averages rule, use a *predictive forecasting rule* according to Definition 4.8.4 with the predictive model comprising the space

[39] In a hypothetical Bayesian interpretation of our framework, this would be because such followers contradict "expected" utility theory.

of all $K$-periodic sequences for some finite $K > 0$. Note that the leader sequence, which is 2-periodic, is deterministically predictable under this parameterization.

Recall that $P_t = \mathbb{I}[I_t = 1]$, i.e. indicator that defender defended target 1 in round $t$. With the above predictive forecast, the attacker will be able to *exactly* predict the value of $I_t$ in round $t$ given history $\mathcal{H}_{t-1}$, and will always respond with the *opposite* target, i.e.

$$J_t = \begin{cases} 2 \text{ if } t \text{ is odd} \\ 1 \text{ if } t \text{ is even} \end{cases}.$$

This obviously maximizes the attacker's payoff because it always succeeds in compromising a target. We denote the defender's expected payoff from a (randomized) rule $(p_1, \ldots, p_T)$ against attackers who use the predictive model $\Omega$ by $f_{T,\mathsf{pred}}(p_1, \ldots, p_T)$. For the deterministic leader rule in Equation (4.25), observe that

$$f_{T,\mathsf{pred}}(P_1, \ldots, P_T) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[t \text{ even}](0) + \mathbb{I}[t \text{ odd}](-2) = \frac{1}{2}(0-2) = -1,$$

which is strictly sub-optimal compared to the ideal Stackelberg payoff of 1/2 (or the payoff achieved by the randomized leader rule, which is $\frac{1}{2} - \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$). Thus, *the uniquely optimal solution to the dynamic program against naive attackers, in Equation* (4.24) *is extremely unstable to a smarter attacker response.* This is a generic property of the strictly-deception-dominant ensemble, which we characterize in Theorem 4.9.5. Figure 4.14 shows the stark contrast in the expected defender payoff from the periodic defender rule against naive followers (the black line) and intelligent followers (the red line), conveying the brittleness of this rule. In contrast to this brittleness, the randomized defender rule (the blue line) is not deterministically predictable, and thus robust to such attacker exploitation.

The brittleness of the optimal rule against naive followers is also a property in the persuasion example, as we see below.

**Example 5.** Recall that in the persuasion game, the only non-trivial signalling decisions are made at epochs of innocent defender arrivals, $s_1, \ldots, s_{N_T}$. All optimal solutions are *deterministic* and satisfy the properties

$$\frac{1}{N_T}\sum_{j=1}^{N_T} P_{s_j} = \frac{1}{2}$$

$$\widehat{P}_{g,0}(t-1) \leq \frac{1}{2} \text{ for all } t.$$

We considered a *periodic* (over innocent epochs) prosecutor rule that satisfied this property:

$$P_{s_j} = \begin{cases} 0 \text{ if } j \text{ odd} \\ 1 \text{ if } j \text{ even.} \end{cases}$$

While such rules are extremely effective against judges who naively use an empirical averages forecast, they are also extremely non-robust to smart judges that use a predictive forecast. Intelligent judges will quickly recognize the existence of a periodic prosecutor rule, for example; and be able to anticipate at every round *exactly* how the prosecutor would signal if the defendant were actually innocent. To see this, consider every round $t$ in which the prosecutor deterministically signals 'g' for guilty. There are two possibilities:

1. At round $t$, we have $\widehat{P}_{g,0}(t-1) = 1 > \frac{1}{2}$. In this case, there have been an odd number of innocent defendants so far. So the judge expects the prosecutor to next signal 'g' even if the defendant were innocent, and will *always acquit* regardless of which signal is sent[40].

2. At round $t$, we have $\widehat{P}_{g,0}(t-1) = 0 < \frac{1}{2}$. In this case, there have been an even number of innocent defendants so far. So the judge expects the prosecutor to next signal truthfully, and will *acquit* if $\Pi_t = 0$, i.e. if the defendant is truly innocent.

*The result of this is that intelligent judges never convict innocent defendants when faced with a periodic prosecutor. The prosecutor's persuasion power can also be compromised on several guilty defendants.*

Formally, the optimal judge response for every round $t \in (s_{j-1}, s_j]$ is as follows:

$$J_t = \begin{cases} \text{'a' if } P_{s,j} = 1 \\ \text{'a' if } P_{s,j} = 0, \Pi_t = 0 \\ \text{'c' if } P_{s,j} = 0, \Pi_t = 1. \end{cases}$$

and because the prosecutor can never expect payoff on innocent defendants, her expected payoff over time is at most $\frac{1}{3}$, the fraction of guilty defendants. In fact, one can show that she also loses her persuasion power on half of the guilty defendants on average, yielding expected payoff only $\frac{1}{6}$. This is sub-optimal compared to the optimal persuasion power of 2/3 commitments on average, and represents a situation in which the prosecutor overplays her hand in attempting to persuade – by resorting to deterministic rules, her incremental persuasion power on innocent defendants vanishes completely, and she even loses the power to facilitate conviction of half the guilty defendants. We see through this example that the brittleness of deterministic rules that was observed in security also manifests concretely in Bayesian persuasion.

### Lessons learned

The examples of security and persuasion are an instructive exercise in understanding how reputation could be *credibly* built up. Before we turn to formal statements of our results, some broad takeaways are summarized below:

[40] Because the prior tells him that innocence was more likely, and the prosecutor's signal is uninformative.

1. The robust commitment constructions that were used for the limited-observability can be also be leveraged to establish belief in leader commitment through repeated interaction by choosing different mixed strategies in every round. These strategies gradually approach the best-response-boundary – thus realizing the ideal Stackelberg payoff at a rate of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$. We show that this is generally possible for $m \times n$ games in Proposition 4.9.2, which builds on Theorem 4.7.4 from the limited-observation case.

2. Leader rules that are "optimal" against naive followers that respond to the empirical averages no matter what can achieve exactly the Stackelberg performance, *or* even higher performance through follower deception. However, they are extremely sub-optimal against intelligent followers.

3. This is because, in the examples considered, staying arbitrary close to the boundary in finitely repeated interaction *requires* determinism in the leader strategy, which can then be exploited by intelligent followers that use a predictive model. This statement is formalized in Theorem 4.9.5.

4. This suggests that leaders likely *need to* incorporate some randomization in their rules to truly build credibility.

While we saw that *naive* leader rules to induce follower deception are a bad idea, one can ask whether the leader is incentivized to deceive the follower in a more sophisticated manner[41]. In Proposition 4.9.6, we show that there is some potential for a benefit over and above ideal Stackelberg performance through the abstract ability to deceive – but this benefit is limited.

## Robust, randomized leader rules

First, we generalize the definition of the leader's expected payoff, averaged over time, when the followers respond using an empirical averages forecast according to Definition 4.8.3.

**Definition 4.9.1.** *Let the leader choose, as her rule,* ***randomized*** *strategies* $\mathbf{x}_1, \ldots, \mathbf{x}_T$ *with* ***independent deployments****. Then, against followers who respond using an empirical averages forecast, she will expect time-averaged payoff*

$$f_{T,\mathsf{avg}}(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \mathbb{E}_{(I_1,\ldots,I_T)\sim(\mathbf{x}_1,\ldots,\mathbf{x}_T)}\left[\frac{1}{T}\sum_{t=1}^{T} A_{I_t,j^*(\widehat{\mathbf{x}}_{t-1})}\right], \tag{4.28}$$

*where the expectation is only taken over the realizations of the randomized strategies (and ensuing follower responses).*

[41] For e.g., deceive without the follower being aware that he is being deceived.

As we saw in Examples 4 and 5, the robust commitment constructions in Theorem 4.7.4 can be naturally applied to help the leader approach her ideal Stackelberg performance at a specific rate. We provide a formal statement below.

**Proposition 4.9.2.** *Let the number of rounds $T = \mathcal{O}(m)$, and fix $0 < \eta < 1/2$. Then, any leader rule that uses the* ***robust commitment*** $\mathbf{x}_{t,\eta}$ *meant for round $t$ satisfies*

$$f^*_\infty - f_{T,\mathsf{avg}}(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \widetilde{\mathcal{O}}\left(\left(\frac{m}{T}\right)^\eta\right), \tag{4.29}$$

*where the $\widetilde{\mathcal{O}}(\cdot)$ contains constant factors that depend on both the local and global geometry of the best-response-region $\mathcal{R}_{j^*}$. For a formal statement that includes these factors, see Equation* (4.62)*.*

The proof of Proposition 4.9.2 is a fairly simple consequence of the properties of Theorem 4.7.4 together with assured independence in deployment, and is deferred to Appendix 4.12.

The implication of this result is that we have progressed from robust commitment constructions meant for a one-shot game with limited observability to a robust randomized leader rule meant for repeated interaction, during which both observability and belief have to be built up. In the limited-observability setting, the requirement for robustness was only that the expected follower response should be preserved under limited observability. Now, the randomized leader rule enjoys robustness in a much broader sense to strategic manipulation by followers. We justify this through two observations:

1. When the randomized leader rule is used, it is asymptotically good for rational follower $t$ to use the empirical averages forecast, and respond with $j^*(\widehat{\mathbf{X}}_{t-1})$ in the sense of asymptotic consistency[42], i.e. for the given rule we have

$$\lim_{t\to\infty}(\widehat{\mathbf{X}}_{t-1} - \mathbf{x}_t) = \mathbf{0} \text{ almost surely.} \tag{4.30}$$

   Equation (4.30) is justified through a non-asymptotic concentration bound on the total variation of $(\widehat{\mathbf{X}}_{t-1} - \mathbf{x}_t)$, which follows from a generalized version of Devroye's lemma (Lemma 4.12.2).

2. Because of the independence of randomness of deployments across rounds, follower $t$'s estimate of the mixed strategy at time $t$, $\widehat{\mathbf{X}}_{t-1}$, is *minimax-optimal* in the traditional information-theoretic/statistical sense [195, Fano's inequality]. Thus, follower $t$ does not even benefit in a *non-asymptotic sense* from deviating from the empirical averages forecast as in Definition 4.8.3.

[42] A similar condition is invoked by Fudenberg and Kreps [233] to justify the use of experimentation by agents; there it is called asymptotic myopic Bayes.

Because of this, not only does the leader expect favorable performance *provided that the followers use an empirical averages forecast*, the leader has no reason to believe that rational followers *would deviate from using an empirical averages forecast* when she is known to be using this randomized rule.

## Dis-incentives for determinism and exploitation of naivete

We saw in Examples 4 and 5 that the leader may have a temptation to exploit naive followers who *always* use the empirical averages forecasting rule, i.e. always play $J_t := j^*(\widehat{\mathbf{X}}_{t-1})$ regardless of the leader rule. More precisely, the leader could maximize her payoff by solving the ensuing dynamic objective $f_{T,\mathsf{avg}}(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ as defined in Equation (4.28).

In both examples, the leader rules that maximized this objective were deterministic, and they thus lost their credibility. Even for general $2 \times 2$ games, the dynamic program in Equation (4.28) does not have a closed form solution, nor is the optimal payoff easily evaluatable. However, we are able to show, for two broad ensembles of $2 \times 2$ games, that all leader rules that maximize the time-averaged payoff are indeed deterministic. We briefly describe these ensembles below. As before, the leader strategy is expressed as her probability of playing strategy 1, denoted by $p \in [0, 1]$. We denote the leader payoff as a function of strategy $p$ when follower responds with pure strategy $j$ by $f(p; j)$ and, without loss of generality, choose the best-response-function

$$j^*(p) = \begin{cases} 1 \text{ if } p \leq p^*_\infty \\ 2 \text{ otherwise.} \end{cases}$$

We also assume that the Stackelberg commitment is equal to $p^*_\infty \in (0, 1)$, i.e. it is mixed. The ideal Stackelberg payoff is denoted by $f^*_\infty$.

Our first ensemble is called a *strictly-deception-dominant ensemble.*

**Definition 4.9.3** (Strictly-deception-dominant ensemble.)**.** *A $2 \times 2$ leader-follower game, as defined above, is* ***strictly-deception-dominant*** *if we have*
$\max_{p\in[0,1]} f(p; 1) = \max_{p\in[0,1]} f(p; 2) > f^*_\infty$.

The strictly-deception-dominant ensemble is paradigmatic of the security game in Example 4. As the name suggests, for any game in this ensemble the leader is incentivized to systematically deceive *naive* followers into responding sub-optimally. This is to try and realize the payoff $\max_{p\in[0,1]} f(p; 1) = \max_{p\in[0,1]} f(p; 2)$, which is strictly greater than the ideal Stackelberg payoff. A more detailed description of this ensemble, and an intuitive illustration, is contained in Appendix 4.12.

Our second ensemble is called a one-response-obviously-dominant ensemble.

**Definition 4.9.4** (One-response-obviously-dominant ensemble.)**.** *A $2 \times 2$ leader-follower game is* ***one-response-obviously-dominant*** *if we have*
$\min_{p\in[0,1]} f(p; 1) > \max_{p\in[0,1]} f(p; 2)$.

The one-response-obviously-dominant ensemble is paradigmatic of the persuasion game in Example 5. In this ensemble, the leader has zero incentive to deceive the follower into responding sub-optimally, and simply wants to realize her maximal power of mixed commitment. A more detailed description of this ensemble, and an intuitive illustration, is contained in Appendix 4.12.

It turns out that in both of these ensembles, not only are followers no longer incentivized to use the empirical averages forecast; but also they can use a particular class of predictive forecasts to exploit the leader rule and make it highly sub-optimal.

**Theorem 4.9.5.** *For two continuous ensembles of* $2 \times 2$ *leader-follower games in which* $p^*_\infty \in (0,1)$ *and* $f_\infty(p)$ *is discontinuous at* $p^*_\infty$*, we have the following characterization of the naive dynamic program:*

1. *Any strategy* $(p^*_1, \ldots, p^*_T)$ *that maximizes* $f_{T,\mathsf{avg}}(p_1, \ldots, p_T)$ *is a deterministic strategy, i.e.* $(p^*_1, \ldots, p^*_T) \in \{0,1\}^T$.

2. *There exists a predictive forecast parameterized by* $\Omega$*, such that the expected leader payoff, averaged over time, against a follower using this predictive forecast is given by*

$$f_{T,\mathsf{pred}}(p^*_1, \ldots, p^*_T) \leq \max_{i \in \{0,1\}} f_\infty(i) < f^*_\infty.$$

Informally, Theorem 4.9.5 says that for these classes of games, any leader rule that is optimal against naive followers, who use the empirical averages forecast regardless of the leader rule, is strictly sub-optimal against intelligent followers, who use a predictable forecast on a class of deterministic sequences. Because of this strong predictability, the leader is restricted to attaining at most her *pure strategy Stackelberg payoff*, which is always strictly sub-optimal when the Stackelberg commitment is mixed.

Theorem 4.9.5 is proved for both the strictly-deception-dominant and the one-response-obviously-dominant ensembles. The proofs are contained in Appendix 4.12. For the strictly-deception-dominant ensemble, the leader's optimal strategy does deceive naive followers, but in an extremely brittle way – not only is the optimal strategy unique and deterministic, it turns out to be finitely periodic, making it extremely sub-optimal against smarter followers who use predictive forecasts. Here, Theorem 4.9.5 tells us that while there may be an incentive to deceive followers in their learning attempt, any attempt to realize this deception *naively* can be catastrophic for the leader[43].

For the one-response-obviously-dominant ensemble, the optimal payoff even against naive followers is exactly the ideal Stackelberg payoff. For these games (and only these games), Theorem 4.9.5 gives a formal framework to show the impossibility of *credibly* achieving the

[43]In fact, such naive attempts to deceive a follower are reminiscent of the "bad reputations" encountered by Ely, Valimaki, Fudenberg and Levine [192, 199] for the pure strategy model: in the SPNE of their repeated game(s), the followers opt out of participating with leaders who resemble a "bad type" too closely; and the leader plays in a manner to avoid resemblance to the bad type.

ideal Stackelberg payoff even against naive followers, and shows the necessity of some amount of independent randomization in the leader rule[44]. This impossibility result parallels the very first result in this chapter showing non-robustness of the traditional Stackelberg commitment to observational uncertainty (Proposition 4.7.1).

## Limited benefit of deception

We saw that for the *one-response-obviously-dominant* games, such as the persuasion game in Example 5, there is no incentive for the leader to elicit a follower mismatch through deception – and we have already seen that the leader incurs some sub-optimality from establishing a credible partial reputation. For others (like the security game in Example 4 and the $2 \times 3$ non-zero-sum game in Example 4), the leader could be incentivized to deceive naive followers. While strategies that deceive a naive follower may be deterministic and highly sub-optimal (like the ones that we saw in Example 4, more sophisticated attempts at deception need not have this brittle property. The more general test for whether deception can ever be a good idea for the leader would be to test an arbitrary leader rule against pessimistic followers, who do not expect predictability from the leader and follow a rule in the spirit of Hannan consistent/"no-regret" learning. Since we have defined all follower rules through sequential *forecasts*, the natural case to consider is an aggregate of followers who follow *universally calibrated forecasts*[45] as in Definition 4.8.5. We show that against such followers, the leader *could* obtain a benefit over and above the ideal Stackelberg payoff, but this benefit will necessarily be minimal.

**Proposition 4.9.6.** *For any leader rule $I_t := i_t(\mathcal{H}_{t-1})$ played against a sequence of followers using a universally calibrated forecast $\{\mathbf{r}_t\}_{t=1}^T$, we denote the* ***realized, time-averaged leader payoff*** *by $f_{T,\mathsf{calib}}(I_1, \ldots, I_T)$. Also let $N_T(\mathbf{r}) := \sum_{t=1}^T \mathbb{I}[\mathbf{r}_t = \mathbf{r}]$ denote the number of times the followers used forecast $\mathbf{r}$. Then, we have*

$$f_{T,\mathsf{calib}}(I_1, \ldots, I_T) \leq \frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \langle \mathbf{r}, \mathbf{a}_{j^*(\mathbf{r})} \rangle + f_{max} \frac{o(T)}{T} \text{ almost surely} \tag{4.31}$$

[44]This is similar in spirit to how we think of Hannan-consistent rules needing to be randomized [80]. Whether the randomization truly needs to be independent from round to round, necessitating a $\widetilde{\Theta}\left(\frac{1}{\sqrt{T}}\right)$ rate of approaching the ideal Stackelberg payoff, remains open and is an interesting question for future work. In the context of the rich literature on de-randomization through pseudo-random number generators, the answer to this question may depend on the nature of computational limitations imposed on leader and followers.

[45]Foster and Vohra [208] establish a general link between the ideas of (asymptotic) no-regret and calibration. It is interesting to note the necessity of the stronger condition of *calibrated forecasts*, as opposed to *no-regret payoffs*, on followers, to prove our forthcoming result – we were not able to get a result for the latter.

*for all realizations of randomness in the leader rule and follower forecasts. This can be upper bounded by*

$$f_{T,\mathsf{calib}}(I_1, \ldots, I_T) \leq f^*_\infty + f_{max}\frac{o(T)}{T}. \tag{4.32}$$

Proposition 4.9.6 follows quite directly from the fundamental definition of calibration and is deferred to Section 4.12. Equation (4.32) tells us what the best case for a leader could be against a universally calibrated forecast – not much more than ideal Stackelberg. The first term in the expression in Equation (4.31), which preserves the dependence of leader payoff on the follower forecasts, suggests that leader payoff could be sub-optimal if she incentivizes follower forecasts to deviate significantly from the Stackelberg payoff – indeed, she could gain a small amount from inducing more calibration error – but this small amount decreases as $T$ increases. This suggests that the leader has limited incentive, and possibly *dis-incentive*, to deceive even in a sophisticated manner, if she is facing pessimistic followers who use a universally calibrated rule. One can think of this result as the repeated-game analog of Theorem 4.7.5, which reached a similar conclusion for the much simpler one-shot model with observational uncertainty.

Taken in conjunction with Propositions 4.9.2 and Theorem 4.9.5, the overall flavor of our results for repeated interaction can be summed up in a few qualitative sentences:

1. For games in which the best possible "naive" payoff is ideal Stackelberg (i.e. there is no incentive to deceive followers), there is a fundamental, non-zero price of establishing partial reputation. Randomized leader rules achieve reputation at the rate of $\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)$ in a manner that is strategy-proof to follower manipulation, and are thus approximately optimal.

2. There is no trivial way of achieving an additional deception benefit, even if it exists – naive attempts, tailored to naive followers, can lead to strict sub-optimality. Maximally sophisticated approaches yield only a minimal benefit against universally calibrated followers, and could also be sub-optimal depending on the nature of the forecasts elicited.

## 4.10 Conclusions and future work

In this chapter, we have used the fact that repeated-game interaction with *one-sided learning* can be intricately connected to models of reputation building in the Bayesian setup. We then introduced a novel frequentist framework to posit explicit strategies for the designated leader and follower. While we uncovered a number of desirable properties in these strategies — namely, that adaptive followers are approximately optimal against a host of leader strategies, and simple randomized leader rules are approximately optimal against adaptive followers — this constitutes as yet partial progress towards the formal definition of a frequentist

SPNE. This is an important immediate goal for future work. Moreover, the reputation games constitute some of the simplest manifestations of repeated games with incomplete information. It is of obvious interest to develop a frequentist theory for repeated games with incomplete information at large, and for far more complex scenarios that might arise in modern online marketplaces. These scenarios could include two-sided learning, which we will discuss at a preliminary level in Chapter 5.

## 4.11 Proofs for limited observability

Before moving into the proofs themselves, we define some additional notation.

**Definition 4.11.1.** *The set of* ***alternate follower best response*** *to the mixed commitment* $\mathbf{x}$ *is denoted by*

$$\mathcal{K}^*_{\mathsf{alt}}(\mathbf{x}) := \mathcal{K}^*(\mathbf{x}) - \{j^*\}.$$

We will be particularly interested in this set for the Stackelberg commitment, that is, $\mathcal{K}^*_{\mathsf{alt}}(\mathbf{x}^*_\infty)$. In general, the set will be non-empty as the follower could be agnostic between more than one pure strategy in response – it is only responding with the pure strategy $j^*$ to break ties in the leader's favor. Figure 4.15 shows this demarcation of follower responses into the expected response $j^*$, and alternate responses to the Stackelberg commitment $\mathbf{x}^*_\infty$.

Further, we denote maximum and minimum obtainable leader payoffs respectively by

$$f_{max} := \max_{i\in[m],j\in[n]} A_{ij}$$
$$f_{min} := \min_{i\in[m],j\in[n]} A_{ij}.$$

### Proof of Proposition 4.7.1

We consider a general $2 \times n$ game and denote the Stackelberg probability of leader playing pure strategy 1 by $p^*_\infty$. Recall that $p^*_\infty \in (0,1)$ (since we have assumed for the proof that the Stackelberg commitment is mixed). Let $j_{\mathsf{alt}}$ be *the* alternate response to the Stackelberg commitment, i.e. we have $\mathcal{K}^*(p^*_\infty) = \{j^*_\infty, j_{\mathsf{alt}}\}$. Without loss of generality, the best-response regions can be described as

$$\mathcal{R}_{j^*_\infty} = [p^-, p^*_\infty]$$
$$\mathcal{R}_{j_{\mathsf{alt}}} = (p^*_\infty, p^+].$$

Finally, we define $f^{(2)} := \lim_{\epsilon\to 0} f_\infty(p^*_\infty + \epsilon)$. Since we are considering leader-follower games for which the function $f_\infty(.)$ is discontinuous at $p^*_\infty$, by the tie-breaking assumption on Stackelberg commitment we will have $f^{(2)} < f^*_\infty$.

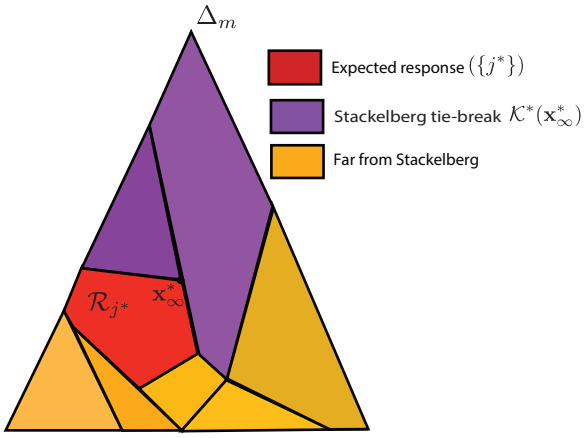



Figure 4.15: Illustration of partition of the set of follower responses, $[n]$, into sets $\{k^*\}$ (red region), *alternate best responses* (purple regions) and everything else (orange regions). Figure from [90].

Now, we consider the quantity $f_N(p^*_\infty)$. Denoting $\widehat{P}_N$ as the empirical estimate of the quantity $p^*_\infty$, we have

$$\begin{aligned}
f_N(p^*_\infty) &\leq \Pr\left[\widehat{P}_N \in \mathcal{R}_{j^*_\infty}\right] f^*_\infty + \Pr\left[\widehat{P}_N \in \mathcal{R}_{j_{\text{alt}}}\right] f^{(2)} \\
&\quad + \left(1 - \Pr\left[\widehat{P}_N \in \mathcal{R}_{j^*_\infty}\right] - \Pr\left[\widehat{P}_N \in \mathcal{R}_{j_{\text{alt}}}\right]\right) f_{max} \\
&= \Pr\left[\widehat{P}_N \in (p^-, p^*_\infty]\right] f^*_\infty + \Pr\left[\widehat{P}_N \in (p^*_\infty, p^+]\right] f^{(2)} \\
&\quad + \Pr\left[\widehat{P}_N \in [0, p^-] \cup (p^+, 1]\right] f_{max} \\
&= f^*_\infty - \underbrace{\Pr\left[\widehat{P}_N \in (p^*_\infty, p^+]\right]}_{T_1(N)} (f^*_\infty - f^{(2)}) + \underbrace{\Pr\left[\widehat{P}_N \in [0, p^-] \cup (p^+, 1]\right]}_{T_2(N)} (f_{max} - f^*_\infty).
\end{aligned}$$

We will now proceed to bound the probabilities $T_1(N)$ and $T_2(N)$.

First, we deal with the quantity $T_2(N)$, which reflects the probability of a mismatched response that is neither Stackelberg nor the alternate response on the boundary. By the Hoeffding bound, we have

$$\begin{aligned}
T_2(N) &:= \Pr\left[\widehat{P}_N \in [0, p^-] \cup (p^+, 1]\right] \\
&= \Pr\left[\widehat{P}_N \in [0, p^-]\right] + \Pr\left[\widehat{P}_N \in (p^+, 1]\right] \\
&\leq \exp\{-2N(p^*_\infty - p^-)^2\} + \exp\{-2N(p^+ - p^*_\infty)^2\}.
\end{aligned}$$

Denoting $C'' := 2\left(\min\{p^+ - p^*_\infty, p^*_\infty - p^-\}\right)^2$, we then have

$$T_2(N) \leq 2\exp\{-NC''\} \tag{4.33}$$

and as expected, this probability decays exponentially with $N$.

Next, we deal with the quantity $T_1(N)$, which reflects the probability of eliciting the alternate response on the Stackelberg boundary. We show that this event is non-vanishingly probable.

We define the following quantities

$$S_N := N\widehat{P}_N \tag{4.34}$$

$$Z_N := \frac{S_N - Np^*_\infty}{\sqrt{Np^*_\infty(1-p^*_\infty)}}. \tag{4.35}$$

Recall that $Z_N$ is a real-valued random variable. We denote its cumulative distribution function by $F_N(.)$.

By a simple change of variables, we then have

$$\begin{aligned} T_1(N) &= \Pr\left[\widehat{P}_N \in (p^*_\infty, p^+]\right] \\ &= \Pr\left[Z_N \in \left(0, \frac{\sqrt{N}(p^+ - p^*_\infty)}{\sqrt{p^*_\infty(1-p^*_\infty)}}\right]\right] \\ &= F_N(\frac{\sqrt{N}(p^+ - p^*_\infty)}{\sqrt{p^*_\infty(1-p^*_\infty)}}) - F_N(0). \end{aligned}$$

Now, recall that $S_N = \sum_{j=1}^N I_j$ for iid random variables $I_j \sim \text{Ber}(p^*_\infty)$. Also note that since we have considered games with mixed Stackelberg commitment, we have $0 < p^*_\infty < 1$. We now invoke the first half of the classical Berry-Esseen theorem [230, 231] stated here as a lemma.

**Lemma 4.11.2.** *There exists a positive constant $C$ such that if $I_1, I_2, \ldots$ are iid random variables with $\mathbb{E}[I_1] = \mu < \infty$, $\text{var}(I_1) = \sigma^2 > 0$ and $\mathbb{E}[|I_1 - \mu|^3] = \rho < \infty$, we have*

$$|F_N(x) - \phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{N}}$$

*for all $x \in \mathbb{R}$, where $\phi(.)$ denotes the CDF of the standard normal distribution $\mathcal{N}(0,1)$.*

It is easy to verify that the distribution $I_1 \sim \text{Ber}(p^*_\infty)$ satisfies the above conditions. Therefore, we can directly apply Lemma 4.11.2 and get

$$F_N\left(\frac{\sqrt{N}(p^+ - p^*_\infty)}{\sqrt{p^*_\infty(1-p^*_\infty)}}\right) \geq \phi(C\sqrt{N}) - \frac{C'}{\sqrt{N}} \text{ and}$$

$$F_N(0) \leq \frac{1}{2} + \frac{C'}{\sqrt{N}}$$

for positive constant $C > 0$, thus giving

$$T_1(N) \geq \left(\phi(C'\sqrt{N}) - \frac{1}{2}\right) - \frac{C'}{\sqrt{N}}. \tag{4.36}$$

Substituting for the expressions for $T_1(N)$ and $T_2(N)$, we now have

$$f^*_\infty - f_N(p^*_\infty) \geq \left(\left(\phi(C'\sqrt{N}) - \frac{1}{2}\right) - \frac{C'}{\sqrt{N}}\right) C - 2C\exp\{-NC''\},$$

which corresponds exactly to Equation (4.5). Clearly, the right hand side of this equation is decreasing in $N$ and so the first corollary – that $f_N(p^*_\infty) \leq f^*_\infty$ for $N \geq N_0$ – holds. Precisely, we have

$$\begin{aligned}\lim_{N\to\infty} \phi(\sqrt{N}C') &= 1\\ \lim_{N\to\infty} \frac{C'}{\sqrt{N}} &= 0\\ \lim_{N\to\infty} 2C\exp\{-NC''\} &= 0,\end{aligned}$$

and so we have

$$f^*_\infty - \lim_{N\to\infty} f_N(p^*_\infty) \geq \frac{f^*_\infty - f^{(2)}}{2}.$$

This is the second corollary from Proposition 4.7.1 and completes the proof. □

## Proof of Theorem 4.7.4

### Notation

For this proof, it will be convenient to consider the $(m-1)$-dimensional representation of the probability simplex, i.e.

$$\Delta_{m-1} := \{\mathbf{y} \succeq \mathbf{0} \text{ and } \langle \mathbf{y}, \mathbf{1}\rangle \leq 1\}.$$

Then, we can represent a commitment $\mathbf{x} \in \Delta_m$ by its $(m-1)$-dimensional representation $\mathbf{y} = \begin{bmatrix} x_1 & x_2 & \dots & x_{m-1}\end{bmatrix}$, and the *leader payoff* if the follower were to respond with pure strategy $j \in [n]$ by

$$\langle \mathbf{y}, \mathbf{c}_j\rangle + d_j$$

where we have

$$\mathbf{c}_j := \begin{bmatrix} a_{j,1} - a_{j,m} \\ a_{j,2} - a_{j,m} \\ \vdots \\ a_{j,m-1} - a_{j,m} \end{bmatrix}$$
$$d_j = a_{j,m}.$$

Similarly, we can represent the corresponding *follower payoff* by

$$\langle \mathbf{y},\, \mathbf{b}'_j \rangle + d'_j$$

where we have

$$\mathbf{b}'_j := \begin{bmatrix} b_{j,1} - b_{j,m} \\ b_{j,2} - b_{j,m} \\ \vdots \\ b_{j,m-1} - b_{j,m} \end{bmatrix}$$
$$d'_j = b_{j,m}.$$

We can also represent this representation of the empirical estimate of $\mathbf{y}$ from $N$ samples by $\widehat{\mathbf{Y}}_N$, and this representation Stackelberg commitment by $y^*_\infty$.

Now, we can consider all the functions introduced in Section 4.5 in terms of the commitment $\mathbf{x}$ and equivalently define them in terms of the $(m-1)$-dimensional representation of the commitment, $\mathbf{y}$.

We also denote the $p$th operator norm of a matrix by $\|\!|.|\!\|_p$.

### The commitment construction

We consider the $(m-1)$-dimensional representation of the best-response-region corresponding to the Stackelberg commitment, $\mathcal{R}_{j^*}$. There are many things to consider while constructing a robust commitment. The first, and obvious, one would be that the follower should respond the same way as it would to Stackelberg when it observes the full mixture. That is, we would have $j^*(\mathbf{y}_N) = j^*$ or alternatively stated, $\mathbf{y}_N \in \mathcal{R}_{j^*}$.

Intuitively, the expected payoff of a leader commitment under observational uncertainty, particularly in terms of gap to the optimal Stackelberg payoff, will depend on two factors: one, how likely the follower is to respond the same as it would if it observed the full commitment; and two, how "far" the leader commitment mixture is from the optimal Stackelberg commitment mixture. We qualitatively show this dependence in the following lemma.

**Lemma 4.11.3.** *Consider a commitment $\mathbf{y}_N$ for which we can provide the following guarantee:*

$$\Pr[\widehat{\mathbf{Y}}_N \notin \mathcal{R}_{j^*}] \leq \epsilon_N.$$

*We then have*

$$f_\infty^* - f_N(\mathbf{y}_N) \leq 2(1-\epsilon_N) f_{max} \|\mathbf{y}_N - y_\infty^*\|_1 + \epsilon_N (f_\infty^* - f_{min})$$

*Proof.* We have

$$\begin{aligned} f_N(\mathbf{y}_N) &= \sum_{j=1}^{n} \Pr[\widehat{\mathbf{Y}}_N \in \mathcal{R}_j](\langle \mathbf{y}_N, \mathbf{c}_j \rangle + d_j) \\ &\geq \Pr[\widehat{\mathbf{Y}}_N \in \mathcal{R}_{j^*}](\langle \mathbf{y}_N, \mathbf{c}_{j^*} \rangle + d_{j^*}) + (1 - \Pr[\widehat{\mathbf{Y}}_N \in \mathcal{R}_{j^*}]) f_{min} \\ &\geq (1-\epsilon_N)\left(\langle \mathbf{y}_N, \mathbf{c}_{j^*} \rangle + d_{j^*} - f_{min}\right) + f_{min} \\ &= (1-\epsilon_N)\left(\langle \mathbf{y}_N, \mathbf{c}_{j^*} \rangle + d_{j^*}\right) + \epsilon_N f_{min}. \end{aligned}$$

Recall that we have $f_\infty^* = \langle y_\infty^*, \mathbf{c}_{j^*} \rangle + d_{j^*}$. Therefore, the gap from Stackelberg is bounded as

$$\begin{aligned} f_\infty^* - f_N(\mathbf{y}_N) &\leq (1-\epsilon_N)\langle y_\infty^* - \mathbf{y}_N, \mathbf{c}_{j^*} \rangle + \epsilon_N (f_\infty^* - f_{min}) \\ &\leq (1-\epsilon_N)\|\mathbf{c}_{j^*}\|_\infty \|\mathbf{y}_N - y_\infty^*\|_1 + \epsilon_N (f_\infty^* - f_{min}) \\ &\leq 2(1-\epsilon_N) f_{max} \|\|\mathbf{y}_N - y_\infty^*\|_1 + \epsilon_N (f_\infty^* - f_{min}), \end{aligned}$$

where the second inequality follows from Holder's inequality. This proves the lemma. □

This lemma implies that we want a commitment construction $\mathbf{y}_N$ with the following two-fold guarantee[46].

1. $\|\mathbf{y}_N - y_\infty^*\|_1$ is bounded (and ideally vanishes with $N$).
2. $\widehat{\mathbf{Y}}_N \in \mathcal{R}_{j^*}$ with high probability.

### Commitment construction using localized geometry

We will leverage the special structure of the Dikin ellipsoid [234] used in interior-point methods to make our commitment constructions. Observe that $y_\infty^*$ is always going to be on an extreme point (vertex) of the best-response-polytope[47] $\mathcal{R}_{j^*}$. We now collect the $k = |\mathcal{K}^*(\mathbf{x}_\infty^*)|$ constraints that are satisfied *with equality* at $\mathbf{x}_\infty^*$:

$$\langle \mathbf{y}, \mathbf{b}'_j \rangle + d'_j \leq \langle \mathbf{y}, \mathbf{b}'_{j^*} \rangle + d'_{j^*} \text{ for all } j \in \mathcal{K}^*(\mathbf{x}_\infty^*).$$

[46]Interestingly, the fact that $y_\infty^*$ is on an extreme point of $\mathcal{R}_{j^*}$ will imply that the two conditions are at odds with one another, and we will need to trade them off. For instance, choosing $\mathbf{y}_N = y_\infty^*$ would satisfy the second condition perfectly by being as close as possible to the Stackelberg commitment, but there would be no guarantee on the best-response as it lies on the boundary of the best-response region.

[47]Recall that the Stackelberg equilibrium is the solution to the LP defined on the best-response-polytope [214].

This is simply the constraint set for commitments such that the follower prefers to respond with pure strategy $j^*$ over any pure strategy $j \in \mathcal{K}^*(y^*_\infty)$ (i.e. any pure strategy whose corresponding best-response-polytope shares a boundary with the Stackelberg best-response-polytope at point $y^*$), and can be thought of as the set of *local constraints to the Stackelberg vertex* in the best-response polytope $\mathcal{R}_{j^*}$. We also collect the other constraints that describe $\mathcal{R}_{j^*}$:

$$\begin{aligned} \langle \mathbf{y},\, \mathbf{b}'_j \rangle + d'_j &\leq \langle \mathbf{y},\, \mathbf{b}'_{j^*} \rangle + d'_{j^*} \text{ for all } j \notin \mathcal{K}^*(\mathbf{x}^*_\infty) \cup \{j^*\} \\ \mathbf{y} &\succeq 0 \\ \langle \mathbf{1},\, \mathbf{y} \rangle &\leq 1, \end{aligned}$$

and together with the local constraints at the Stackelberg vertex, these describe the global constraints for the polytope.

We represent the system of inequalities in matrix form as: $B\mathbf{y} \preceq \mathbf{c}$ for some $B \in \mathbb{R}^{k \times (m-1)}$ and some $\mathbf{c} \in \mathbb{R}^k$. We leverage the following useful fact about a general set of linear constraints.

**Fact 4.11.4.** *For any parameterization of linear constraints $(B, \mathbf{c})$, there exists an affine transformation $\mathbf{y}' = T_1\mathbf{y} + T_2$ (where $T_1 \in \mathbb{R}^{(m-1)\times(m-1)}$ is invertible and $T_2 \in \mathbb{R}^{m-1}$) and a matrix $B' \in \mathbb{R}^{k\times(m-1)}$ such that*

$$B\mathbf{y} \preceq \mathbf{c} \iff B'\mathbf{y}' \preceq \mathbf{1}.$$

*We denote the transformation function by $T(\cdot)$ and its inverse by $T^{-1}(\cdot)$. In particular, we note the relationship $B = B'T_1$.*

The above fact is useful[48] because it is most convenient to define our class of commitments in the transformed space $\mathbf{y}' = T(\mathbf{y})$.

**Definition 4.11.5.** *For a particular value of $\delta \in (0, 1)$, Stackelberg commitment $y^*_\infty$, and local constraints modeled by $(B, \mathbf{c})$, we define a $\delta$-deviation commitment by*

$$\begin{aligned} \mathbf{y}(\delta; y^*_\infty) &:= T^{-1}(\mathbf{y}'(\delta; (y^*)'_\infty)) \text{ where} \\ \mathbf{y}'(\delta; (y^*)'_\infty) &:= (1-\delta)(y^*)'_\infty. \end{aligned}$$

Our robust commitments $\{\mathbf{y}_N\}_{N\geq 1}$ are going to be taken out of the set of $\delta$-deviation commitments, with appropriately chosen values of $\{\delta_N\}_{N\geq 1}$. *Clearly, the computational complexity of constructing any $\delta$-deviation commitment is equivalent to the complexity of computing the Stackelberg equilibrium itself.*

[48] A subtle point is that there do exist special cases of polytope constraints for which Fact 4.11.4 is true only with an augmentation of the variable space from $m$ to $2m$ dimensions. Then, defining the invertible map becomes trickier. Nevertheless, for ease of exposition and clarity in the proof, we assume that we can indeed carry out the affine transformation without augmenting the dimension.

To understand how to set these values, we will turn to the question of how to satisfy the three conditions above.

First, we observe that $\mathbf{y}(\delta; y^*_\infty)$ satisfies the *local constraints* $B\mathbf{y} \preceq \mathbf{c}$ for any $\delta \in (0,1)$. Because of Fact 4.11.4, it suffices to show that its affine transformation $\mathbf{y}'(\delta; (y^*)'_\infty)$ satisfies the local constraints $B'\mathbf{y}' \preceq \mathbf{1}$. Recall that $(y^*)'_\infty$ satisfies *all the local constraints* with equality, i.e. we have $B'(y^*)'_\infty = \mathbf{1}$. From the definition of the commitment, we thus have

$$\begin{aligned} B'\mathbf{y}'(\delta; (y^*)'_\infty) &= (1-\delta)B'(y^*)'_\infty \\ &= (1-\delta)\mathbf{1} \preceq \mathbf{1}. \end{aligned}$$

Next, we turn to the question of how close such a defined commitment would be from the Stackelberg commitment $y^*_\infty$, in terms of the $\ell_1$ norm. For this, we have

$$\begin{aligned} \|\mathbf{y}(\delta; y^*_\infty) - y^*_\infty\|_1 &= \|T_1^{-1}(\mathbf{y}'(\delta; (y^*)'_\infty) - (y^*)'_\infty)\|_1 \\ &= \delta\|T^{-1}(y^*)'_\infty\|_1 \\ &= \delta\|y^*_\infty\|_1 \leq \delta. \end{aligned}$$

Therefore, we have

$$\|\mathbf{y}(\delta; y^*_\infty) - y^*_\infty\|_1 \leq \delta. \tag{4.37}$$

In lieu of Lemma 4.11.3, we wish to choose values $\{\delta_N\}_{N\geq 1}$ (to create commitments $\{\mathbf{y}_N\}_{N\geq 1}$) such that $\delta_N$ decreases with $N$ sufficiently fast, while maintaining a high probability of staying in the best-response polytope $\mathcal{R}_{j^*}$. To understand the rate at which we can decrease $\delta_N$, we need to prove a high-probability best-response guarantee.

### Using the local Dikin ellipsoid as a confidence ball

For a (affine-transformed) commitment $\mathbf{y}'(\delta; (y^*)'_\infty)$, we make use of the *local Dikin ellipsoid* centered at $\mathbf{y}'(\delta; (y^*)'_\infty)$, defined below for an arbitrary point $\mathbf{y}'$.

**Definition 4.11.6** ( [234]). *For constraint set $B'\mathbf{y}' \preceq \mathbf{1}$, the **Dikin ellipsoid** of radius $r$ centered at $\mathbf{y}'$ is given by*

$$\mathbb{B}_{B',\mathbf{1},\mathbf{y}'}(r) := \{\mathbf{z}' : (\mathbf{z}' - \mathbf{y}')^\top H(\mathbf{y}')(\mathbf{z}' - \mathbf{y}') \leq r\}, \tag{4.38}$$

*where we define*

$$H(\mathbf{y}') := \sum_{i=1}^{k} \frac{(\mathbf{b}')_i (\mathbf{b}')_i^\top}{(1 - \langle (\mathbf{b}')_i, \mathbf{y}' \rangle)^2}. \tag{4.39}$$

*The Dikin ellipsoid has two special properties [234]:*

1. ***Affine invariance****: (using the notation from Fact 4.11.4) For transformation* $\mathbf{y}' = T(\mathbf{y})$*, the Dikin ellipsoid of radius* $r$ *centered at the point* $\mathbf{y}$ *for the polytope* $B\mathbf{y} \preceq \mathbf{c}$ *is* $\mathbb{B}_{B,\mathbf{c},\mathbf{y}'}(r) = T^{-1}(\mathbb{B}_{B',\mathbf{1},\mathbf{y}'}(r))$.

2. ***Interior guarantee****: For any interior point* $\mathbf{y}'$ *(according to the constraint set* $B'\mathbf{y}' \preceq \mathbf{1}$*), the Dikin ellipsoid of radius* 1 *centered at* $\mathbf{y}'$ *is contained in the feasibility set, that is,*

$$\mathbf{z}' \in \mathbb{B}_{B',\mathbf{1},\mathbf{y}'}(1) \implies B'\mathbf{z}' \preceq \mathbf{1}.$$

We center our Dikin ellipsoid at $\mathbf{y}'(\delta; (y^*)'_\infty)$, and observe that the constraint takes on a particularly nice form, as stated by the following simple lemma.

**Lemma 4.11.7.** *For any* $\delta \in (0, 1)$*, the Dikin ellipsoid can be expressed as*

$$\mathbb{B}_{B',\mathbf{1},\mathbf{y}'(\delta;(y^*)'_\infty)}(1) = \{\mathbf{z}' : \|B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty))\|_2 \leq \delta\}. \tag{4.40}$$

*Furthermore, in the original space we can write*

$$\mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1) = \{\mathbf{z} : \|B(\mathbf{z} - \mathbf{y}(\delta; y^*_\infty))\|_2 \leq \delta\}. \tag{4.41}$$

*Proof.* From Definition 4.11.5, we observe that $B'\mathbf{y}'(\delta; (y^*)'_\infty) = (1-\delta)B'(y^*)'_\infty = (1-\delta)\mathbf{1}$. This implies that

$$1 - \langle (\mathbf{b}')_i, \mathbf{y}'(\delta; (y^*)'_\infty)\rangle = 1 - (1-\delta) = \delta,$$

and thus we have

$$\begin{aligned} H(\mathbf{y}'(\delta; (y^*)'_\infty)) &= \frac{\sum_{i=1}^k (\mathbf{b}')_i (\mathbf{b}')_i^\top}{\delta^2} \\ &= \frac{(B')^\top B'}{\delta^2} \end{aligned}$$

where in the last equality step, we have used $(B')^\top B' = \sum_{i=1}^k (\mathbf{b}')_i (\mathbf{b}')_i^\top$, noting that $(\mathbf{b}')_i$ denotes the $i^{th}$ row of $B'$.

Thus, the ellipsoid constraint in Equation (4.38) can be rewritten as

$$\begin{aligned} \frac{1}{\delta^2}(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty))^\top (B')^\top B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty)) \leq 1 \\ \implies \|B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty))\|_2^2 \leq \delta^2 \\ \implies \|B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty))\|_2 \leq \delta, \end{aligned}$$

thus completing the first part of the proof (Equation (4.40)).

For the second part of the proof, we use the affine invariance property of the Dikin ellipsoid, which tells us that

$$\begin{aligned}\mathbf{z} \in \mathbb{B}_{B,\mathbf{c},\mathbf{y}}(1) &\implies \mathbf{z}' = T_1\mathbf{z} + T_2 \in \mathbb{B}_{B',\mathbf{1},\mathbf{y}'}(1) \\ &\implies \|B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty))\|_2 \leq \delta.\end{aligned}$$

Now, observe that

$$\begin{aligned}B'(\mathbf{z}' - \mathbf{y}'(\delta; (y^*)'_\infty)) &= B'(T_1\mathbf{z} + T_2 - T_1\mathbf{y}(\delta; y^*_\infty) - T_2) \\ &= (B'T_1)(\mathbf{z} - \mathbf{y}(\delta; y^*_\infty)) \\ &= B(\mathbf{z} - \mathbf{y}(\delta; y^*_\infty))\end{aligned}$$

where in the last step we have used the relationship $B = B'T_1$ from Fact 4.11.4. Putting these observations together, we have

$$\mathbf{z} \in \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1) \implies \|B(\mathbf{z} - \mathbf{y}(\delta; y^*_\infty))\|_2 \leq \delta,$$

completing the second part of the proof. □

At this stage, it is worth remembering that the commitment is *mixed*, and the payoff from using a $\delta$-deviation commitment $\mathbf{y}(\delta; y^*_\infty) \in \Delta_{m-1}$ under a finite number of observations $N$ depends on the guarantee that its observed empirical distribution $\widehat{\mathbf{Y}}_N$ (typically) stays inside the best-response region. As a starting point we need to guarantee that at least the *local vertex constraints* are not violated.

Note that $\mathbf{y}(\delta; y^*_\infty) \in \Delta_{m-1}$ is an interior point for any $\delta > 0$, and thus the interior guarantee property of the Dikin ellipsoid can be applied. We thus know that if the empirical distribution of the commitment stays inside the Dikin ellipsoid centered at the actual commitment, it will stay inside the local constraint feasibility set. Thus, it makes sense to use the Dikin ellipsoid as a confidence ball and tail bound the probability that the empirical estimate lies outside this ball. Because of the weighted $\ell_2$-ball structure on the particular ellipsoid corresponding to a $\delta$-deviation commitment that we proved in Lemma 4.11.7, this is not difficult to do. We state this formally in the following lemma.

**Lemma 4.11.8.** *For a given $\delta > 0$, let $\widehat{\mathbf{Y}}_N$ be the empirical distribution of $N$ samples drawn from the $\delta$-deviation commitment $\mathbf{y}(\delta; y^*_\infty)$. Then, we have*

$$\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1)\right] \leq 3\exp\{-\frac{N\delta^2}{25|\!|\!|B|\!|\!|^2_{\mathsf{op}}}\}$$

*provided that $N \geq \frac{20m|\!|\!|B|\!|\!|^2_{\mathsf{op}}}{\delta^2}$.*

*Proof.* The proof is a simple consequence of Devroye's lemma [235], which tail bounds the total variation between the empirical estimate of a discrete distribution and the true distribution.

**Lemma 4.11.9** ([235]). *Let $\widehat{\mathbf{Y}}_N$ be the empirical distribution of $N$ samples drawn from any distribution $\mathbf{y} \in \Delta_{m-1}$. Then, as long as $\delta \geq \sqrt{\frac{20m}{N}}$ we have*

$$\Pr\left[\|\widehat{\mathbf{Y}}_N - \mathbf{y}\|_1 \geq \delta\right] \leq 3\exp\{\frac{-N\delta^2}{25}\}.$$

We note from Lemma 4.11.7 that

$$\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1) \implies \|B(\widehat{\mathbf{Y}}_N - \mathbf{y}(\delta;y^*_\infty))\|_2 > \delta,$$

and thus, we have

$$\begin{aligned}
\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1)\right] &= \Pr\left[\|B(\widehat{\mathbf{Y}}_N - \mathbf{y}(\delta;y^*_\infty))\|_2 > \delta\right] \\
&\overset{\text{(i)}}{\leq} \Pr\left[\|B\|_{op}\|\widehat{\mathbf{Y}}_N - \mathbf{y}(\delta;y^*_\infty))\|_2 > \delta\right] \\
&\overset{\text{(ii)}}{\leq} \Pr\left[\|B\|_{op}\|\widehat{\mathbf{Y}}_N - \mathbf{y}(\delta;y^*_\infty))\|_1 > \delta\right] \\
&= \Pr\left[\|\widehat{\mathbf{Y}}_N - \mathbf{y}(\delta;y^*_\infty))\|_1 > \delta/\|B\|_{op}\right]
\end{aligned}$$

where inequality (i) uses the definition of the operator norm and inequality (ii) uses the fact that $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1$ for any finite-dimensional vector $\mathbf{v}$. Applying Lemma 4.11.9 directly then gives us

$$\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1)\right] \leq 3\exp\{-\frac{N\delta^2}{25|||B|||^2_{\mathsf{op}}}\}$$

as long as

$$\begin{aligned}
&\frac{\delta}{|||B|||_{\mathsf{op}}} \geq \sqrt{\frac{20m}{N}} \\
\implies &N \geq \frac{20m|||B|||^2_{\mathsf{op}}}{\delta^2}.
\end{aligned}$$

This completes the proof.

□

**Completing proof of Theorem 4.7.4: Ensuring global constraint satisfiability**

Let us recap what we have proved so far about a $\delta$-deviation commitment $\mathbf{y}(\delta;y^*_\infty)$ for any $\delta \in (0,1)$.

1. For $N$ samples from $\mathbf{y}(\delta;y^*_\infty)$, we have $\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1)\right] \leq 3\exp\{-\frac{N\delta^2}{25|||B|||^2_{\mathsf{op}}}\}$ (from Lemma 4.11.8).

2. $\|\mathbf{y}(\delta; y^*_\infty) - y^*_\infty\|_1 \leq \delta$.

Thus, from Lemma 4.11.3 we have for any $\delta$-deviation commitment,

$$f^*_\infty - f_N(\mathbf{y}(\delta; y^*_\infty)) \leq 2\delta f_{max} + \Pr\left[\widehat{\mathbf{Y}}_N \notin \mathcal{R}_{j^*}\right](f^*_\infty - f_{min})$$

Thus, if we had $\mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1) \subset \mathcal{R}_{j^*}$, we would have

$$\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathcal{R}_{j^*}\right] \leq \Pr\left[\widehat{\mathbf{Y}}_N \notin \mathbb{B}_{B,\mathbf{c},\mathbf{y}(\delta;y^*_\infty)}(1)\right] \leq 3\exp\{-\frac{N\delta^2}{25\|B\|^2_{\mathsf{op}}}\}.$$

However, the set $\mathcal{R}_{j^*}$ includes *global constraints* in addition to the local constraints $B\mathbf{y} \preceq \mathbf{c}$, and all points in the *local* Dikin ellipsoid need not satisfy these constraints. This is the final technicality in the proof that we now deal with. We will see that for a small enough value of $\delta$ (that depends on how the local geometry of the polytope relates to the global geometry), we can guarantee global satisfiability. Let the constraints corresponding to the convex polytope $\mathcal{R}_{j^*}$ be represented by $C\mathbf{y} \preceq \mathbf{d}$, and the corresponding constraints after the affine transformation $(T_1, T_2)$ be represented as $C'y' \preceq \mathbf{d}'$ (where the values of $\mathbf{d}'$ corresponding the local constraints are 1). Thus, for the vertex $(y^*)'_\infty$, we can define the quantity

$$\mathcal{Z}(\mathcal{R}_{j^*}; (y^*)'_\infty) := \sup\{\delta > 0 : \mathbf{z}' \in \mathbb{B}_{B',\mathbf{1},y'(\delta;(y^*)'_\infty)}(1) \implies C'\mathbf{z}' \preceq \mathbf{d}'\}. \tag{4.42}$$

Because $\mathcal{R}_{j^*}$ is *non-empty and convex*, we have $\mathcal{Z}(\mathcal{R}_{j^*}; (y^*)'_\infty) > 0$.
From this definition, under the condition $\delta < \mathcal{Z}(\mathcal{R}_{j^*}; (y^*)'_\infty)$ we have

$$\begin{aligned}&\mathbb{B}_{B',\mathbf{1},y'(\delta;(y^*)'_\infty)}(1) \subset T(\mathcal{R}_{j^*})\\ &\implies \mathbb{B}_{B,\mathbf{1},y(\delta;y^*_\infty}(1) \subset \mathcal{R}_{j^*},\end{aligned}$$

where the last implication is because of the affine-invariance property of the Dikin ellipsoid.

On the other hand, we used the condition $N \geq \frac{20m\|B\|^2_{\mathsf{op}}}{\delta^2}$ to prove Lemma 4.11.8. Combining these inequalities tells us that we require $N > \frac{20m\|B\|^2_{\mathsf{op}}}{\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2} = \mathcal{O}(m)$ to prove our result.

Then, we formally define our robust commitment for a particular value of $N$ below, and prove this final lemma which is essentially a formal statement of Theorem 4.7.4.

**Lemma 4.11.10.** *For every $N > \frac{20m\|B\|^2_{\mathsf{op}}}{\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2}$, and every $\eta < 1/2$, we define the $\eta$-robust commitment as a $\delta_{N,\eta}$-deviation commitment $y_{N,\eta} := y(\delta_{N,\eta}; y^*_\infty)$, where*

$$\delta_{N,\eta} := \mathcal{Z}(\mathcal{R}_{j^*}; (y^*)'_\infty)\left(\frac{m}{N}\right)^\eta. \tag{4.43}$$

*We then have*

$$\begin{aligned} f_N(y_{N,\eta}) &\leq \frac{2f_{max}}{\cdot}\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)\cdot\left(\frac{m}{N}\right)^\eta \\ &\quad + 3\exp\{-\frac{m^{2\eta}\cdot\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2\cdot N^{1-2\eta}}{25\|B\|^2_{\mathsf{op}}}\}(f^*_\infty - f_{min}) \\ &= \mathcal{O}\left(\left(\frac{m}{N}\right)^\eta + \exp\{-C\cdot N^{1-2\eta}\}\right). \end{aligned}$$

*Proof.* This is a simple consequence of everything put together. Since $N > m$, we have $\delta_{N,\eta} < \mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)$ and thus we have $\mathbb{B}_{B,\mathbf{1},y(\delta_{N,\eta};y^*_\infty}(1) \subset \mathcal{R}_{j^*}$. This tells us that

$$\Pr\left[\widehat{\mathbf{Y}}_N \notin \mathcal{R}_{j^*}\right] \leq 3\exp\{-\frac{N\delta^2_{N,\eta}}{25\|B\|^2_{\mathsf{op}}}\}.$$

and thus from Lemma 4.11.3 we get the following expression:

$$f^*_\infty - f_N(y_N) \leq 2\delta_{N,\eta}f_{max} + 3\exp\{-\frac{N\delta^2_{N,\eta}}{25\|B\|^2_{\mathsf{op}}}\}(f^*_\infty - f_{min}).$$

Directly substituting the expression for $\delta_{N,\eta}$ in Equation (4.43) into the above expression completes the proof. □

## Proof of Theorem 4.7.5

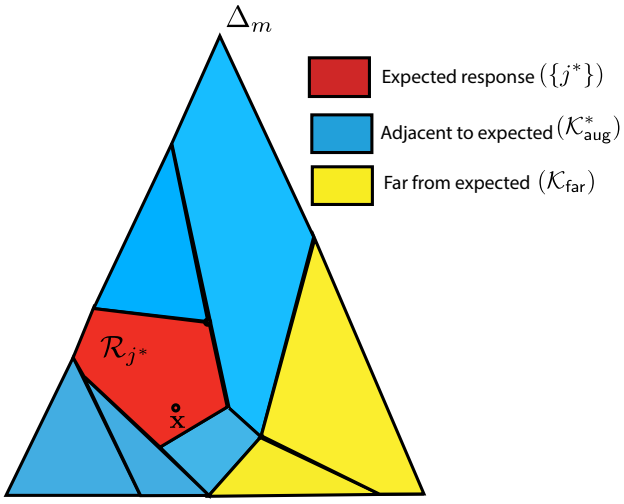



Figure 4.16: Illustration of partition of the set of follower responses, $[n]$, into sets $\{j^*\}$ (red region), $\mathcal{K}^*_{\mathsf{aug}}$ (blue regions) and $\mathcal{K}^*_{\mathsf{far}}$ (yellow regions). Figure from [90].

Recall that $f_N^* := \max_{\mathbf{x}\in\Delta_m} f_N(\mathbf{x})$. To prove an upper bound on $f_N^*$, we will upper bound $f_N(\mathbf{x})$ for every $\mathbf{x} \in \Delta_m$.

Without loss of generality the same proof method will extend to all $\mathbf{x} \in \Delta_m$. Denoting as shorthand $p_j(\mathbf{x}) := \Pr\left[\widehat{\mathbf{X}}_N \in \mathcal{R}_j\right]$, we have

$$f_N(\mathbf{x}) = \sum_{j=1}^{n} p_j(\mathbf{x})\langle \mathbf{a}_j, \mathbf{x}\rangle \tag{4.44}$$

$$= \sum_{j=1}^{n} T_j(\mathbf{x}) \tag{4.45}$$

where we denote $T_j(\mathbf{x}) := p_j(\mathbf{x})\langle \mathbf{a}_j, \mathbf{x}\rangle$. We will proceed to upper bound the quantity $T_j(\mathbf{x})$ for every $\mathbf{x} \in \Delta_m$ and every $j \in [n]$.

To do this, we will see that it is natural to divide all the pure strategy responses to a commitment $\mathbf{x}$ into three categories. The first is the expected response $j^*(\mathbf{x})$. The second is the set of responses whose regions are *adjacent* to the expected response as defined below.

**Definition 4.11.11.** *For a particular commitment $\mathbf{x} \in \Delta_m$, the set of **adjacent to expected** responses $\mathcal{K}^*_{aug}(\mathbf{x})$ is the set of all best-responses whose corresponding best-response-regions share a boundary with the best-response-region corresponding to the best response to $\mathbf{x}$. Formally, we have*

$$\mathcal{K}^*_{\mathsf{aug}}(\mathbf{x}) := \{j \in [n] : j \neq j^*(\mathbf{x}) \text{ and } cl(\mathcal{R}_{j^*(\mathbf{x})}) \cap cl(\mathcal{R}_j) \neq \emptyset\}.$$

*We also define $\mathcal{K}_{\mathsf{far}} := [n] - (\{j^*(\mathbf{x})\} \cup \mathcal{K}^*_{\mathsf{aug}}(\mathbf{x}))$ as the set of all follower responses that are "far" from the expected response in this sense.*

The illustration in Figure 4.16 shows this division.

For the rest of the proof, we will drop the term $\mathbf{x}$ from the notation and denote $\mathcal{K}^*_{\mathsf{aug}} := \mathcal{K}^*_{\mathsf{aug}}(\mathbf{x})$ as well as $j^* := j^*(\mathbf{x})$. This is done for notational simplicity.

It is first easy to show a bound on $T_{j^*}(\mathbf{x})$. In particular, we can directly use the definition of the function $f_\infty(.)$ to obtain

$$T_{j^*}(\mathbf{x}) = p_{j^*}(\mathbf{x})\langle \mathbf{a}_{j^*}, \mathbf{x}\rangle \tag{4.46}$$

$$= p_{j^*}(\mathbf{x}) f_\infty(\mathbf{x}) \tag{4.47}$$

$$\leq p_{j^*}(\mathbf{x}) f_\infty^*. \tag{4.48}$$

This inequality is also intuitive because the leader would only hope to gain from eliciting a different-than-expected response. Next, we deal with this cases.

**"Far"-from-expected responses**

We collect the set of commitments that (if observed fully) would elicit a response far away from the actual expected response. Formally, we denote $\mathcal{R}_{\mathsf{far}} := \cup_{j \in \mathcal{K}_{\mathsf{far}}} \mathcal{R}_j$. Now, we wish to bound the term

$$T_{far} := \sup_{\mathbf{x} \in \Delta_m} \sum_{j \in \mathcal{K}_{\mathsf{far}}} T_j(\mathbf{x}).$$

By definition, we have $\mathrm{cl}(\mathcal{R}_{j^*}) \cap \mathrm{cl}(\mathcal{R}_{far}) = \emptyset$. Because we are considering *finite* games, i.e. $n < \infty$, there exists a constant $C > 0$ that depends solely on the parameters of the game such that

$$\inf_{\mathbf{x} \in \mathcal{R}_{j^*}, \mathbf{x}' \in \mathcal{R}_{far}} D(\mathbf{x}' \parallel \mathbf{x}) \geq C. \tag{4.49}$$

Geometrically, Figure 4.17 shows this separation between the expected-response-region and any far-from-expected-response-region.

To understand the probability of eliciting such responses, we invoke a classical result from large-deviations theory, Sanov's theorem [229]. The upper bound part of the theorem is restated here as a lemma and with appropriate notation.

**Lemma 4.11.12.** *Let $I_1, I_2, \ldots, I_N$ be i.i.d $\sim \mathbf{x}$ for any $\mathbf{x} \in \Delta_m$ and $\widehat{\mathbf{X}}_N$ denote the empirical estimate. Then, for any region $\mathcal{R} \subseteq \Delta_m$, we have*

$$\Pr\left[\widehat{\mathbf{X}}_N \in \mathcal{R}\right] \leq (N+1)^m 2^{-N \inf_{\mathbf{x}' \in \mathcal{R}} D(\mathbf{x}' \parallel \mathbf{x})}. \tag{4.50}$$

Combining equations (4.50) and (4.49), we therefore get

$$\begin{aligned} T_{far} &\leq \left[\sup_{\mathbf{x} \in \Delta_m} \sum_{j \in \mathcal{K}_{\mathsf{far}}} p_j(\mathbf{x})\right] f_{max} && (4.51)\\ &\leq \left[(N+1)^m 2^{-N \inf_{\mathbf{x} \in \mathcal{R}_{j^*}, \mathbf{x}' \in \mathcal{R}_{far}} D(\mathbf{x}' \parallel \mathbf{x})}\right] f_{max} && (4.52)\\ &\leq (N+1)^m 2^{-NC} f_{max} && (4.53)\\ &= C(N+1)^m \exp\{-NC\} f_{max}. && (4.54) \end{aligned}$$

The rationale for calling these responses *far-from-expected* is now clear: there is a minimum constant separation in terms of the KL-divergence from the expected best response, and so the probability of realizing these responses decreases exponentially with $N$.

Dealing with the adjacent-to-expected responses is more delicate. We turn to this case next.

**Adjacent-to-expected responses**

Consider the set of adjacent-to-expected response $\mathcal{K}^*_{\mathsf{aug}}$. We wish to bound the term $\sum_{j\in\mathcal{K}^*_{\mathsf{aug}}} T_j(\mathbf{x})$. It turns out that we can no longer control the probability that one of these responses is elicited for all choices of $\mathbf{x} \in \mathcal{R}_{j^*}$ – this is because the commitment $\mathbf{x}$ could be chosen arbitrarily close to a boundary of its expected-response-region. However, we can bound the ensuing payoff as a function of how close the commitment is to a boundary. This notion of closeness is defined in terms of the $\ell_1$-norm below.

**Definition 4.11.13.** *For a commitment $\mathbf{x} \in \mathcal{R}_{j^*}$ and a particular adjacent response $j \in \mathcal{K}^*_{\mathsf{aug}}$, we define its minimum distance to the boundary by*

$$\delta_1(\mathbf{x}; j) := \inf_{\mathbf{x}'\in cl(\mathcal{R}_j)} \|\mathbf{x} - \mathbf{x}'\|_1.$$

First, we use this notion to bound the maximum possible payoff that could be elicited.

**Lemma 4.11.14.** *For any commitment $\mathbf{x} \in \mathcal{R}_{j^*}$, we have*

$$T_j(\mathbf{x}) \leq p_j(\mathbf{x}) \left[f^*_\infty + f_{max}\delta_1(\mathbf{x}; j)\right].$$

*Proof.* Let $\widetilde{\mathbf{x}} \in \arg\min_{\mathbf{x}'\in \mathrm{cl}(\mathcal{R}_j)} \|\mathbf{x}-\mathbf{x}'\|_1$. (Note that the minimum exists because we've taken the closure of the region.) Using Holder's inequality, we have

$$\begin{aligned}\langle \mathbf{a}_j, \mathbf{x} - \widetilde{\mathbf{x}}\rangle &\leq \|\mathbf{a}_j\|_\infty \|\mathbf{x} - \widetilde{\mathbf{x}}\|_1 \\ &\leq f_{max}\delta_1(\mathbf{x}; j).\end{aligned}$$

For every $j \in \mathcal{K}^*_{aug}$ we have

$$\begin{aligned}\langle \mathbf{a}_j, \mathbf{x}\rangle &\leq \langle \mathbf{a}_j, \widetilde{\mathbf{x}}\rangle + f_{max}\delta_1(\mathbf{x}; j) \\ &\leq f_\infty(\widetilde{\mathbf{x}}) + f_{max}\delta_1(\mathbf{x}; j) \\ &\leq f^*_\infty + f_{max}\delta_1(\mathbf{x}; j).\end{aligned}$$

where we are crucially using the fact that $\widetilde{\mathbf{x}}$ lies on the boundary and the tie-breaking assumption, to tie its payoff to the function $f_\infty(.)$. Substituting the above bound into the definition of $T_j(\mathbf{x})$ completes the proof. □

Lemma 4.11.14 is important because it limits the potential of leader gain from eliciting an adjacent follower response, even if she is able to do this with high probability, i.e. by committing very close to a boundary. Figure 4.17 clearly illustrates this for a $2 \times 2$ game: here, the leader might wish to elicit different-than-expected response 2 with high probability. However, to do this she would have to commit close to the boundary between regions expecting responses 1 and 2, resulting in her payoff being close to an objective function value of $f_\infty(.)$ (in the figure, depicted as the optimum payoff $f^*_\infty$). For a general $m \times n$ game, the picture stays the same.

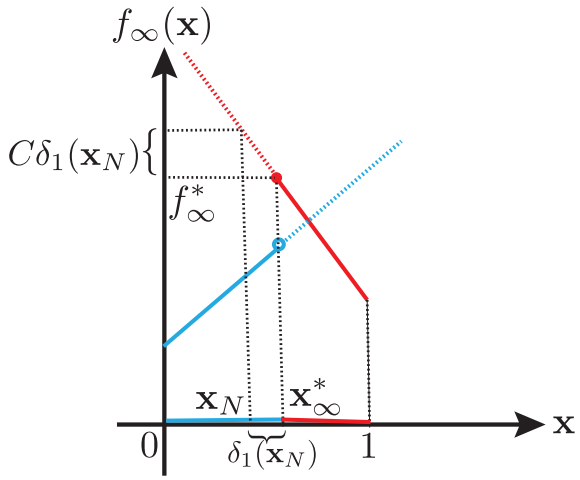



Figure 4.17: Illustration showing the potential gain in payoff obtainable by eliciting a different-than-expected response for a $2 \times 2$ game. Figure from [90].

Since the quantity $\delta_1(\mathbf{x})$ can take values anywhere in the interval $[0, 2]$ (by the triangle inequality), we will still want to control the quantity $p_j(\mathbf{x})$ for large enough values of $\delta$. We will again use Devroye's lemma (Lemma 4.11.9) for tail bounding the total variation between the empirical estimate of a distribution and a true distribution. Recall that the condition required for it to be applied was $\delta \geq \sqrt{\frac{20m}{N}}$.

It is natural to further divide the set $\mathcal{K}^*_{\mathsf{aug}}$ into two subsets, defined by the commitment $\mathbf{x}$.

$$\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x}) := \{j \in \mathcal{K}^*_{\mathsf{aug}} : \delta_1(\mathbf{x}) \leq \sqrt{\frac{20m}{N}}\}$$
$$\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x}) := \{j \in \mathcal{K}^*_{\mathsf{aug}} : \delta_1(\mathbf{x}) > \sqrt{\frac{20m}{N}}\}.$$

Let's consider these subsets one-by-one. First, we use Lemma 4.11.14 and the definition

of the subset $\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})$ to get

$$\begin{aligned}\sum_{j\in\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})} T_j(\mathbf{x}) &= \sum_{j\in\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})} p_j(x)\langle \mathbf{a}_j,\, \mathbf{x}\rangle \\ &\le \sum_{j\in\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})} p_j(x)\left[f^*_\infty + f_{max}\delta_1(\mathbf{x};j)\right] \\ &\le \sum_{j\in\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})} p_j(x)\left[f^*_\infty + f_{max}\sqrt{\frac{20m}{N}}\right] \\ &\le \left[\sum_{j\in\mathcal{K}^*_{\mathsf{aug},1}(\mathbf{x})} p_j(x)\right] f^*_\infty + f_{max}\sqrt{\frac{20m}{N}}. \end{aligned} \tag{4.55}$$

Next, we consider the term $\sum_{j\in\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})} T_j(\mathbf{x})$. We state and prove the following lemma.

**Lemma 4.11.15.** *For any commitment $\mathbf{x}\in\Delta_m$, we have*

$$\sum_{j\in\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})} T_j(\mathbf{x}) \le \left[\sum_{j\in\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})} p_j(\mathbf{x})\right] f^*_\infty + 3|\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})| f_{max}\sqrt{\frac{20m}{N}}. \tag{4.56}$$

*Proof.* Consider any $j\in\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})$. Now note that by the definition of $\delta_1(\mathbf{x};j)$, we can denote the open $\ell_1$ ball with center $\mathbf{x}$ and radius $\delta_1(\mathbf{x})$ by $B_1(\mathbf{x};\delta_1(\mathbf{x}))$. By the definition of $\delta_1(\mathbf{x};j)$, it follows that $B_1(\mathbf{x};\delta_1(\mathbf{x};j))\cap\mathcal{R}_j=\emptyset$. Therefore, we have

$$\begin{aligned} p_j(\mathbf{x}) &= \Pr\left[\widehat{\mathbf{X}}_N\in\mathcal{R}_j\right] \\ &\le \Pr\left[\widehat{\mathbf{X}}_N\notin B_1(\mathbf{x};\delta_1(\mathbf{x};j))\right] \\ &= \Pr\left[\|\widehat{\mathbf{X}}_N-\mathbf{x}\|_1\ge\delta_1(\mathbf{x};j)\right] \\ &\le 3\exp\{-\frac{N\delta_1(\mathbf{x})^2}{25}\} \end{aligned}$$

where we used Lemma 4.11.9 in the last inequality since we have $\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})$, we have $\delta_1(\mathbf{x})\ge\sqrt{\frac{20m}{N}}$.

Combining this with Lemma 4.11.14, we then have

$$T_j(\mathbf{x}) \le p_j(\mathbf{x})f^*_\infty + f_{max}3\delta_1(\mathbf{x};j)\exp\{-\frac{N\delta_1(\mathbf{x};j)^2}{25}\}.$$

Next, it is easy to verify that the function $g_2(\delta) = \delta \exp\{-\frac{N\delta^2}{25}\}$ is decreasing in $\delta$ over the domain $\delta \geq \sqrt{\frac{20m}{N}}$ for all $m \geq 1$. This tells us that

$$\begin{aligned} 3\delta_1(\mathbf{x}; j) \exp\{-\frac{N\delta_1(\mathbf{x}; j)^2}{25}\} &\leq 3\sqrt{\frac{20m}{N}} \exp\{-\frac{4m}{5}\} \\ &\leq 3\sqrt{\frac{20m}{N}} \end{aligned}$$

and so we have

$$T_j(\mathbf{x}) \leq p_j(\mathbf{x}) f^*_\infty + 3 f_{max} \sqrt{\frac{20m}{N}}. \tag{4.57}$$

Summing over all $j \in \mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})$ and substituting Equation (4.57) then proves the lemma. □

**Putting it all together**

Combining Equations (4.46), (4.51), (4.55) and (4.56) into Equation (4.44), we have

$$\begin{aligned} f_N(\mathbf{x}) &= \sum_{j=1}^{n} T_j(\mathbf{x}) \\ &\leq p_{j^*}(\mathbf{x}) f^*_\infty + C(N+1)^m \exp\{-NC\} f_{max} + \left[ \sum_{j \in \mathcal{K}^*_{aug}} p_j(\mathbf{x}) \right] f^*_\infty \\ &\quad + (3|\mathcal{K}^*_{\mathsf{aug},2}(\mathbf{x})| + 1) f_{max} \sqrt{\frac{20m}{N}} \\ &\leq f^*_\infty + C(N+1)^m \exp\{-NC\} f_{max} + 4n f_{max} \sqrt{\frac{20m}{N}} \\ &\leq f^*_\infty + Cn f_{max} \sqrt{\frac{20m}{N}} \end{aligned}$$

for some constant $C > 0$. This inequality holds for any $\mathbf{x} \in \Delta_m$. This implies that $f^*_N \leq f^*_\infty + Cn\sqrt{\frac{m}{N}}$, thus completing the proof of Theorem 4.7.5. □

# 4.12 Proofs for repeated interaction

## Proof of Proposition 4.9.2

This proof builds off the proof of Theorem 4.7.4 in Section 4.11, so the reader is recommended to read this section concurrently with that one. All notation from Section 4.11 carries over.

To refresh, we will consider the $(m-1)$-dimensional representation of the best-response-region corresponding to the Stackelberg commitment, $\mathcal{R}_{j^*}$. For shorthand, we will denote $Z := \mathcal{Z}(\mathcal{R}_{j^*}; (y^*)'_\infty)$ as defined in Equation (4.42).

For $0 < \eta < 1/2$, we propose the randomized leader rule

$$y_t = \begin{cases} y(\delta_t; y^*_\infty) \text{ for all } t > t_0 := \max\left\{\frac{20m\|B\|^2_{\mathsf{op}}}{Z^2}, m\right\} \\ y(\delta_{t_0}; y^*_\infty) \text{ otherwise,} \end{cases} \tag{4.58}$$

where $\delta_t = Z\left(\frac{m}{t}\right)^\eta$, and $y(\delta; y^*_\infty)$ is a $\delta$-deviation commitment as defined in Definition 4.11.5.

We state and prove the following elementary lemma characterizing the *expected* gap to the ideal Stackelberg payoff. The steps are similar to those in Lemma 4.11.3.

**Lemma 4.12.1.** *Consider the robust leader rule $(y_1, \ldots, y_T)$ as defined in Equation (4.58). We have*

$$f^*_\infty - f_{\mathsf{avg}}(y_1, \ldots, y_T) \le 2f_{max}\frac{1}{T}\sum_{t=1}^T \|y_t - y^*_\infty\|_1 + (f^*_\infty - f_{min})\left(\frac{1}{T}\sum_{t=1}^T \Pr\left[\widehat{\mathbf{Y}}_t \notin \mathcal{R}_{j^*}\right]\right). \tag{4.59}$$

*Proof.* We re-parameterize Equation (4.28) according to the $(m-1)$-dimensional representation of the randomized strategies to get

$$\begin{aligned} f_{T,\mathsf{avg}}(y_1, \ldots, y_T) &= \mathbb{E}_{(I_1,\ldots,I_T)\sim(y_1,\ldots,y_T)}\left[\frac{1}{T}\sum_{t=1}^T A_{I_t, j^*(\widehat{\mathbf{Y}}_{t-1})}\right] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{t-1}\left[A_{I_t, j^*(\widehat{\mathbf{Y}}_{t-1})}\right]\right] \\ &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \langle y_t,\, \mathbf{c}_{j^*(\widehat{\mathbf{Y}}_{t-1})}\rangle + d_{j^*(\widehat{\mathbf{Y}}_{t-1})}\right] \text{ (using independence across rounds)} \\ &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\langle y_t,\, \mathbf{c}_{j^*(\widehat{\mathbf{Y}}_{t-1})}\rangle + d_{j^*(\widehat{\mathbf{Y}}_{t-1})}\right] \text{ (by linearity of expectation)} \\ &= \frac{1}{T}\sum_{t=1}^T \sum_{j=1}^n \Pr\left[\widehat{\mathbf{Y}}_{t-1} \in \mathcal{R}_j\right](\langle y_t,\, \mathbf{c}_j\rangle + d_j), \end{aligned}$$

and in an argument almost identical to the proof of Lemma 4.11.3, we have, for every $t$,

$$\begin{aligned}\sum_{j=1}^{n} \Pr\left[\widehat{\mathbf{Y}}_{t-1} \in \mathcal{R}_j\right](\langle y_t, \mathbf{c}_j\rangle + d_j) &\geq \Pr\left[\widehat{\mathbf{Y}}_{t-1} \in \mathcal{R}_{j^*}\right](\langle y_t, \mathbf{c}_{j^*}\rangle + d_{j^*}) \\ &\quad + (1 - \Pr[\widehat{\mathbf{Y}}_{t-1} \in \mathcal{R}_{j^*}])f_{min} \\ &= \langle y_t, \mathbf{c}_{j^*}\rangle + d_{j^*} - \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](\langle y_t, \mathbf{c}_{j^*}\rangle + d_{j^*} - f_{min}) \\ &\geq \langle y_t, \mathbf{c}_{j^*}\rangle + d_{j^*} - \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](f^*_\infty - f_{min})\end{aligned}$$

where in the last step we have used that $y_t \in \mathcal{R}_{j^*}$, implying that $\langle y_t, \mathbf{c}_{j^*}\rangle + d_{j^*} \leq f^*_\infty$. Recall that we have $f^*_\infty = \langle y^*_\infty, \mathbf{c}_{j^*}\rangle + d_{j^*}$. Therefore, the time-averaged gap is bounded as

$$\begin{aligned}f^*_\infty - f_{\mathsf{avg}}(y_1, \ldots, y_T) &\leq \frac{1}{T}\sum_{t=1}^{T} \langle y^*_\infty - y_t, \mathbf{c}_{j^*}\rangle + \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](f^*_\infty - f_{min}) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} \|\mathbf{c}_{j^*}\|_\infty \|y^*_\infty - y_t\|_1 + \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](f^*_\infty - f_{min}) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} 2\|y^*_\infty - y_t\|_1 + \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](f^*_\infty - f_{min}) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} 2\|y^*_\infty - y_t\|_1 + \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}](f^*_\infty - f_{min}) \\ &= 2f_{max}\frac{1}{T}\sum_{t=1}^{T} \|y_t - y^*_\infty\|_1 + (f^*_\infty - f_{min})\left(\frac{1}{T}\sum_{t=1}^{T} \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}]\right),\end{aligned}$$

where the second inequality follows from Holder's inequality. This completes the proof. □

With Lemma 4.12.1 proved, it suffices to bound two quantities to complete the proof:

1. The quantity $\frac{1}{T}\sum_{t=1}^{T} \|y_t - y^*_\infty\|_1$, i.e. the time-averaged gap of the randomized leader rule to Stackelberg commitment.
2. The quantity $\frac{1}{T}\sum_{t=1}^{T} \Pr[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}]$, i.e. the time-average of the (marginal) probabilities that a different-than-expected response is elicited.

We bound these quantities one-by-one.

**Bounding the time-averaged gap to Stackelberg**

Recall that, from Equation (4.37), we have for every $t > t_0$,

$$\|y_t - y^*_\infty\|_1 = \|y(\delta_t; y^*_\infty) - y^*_\infty\|_1 \leq \delta_t,$$

and similarly for every $t \leq t_0$ we have

$$\|y_t - y^*_\infty\|_1 = \|y(\delta_{t_0}; y^*_\infty) - y^*_\infty\|_1 \leq \delta_{t_0}.$$

Thus, we get

$$\frac{1}{T}\sum_{t=1}^{T} \|y_t - y^*_\infty\|_1 = \underbrace{\frac{t_0 \delta_{t_0}}{T}}_{A} + \underbrace{\frac{\sum_{t=t_0+1}^{T} \delta_t}{T}}_{B}.$$

Pick $\delta_t := \delta_{t,\eta}$ according to Equation (4.43) (recall that $\eta < 1/2$). Note that

$$\begin{aligned} A &= \frac{t_0}{T} \cdot Z \cdot \left(\frac{m}{t_0}\right)^\eta \\ &= \frac{Z m^\eta t_0^{1-\eta}}{T} \\ &= \widetilde{\mathcal{O}}\left(\frac{1}{T}\right). \end{aligned}$$

On the other hand, we have

$$\begin{aligned} B &= \frac{Z \cdot m^\eta}{T} \sum_{t=t_0+1}^{T} \frac{1}{t^\eta} \\ &\leq \frac{Z \cdot m^\eta}{T} \sum_{t=1}^{T} \frac{1}{t^\eta} \\ &\leq \frac{2Z \cdot m^\eta}{T} \cdot T^{1-\eta}, \\ &= 2Z \cdot \left(\frac{m}{T}\right)^\eta \end{aligned}$$

where our last inequality follows from Fact 4.14.1 applied to $\eta < 1/2$. Putting these together gives us

$$\frac{1}{T}\sum_{t=1}^{T} \|y_t - y^*_\infty\|_1 \leq \frac{Z m^\eta t_0^{1-\eta}}{T} + 2Z \cdot \left(\frac{m}{T}\right)^\eta = \widetilde{\mathcal{O}}\left(\frac{1}{T^\eta}\right). \tag{4.60}$$

**Bounding the time-averaged probability of mismatched response**

Next, we turn to bounding the quantity $\frac{1}{T}\sum_{t=1}^{T} \Pr\left[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}\right]$. To do this, we consider for every $t > 0$ the quantity

$$\overline{y}_t := \frac{1}{t}\sum_{s=1}^{t} y_s.$$

Observe that, by the definition of $y_s := y(\delta_s; y^*_\infty)$ we have

$$\begin{aligned}\|\overline{y}_t - y^*_\infty\|_1 &= \|\frac{1}{t}\sum_{s=1}^{t} y(\delta_s; y^*_\infty) - y^*_\infty\|_1 \\ &= \|\frac{1}{t}\sum_{s=1}^{t} \delta_s y^*_\infty\|_1 \\ &= \overline{\delta_t}\|y^*_\infty\|_1,\end{aligned}$$

where we note that $\overline{\delta_t} := \frac{1}{t}\sum_{s=1}^{t}\delta_s$ and observe that $\delta_t \leq \overline{\delta_t} \leq \delta_{t_0}$ for $t > t_0$, and $\overline{\delta_t} = \delta_{t_0}$ for $t < t_0$. We define the Dikin ellipsoid as defined in Equation (4.41) with the choice of $\delta = \overline{\delta_t}$, i.e. the ellipsoid $\mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1)$.

Observe that for all values of $t \geq 1$, under the above condition, we have $\mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1) \subset \mathcal{R}_{j^*}$. Thus, for any $t \geq t_0$ we have

$$\Pr\left[\widehat{\mathbf{Y}}_{t-1} \notin \mathcal{R}_{j^*}\right] \leq \Pr\left[\widehat{\mathbf{Y}}_{t-1} \notin \mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1)\right],$$

and it suffices to bound the right hand side. Recall that we have samples $I_t \sim y_t$ and these are drawn independently. To this end, we state a more general form of Devroye's lemma that uses independent, but not identically distributed samples.

**Lemma 4.12.2** ( [235]). *Let $\widehat{\mathbf{Y}}_t$ be the empirical distribution of $t$ samples drawn independently according to $I_s \sim y_s$ and distributions $y_s \in \Delta_{m-1}$ for all $s = 1, 2, \ldots, t$. Then, as long as $\delta \geq \sqrt{\frac{20m}{t}}$, we have*

$$\Pr\left[\|\widehat{\mathbf{Y}}_t - \overline{y}_t\|_1 \geq \delta\right] \leq 3\exp\left\{-\frac{t\delta^2}{25}\right\}.$$

We apply this argument for all $t > t_0$. We note from Lemma 4.11.7 that

$$\widehat{\mathbf{Y}}_t \notin \mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1) \implies \|B(\widehat{\mathbf{Y}}_t - y(\overline{\delta}_t; y^*_\infty))\|_2 > \overline{\delta}_t,$$

and thus, we have

$$\begin{aligned}\Pr\left[\widehat{\mathbf{Y}}_t \notin \mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1)\right] &= \Pr\left[\|B(\widehat{\mathbf{Y}}_t - y(\overline{\delta}_t; y^*_\infty))\|_2 > \overline{\delta}_t\right] \\ &\overset{\text{(i)}}{\leq} \Pr\left[\|B\|_{op}\|\widehat{\mathbf{Y}}_t - y(\overline{\delta}_t; y^*_\infty))\|_2 > \overline{\delta}_t\right] \\ &\overset{\text{(ii)}}{\leq} \Pr\left[\|B\|_{op}\|\widehat{\mathbf{Y}}_t - y(\overline{\delta}_t; y^*_\infty))\|_1 > \overline{\delta}_t\right] \\ &= \Pr\left[\|\widehat{\mathbf{Y}}_N - y(\overline{\delta}_t; y^*_\infty))\|_1 > \overline{\delta}_t/\|B\|_{op}\right]\end{aligned}$$

where inequality (i) uses the definition of the operator norm and inequality (ii) uses the fact that $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1$ for any finite-dimensional vector $\mathbf{v}$. Applying Lemma 4.12.2 directly then gives us

$$\begin{aligned}\Pr\left[\widehat{\mathbf{Y}}_t \notin \mathbb{B}_{B,\mathbf{c},y(\overline{\delta}_t;y^*_\infty)}(1)\right] &\leq 3\exp\left\{-\frac{t\overline{\delta}_t^2}{25\|\!|B|\!\|_{\mathsf{op}}^2}\right\} \\ &\overset{\text{(i)}}{\leq} 3\exp\left\{-\frac{t\delta_t^2}{25\|\!|B|\!\|_{\mathsf{op}}^2}\right\} \\ &= 3\exp\left\{-\frac{m^{2\eta}\cdot\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2\cdot t^{1-2\eta}}{25\|\!|B|\!\|_{\mathsf{op}}^2}\right\},\end{aligned}$$

where inequality (i) follows from $\overline{\delta}_t \geq \delta_t$, and in the last equality we have simply substituted the definition of $\delta_t$.

Note that Lemma 4.12.2 can be applied as long as

$$\frac{\overline{\delta}_t}{\|\!|B|\!\|_{\mathsf{op}}} \geq \sqrt{\frac{20m}{t}}.$$

This last statement is true because $\overline{\delta}_t > \delta_t$ for all $t > t_0$ and we know from the proof of Theorem 4.7.4 that the statement is satisfied for $\delta_t$ for all $t > t_0$.

Putting it all together, and observing that $\Pr\left[\widehat{\mathbf{Y}}_t \notin \mathcal{R}_{j^*}\right] \leq 1$ for $t \leq t_0$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\Pr\left[\widehat{\mathbf{Y}}_t \notin \mathcal{R}_{j^*}\right] &\leq \frac{t_0}{T} + \frac{\sum_{t=t_0+1}^{T} 3\exp\left\{-\frac{m^{2\eta}\cdot\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2\cdot t^{1-2\eta}}{25\|\!|B|\!\|_{\mathsf{op}}^2}\right\}}{T} \\ &\leq \frac{t_0}{T} + \frac{\sum_{t=1}^{T} 3\exp\left\{-\frac{m^{2\eta}\cdot\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2\cdot t^{1-2\eta}}{25\|\!|B|\!\|_{\mathsf{op}}^2}\right\}}{T}\end{aligned}$$

Denote $C := \frac{m^{2\eta}\cdot\mathcal{Z}(\mathcal{R}_{j^*};(y^*)'_\infty)^2}{25\|\!|B|\!\|_{\mathsf{op}}^2}$. Then, by Fact 4.14.2, we can show that

$$\sum_{t=1}^{\infty} 3e^{-C\cdot t^{1-2\eta}} < C' < \infty,$$

i.e. is a convergent series for any $\eta < 1/2$. Thus, for some constant $C'$ that depends on $C$, we get

$$\frac{1}{T}\sum_{t=1}^{T}\Pr\left[\widehat{\mathbf{Y}}_t \notin \mathcal{R}_{j^*}\right] \leq \frac{t_0 + C'}{T}. \tag{4.61}$$

**Completing the proof**

Combining Equations (4.60) and (4.61) into Lemma 4.12.1, we have

$$f_{T,\mathsf{avg}}(y_1,\ldots,y_T) - f_\infty^* \leq 2f_{max}\cdot\left(\frac{Zm^\eta t_0^{1-\eta}}{T} + 2Z\cdot\left(\frac{m}{T}\right)^\eta\right) + (f_\infty^* - f_{min})\left(\frac{t_0+C'}{T}\right) \tag{4.62}$$

$$= \widetilde{\mathcal{O}}\left(\frac{1}{T^\eta}\right), \tag{4.63}$$

which completes the proof of Proposition 4.9.2.

□

## Proof of Theorem 4.9.5

For this proof, it is convenient to work with the following special representation of a $2\times 2$ game.

**Definition 4.12.3.** *We represent a $2\times 2$ game with the following notation:*

1. *Leader strategy is denoted by $p\in[0,1]$, the probability with which she chooses pure strategy 1.*

2. *The expected leader payoff as a function of $p$ if follower responds with strategies 1 and 2 respectively, is given by:*

$$f(p;1) = a_1p + b_1$$
$$f(p;2) = a_2p + b_2.$$

*We assume that $-1\leq a_1,b_1,a_2,b_2\leq 1$ and (without loss of generality[49]) that $a_1>0$, i.e. the function $f(p;1)$ is strictly increasing in $p$.*

3. *The follower has utility function such that his best-response function of $p$ is given by:*

$$j^*(p) = \begin{cases} 1 \text{ if } p\leq p_\infty^* \\ 2 \text{ otherwise.}\end{cases}$$

*The assumption of breaking ties in favor of the leader implies that $f(p_\infty^*;1) > f(p_\infty^*;2)$.*

4. *We assume $f(p_\infty^*;1)\geq f(p;2)$ for all $p\geq p_\infty^*$. Thus, the mixed Stackelberg commitment of the game is $p_\infty^*\in(0,1)$ with follower best response equal to 1. We denote $f_\infty^* := f(p_\infty^*;1)$.*

[49]Similar results will hold for the case where $a_1<0$ as well.

### The strictly-deception-dominant ensemble

We first consider the strictly-deception-dominant ensemble of $2 \times 2$ games that we defined in Definition 4.9.3. Recall that this game, as described by the notation in Definition 4.12.3, has the following properties:

$$a_2 \neq 0$$
$$f(0;2) = f(1;1) > f^*_\infty$$

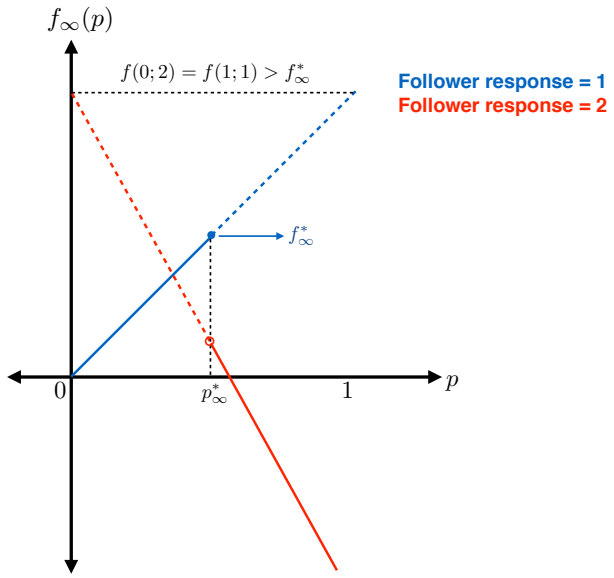



Figure 4.18: A depiction of the leader payoff function $f_\infty(p)$ for the strictly deception dominant ensemble for $2 \times 2$ leader-follower games. Figure from [90].

It is worth discussing in more detail the extra assumptions that we have made for this ensemble of $2 \times 2$ games. We assumed $f(p;1)$ to be strictly increasing in $p$ without loss of generality, so $f(1;1) > f^*_\infty$ by itself is not a new assumption. However, the further assumption that $f(0;2) > f^*_\infty$ together with $a_2 \neq 0$ is new and provides a *strict* incentive for deception by the leader. A depiction of this ensemble, which is provided in Figure 4.18, shows that the Stackelberg payoff is strictly dominated by the payoff the leader could expect if she elicited a sub-optimal response; i.e. if she was able to simultaneously play $p = 0$ with follower responding with strategy 2, or vice-versa. This property is emblematic of situations in security games, like Example 4, where the leader (defender) has a natural incentive to be unpredictable in her defense strategy – note that the leader payoff function depicted in Figure 4.11b is an example of the leader payoff function structure in Figure 4.18.

Note that the situation of a sub-optimal response, that the leader strictly desires in this ensemble, can ensue with a rational follower only *if the leader is able to deceive the follower's learning process* – hence the name "strictly-deception-dominant" for this ensemble.

As a technicality, we have additionally assumed equality of $f(0;2) = f(1;1)$ for ease of exposition[50] in the analysis of the dynamic program of Equation (4.28). Observe that, for this ensemble, $f_{max} = f(0;2) = f(1;1)$. We state the following lemma.

**Lemma 4.12.4.** *For all games in the strictly-deception-dominant $2 \times 2$ ensemble, and for any $T$:*

1. *The strategy $(p_1^*, \ldots, p_T^*)$ that maximizes the objective $f_{T,\mathsf{avg}}(p_1, \ldots, p_T)$ is unique, deterministic, and exactly achieves payoff $f_{max}$.*

2. *If the Stackelberg commitment is a rational number, i.e. $p_\infty^* = \frac{q}{r}$ for positive integers $q, r > 0$, and $q$ does not divide $r$, this strategy is periodic with period equal to $r$.*

We remark that Lemma 4.12.4 illuminates two important features specific to the strictly-deception-dominant $2 \times 2$ ensemble:

1. The optimal strategy of the naive dynamic program defined in Definition 4.9.1 is unique, and deterministic. The realized payoff can be much greater than Stackelberg due to the additional power of deception (encapsulated mathematically by $f_{max} > f_\infty^*$). However, this very property makes the payoff brittle as any deterministic strategy could be exploited by a follower using some predictive forecast according to Definition 4.8.4.

2. The optimal strategy is particularly brittle when the Stackelberg commitment $p_\infty^*$ is a rational fraction. In this case, it turns out to be periodic *with finite period*, and as we saw in Example 4 in Section 4.9, it is not only possible, but also extremely *plausible* that a follower using a very simple predictive forecast of all finitely periodic forecasts would be able to easily exploit this strategy.

Taking Lemma 4.12.4 to be true for the moment, we now concretely show that when the Stackelberg commitment $p_\infty^* = \frac{q}{r}$ is rational the optimal leader strategy $(p_1^*, \ldots, p_T^*)$ is strictly sub-optimal against a follower using a *predictive forecasting rule* according to Definition 4.8.4 with the following specifications:

1. The set of predictors $\Omega = \cup_{j=2}^K \{0,1\}^j$, i.e. the space of all $K$-periodic sequences for some finite $K > 0$.

2. A "predictable" leader sequence with parameter $\theta = (i_1, \ldots, i_l) \in \{0,1\}^l, l \leq j$ would be generated as follows:

$$I_t = i_t \text{ for } t \in [l]$$
$$I_t = I_{t-l} \text{ for all } t > l.$$

[50] More generally, if we had $f(0;2) > f_\infty^*$, the analysis would then require backward induction and become more involved. We conjecture that our results hold more generally for this case, as we have observed these properties in examples.

Observe that under the conditions of Lemma 4.12.4, the optimal leader sequence $(p_1^*, \ldots, p_T^*)$ is generated by this predictive model with $\theta^* = (p_1^*, \ldots, p_r^*)$ as the true parameter. Thus, a predictive forecast will make at most $2r$ errors, and for any $t > 2r$ the follower will respond with $J_t = j^*(p_t^*)$. Therefore, we have

$$\begin{aligned} f_{T,\mathsf{pred}}(p_1^*, \ldots, p_T^*) &\leq \frac{1}{T}\left(2r f_{max} + \sum_{t=2r+1}^{T} A_{p_t^*, j^*(p_t^*)}\right) \\ &\leq \frac{2r}{T} f_{max} + \frac{T-2r}{T} \max_{i \in \{0,1\}} f_\infty(i) \\ &< \frac{2r}{T} f_{max} + \frac{T-2r}{T} f_\infty^* \end{aligned}$$

where the last strict inequality is a consequence of the fact that any pure strategy is strictly sub-optimal in the one-shot Stackelberg game. In Section 4.9, we saw the strong extent of this sub-optimality through the $2 \times 2$ security game example where $p_\infty^* = 1/2$.

We complete this section by proving Lemma 4.12.4.

*Proof.* We specify the dynamic program for leader payoff optimization resulting from followers responding to empirical averages, as defined formally in Definition 4.9.1, in terms of the $2 \times 2$ ensemble below:

$$\max_{p^T \in [0,1]^T} f_{T,\mathsf{avg}}(p_1, \ldots, p_T) = \max_{p^T \in [0,1]^T} \frac{1}{T} \sum_{t=1}^{T} \Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right] f(p_t; 1) + \Pr\left[\widehat{P}_{t-1} > p_\infty^*\right] f(p_t; 2), \tag{4.64}$$

where we assume that the follower breaks ties[51] in favor of response 1 in the first round, i.e. $t = 1$. Remember that we have also assumed tie-breaks in favor of the leader, in the sense that $\widehat{P}_{t-1} = p_\infty^* \implies J_t = 1$.

Observe that for any $p \in [0, 1]$, $f(p; 1) \leq f_{max}$ and $f(p; 2) \leq f_{max}$, and thus the maximum payoff the leader can expect on any round is $f_{max}$. This means that in general, $f_{T,\mathsf{avg}}(p_1, \ldots, p_T) \leq f_{max}$. For the special case of the deception-dominant ensemble, we can construct a greedy strategy that achieves this upper bound with equality, and this is the unique optimal strategy. We describe this construction below: For $t = 1$, we have $p_t^* = 1$, and for every $t \geq 2$, we have

$$p_t^* = \begin{cases} 0 \text{ if } \sum_{s=1}^{t-1} p_s > (t-1)p_\infty^* \\ 1 \text{ otherwise.} \end{cases} \tag{4.65}$$

Note that because $p_t^* \in \{0, 1\}$ for every $t \geq 1$, this is a *deterministic strategy*. We need to prove that this strategy is the *unique optimal strategy*. First, to prove optimality, observe that for every round $t$, we have two cases:

[51] Thus, a unique optimum subject to tie-breaking on the first round.

1. We have $\sum_{s=1}^{t-1} p_s > (t-1)p^*_\infty$, in which case the follower responds with strategy 2 and, according to Equation (4.65), the leader will play $p^*_t = 0$. The leader payoff in these rounds is equal to $f(0;2) = f_{max}$.
2. We have $\sum_{s=1}^{t-1} p_s \leq (t-1)p^*_\infty$ (recall that ties are broken in favor of leader), in which case the follower responds with strategy 1 and, according to Equation (4.65), the leader will play $p^*_t = 1$. The leader payoff in these rounds is equal to $f(1;1) = f_{max}$.

Thus, this strategy achieves *exactly* $f_{T,\mathsf{avg}}(p^*_1, \ldots, p^*_T) = f_{max}$.

Second, to prove uniqueness, recall that our ensemble has $a_1, a_2 \neq 0$. For any distinct strategy $(p_1, \ldots, p_T)$, consider the first round $t_0$ at which $p_{t_0} \neq p^*_{t_0}$. There would again be two cases corresponding to the follower response:

1. We have $\sum_{s=1}^{t_0-1} p^*_s > (t_0-1)p^*_\infty$, in which case the follower responds with strategy 2. Note that the optimal strategy would then be $p^*_{t_0} = 0$. The payoff of the alternative strategy in this round will be $f(p_{t_0};2) < f(0;2) = f_{max}$.
2. We have $\sum_{s=1}^{t-1} p_s \leq (t-1)p^*_\infty$, in which case the follower responds with strategy 1. Note that the optimal strategy would then be $p^*_{t_0} = 1$. The payoff of the alternative strategy in this round will be $f(p_{t_0};1) < f(1;1) = f_{max}$.

In both cases, the payoff in round $t_0$ is strictly less than $f_{max}$, and thus we would have

$$\begin{aligned}
f_{T,\mathsf{avg}}(p_1, \ldots, p_T) &= \frac{1}{T}\Big( \Pr\left[\widehat{P}_{t_0-1} \leq p^*_\infty\right] f(p_{t_0};1) + \Pr\left[\widehat{P}_{t_0-1} > p^*_\infty\right] f(p_{t_0};2) \\
&+ \sum_{t \neq t_0} \Pr\left[\widehat{P}_{t-1} \leq p^*_\infty\right] f(p_t;1) + \Pr\left[\widehat{P}_{t-1} > p^*_\infty\right] f(p_t;2)\Big) \\
&\leq \frac{1}{T}\left(\Pr\left[\widehat{P}_{t_0-1} \leq p^*_\infty\right] f(p_{t_0};1) + \Pr\left[\widehat{P}_{t_0-1} > p^*_\infty\right] f(p_{t_0};2) + (T-1)f_{max}\right) \\
&< \frac{1}{T}\left(f_{max} + (T-1)f_{max}\right) = f_{max},
\end{aligned}$$

showing that the payoff of any alternate strategy is strictly sub-optimal. This proves uniqueness of the greedy construction in Equation (4.65).

Finally, we need to prove the second statement, i.e. when the Stackelberg commitment is equal to $p^*_\infty = \frac{q}{r}$, the greedy construction is periodic with period $r$. To do this, we unravel the expression in Equation (4.65) to provide explicit expressions for the optimal strategy. To do this, we characterize the optimal strategy for the first $r$ steps. In particular, we have the following lemma, whose proof we defer to Section 4.12.

**Lemma 4.12.5.** *For the greedy construction in Equation* (4.65)*, we have* $\widehat{P}_r = \frac{q}{r}$.

We use this lemma to prove periodicity with period $k$. This is equivalent to showing that for every $k \in [r]$ and for all integers $h \geq 1$, we have

$$p^*_{hr+k} = p^*_k \text{ for all } k \in [r] \text{ and } h \in \mathbb{N}. \tag{4.66}$$

We prove this by a two-step induction argument. First, we prove Equation (4.66) for $h = 1$, i.e.

$$p^*_{r+k} = p^*_k \text{ for all } k \in [r]. \tag{4.67}$$

The base case, i.e. $k = 1$ is true because by Lemma 4.12.5, we have $\widehat{P}_r = \frac{q}{r}$ and thus by Equation (4.65) we get $p^*_{r+1} = 1 = p^*_1$. Let Equation (4.67) be true for all $k \in [l]$, where $l \geq 2$. Then, for $k = l + 1$, we have $\widehat{P}_{r+l} = \frac{q+\sum_{k=1}^{l} p^*_{r+l}}{r+l} = \frac{q+\sum_{k=1}^{l} p^*_l}{r+l}$. Recall that $\widehat{P}_l = \frac{\sum_{k=1}^{l} p^*_l}{l}$, and so $\widehat{P}_{r+l} = \frac{q+l\widehat{P}_l}{r+l}$. Then, a simple calculation yields

$$\begin{aligned} \widehat{P}_{r+l} \leq \frac{q}{r} &\iff \frac{q + l\widehat{P}_l}{r+l} \leq \frac{q}{r} \\ \iff qr + rl\widehat{P}_l &\leq qr + ql \\ \iff \widehat{P}_l &\leq \frac{q}{r}. \end{aligned}$$

Thus, we have proved that $\widehat{P}_{r+l} \leq \frac{q}{r} \iff \widehat{P}_l \leq \frac{q}{r}$, which implies by Equation (4.65) that $p^*_{r+l+1} = p^*_{l+1}$. This completes the first induction argument and shows that Equation (4.66) is true for $h = 1$.

A second induction argument over $h \geq 1$, which we omit for brevity, completes the proof of periodicity. □

**One-response-obviously-dominant ensemble**

We now consider the *one-response-obviously-dominant* ensemble which was defined in Definition 4.9.4. Recall that this game, as described by the notation in Definition 4.12.3, has the following property:

$$f(0; 1) > f(p; 2) \text{ for all } p \in [0, 1]. \tag{4.68}$$

Essentially, this property means that eliciting the follower response 1 is an *obviously dominant* strategy for the leader: her expected payoff is strictly higher in the worst case over her mixed strategy if the follower responds 1 (corresponding to $p = 0$), than the best case over her mixed strategy (over all $p$) if the follower responds 2. This is clearly seen in the depiction of the leader payoff function for this ensemble, in Figure 4.19.

This situation is emblematic of building up persuasion power, as in Example 5 – note that the leader payoff function depicted in Figure 4.11a is an example of the leader payoff function structure in Figure 4.19. In this example, the leader is wholly incentivized to elicit the followers to respond with the desired pure strategy 1. In the following lemma, we show that all optimal strategies for the leader for the dynamic program over naive followers involve eliciting the response 1 on all rounds *deterministically*.

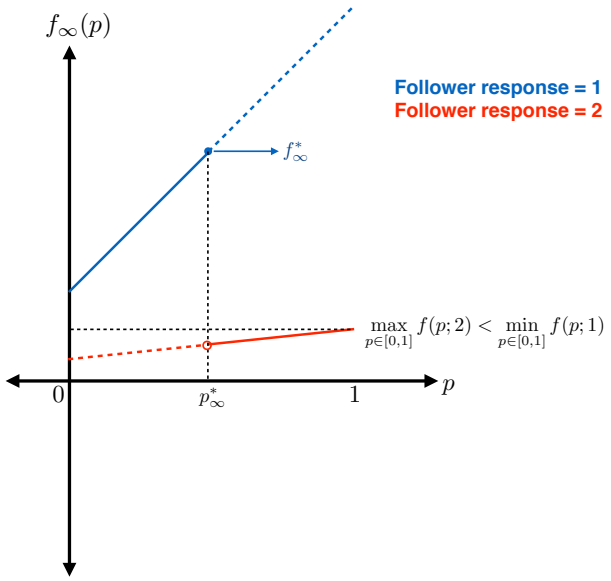



Figure 4.19: A depiction of the leader payoff function $f_\infty(p)$ for the one-response-obviously-dominant ensemble for $2 \times 2$ leader-follower games. Figure from [90].

**Lemma 4.12.6.** *Assume that $p^*_\infty = \frac{q}{r}$ for positive integers $q < r$, and $q$ does not divide $r$. Let $T = Lr$ for some positive integer $L > 0$. Then, for all games in the one-response-obviously-dominant $2 \times 2$ ensemble, all strategies $(p^*_1, \ldots, p^*_T)$ that maximize the objective $f_{T,\mathsf{avg}}(p_1, \ldots, p_T)$ satisfy the following two properties:*

1. $\frac{1}{T}\sum_{t=1}^{T} p^*_t = \widehat{P}_T = p^*_\infty$.
2. $\widehat{P}_t \leq p^*_\infty$ *with probability equal to* 1 *for all* $t = 1, 2, \ldots, T-1$.

We make two conclusions from this lemma:

1. The optimal payoff is $f^*_\infty$, i.e. precisely the ideal Stackelberg payoff.
2. All strategies that achieve the optimal payoff are *deterministic*, i.e. randomization is strictly sub-optimal.

*Proof.* We starting by noting that if a strategy that satisfies the two properties exists, then it achieves payoff equal to $f^*_\infty$. This is because under property 2, the follower always responds with pure strategy 1, yielding

$$\begin{aligned} f_{T,\mathsf{avg}}(p^*_1, \ldots, p^*_T) &= \frac{1}{T}\sum_{t=1}^{T} f(p^*_t; 1) \\ &= f_\infty(\widehat{P}_T) = f_\infty\left(p^*_\infty\right) = f^*_\infty. \end{aligned}$$

To show that this payoff is *achievable*, we construct an explicit strategy that satisfies properties 1 and 2. We can write any $t \in [T]$ as $t = lr + m$ for $l \in \{0, 1, \ldots, L-1\}$ and $m \in [r]$. Then, we pick

$$p_t^* = \begin{cases} 0 \text{ if } m < r - q \\ 1 \text{ otherwise.} \end{cases}$$

It is easy to verify from this that for any $(t-1) = lr + m$, we have

$$\widehat{P}_{t-1} \leq \widehat{P}_{(l+1)r} = \frac{(l+1)q}{(l+1)r} = \frac{q}{r} = p_\infty^*$$

and so the follower will always respond with pure strategy 1. Recalling that $T = Lr$, it is also easy to verify the first property, i.e. that $\frac{1}{T}\sum_{t=1}^{T} p_t^* = \frac{q}{r} = p_\infty^*$.

To show that this is the optimum payoff, it suffices to show that any other strategy expects payoff strictly less than $f_\infty^*$. Note that any other strategy will violate one of properties 1 and 2, so it suffices to show that any strategy violating at least one of the properties is sub-optimal. We do this in two steps.

First, we consider strategies $(p_1, \ldots, p_T)$ for which property 1 holds, but property 2 does not. Because of the obviously dominant property, observe that for any round $t$, we have

$$f_t(p_t) := \Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right] f(p_t; 1) + (1 - \Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right]) f(p_t; 2) \leq f(p_t; 1), \tag{4.69}$$

with strict inequality unless $\Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right] = 1$.

Now, if property 2 does not hold, there exists some $t_0 \in [T]$, and some $q > 0$, for which $\Pr\left[\widehat{P}_{t_0-1} > p_\infty^*\right] = q$. Under this condition, we have

$$\begin{aligned} f_{T,\mathsf{avg}}(p_1, \ldots, p_T) &= \frac{1}{T}\left(\sum_{t=1}^{T} Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right] f(p; 1) + (1 - \Pr\left[\widehat{P}_{t-1} \leq p_\infty^*\right]) f(p; 2)\right) \\ &\leq \frac{1}{T}\left(\sum_{t \neq t_0} f(p_t; 1) + (1-q) f(p_{t_0}; 1) + q f(p_{t_0}; 2)\right) \\ &< \frac{1}{T}\left(\sum_{t \neq t_0} f(p_t; 1) + f(p_{t_0}; 1)\right) \\ &= f(\widehat{P}_T; 1) = f(p_\infty^*; 1) = f_\infty^*, \end{aligned}$$

where the strict inequality follows from Equation (4.69) and because $q > 0$, and the last inequality follows because we are considering strategies for which property 1 holds. This tells us that strategies that satisfy property 1 but not property 2 are strictly sub-optimal.

Second, we show that any strategy (that could be randomized) that does not satisfy property 1 almost surely is also strictly sub-optimal as well. Proving this statement suffices

to prove the statement of the lemma, because it implies that any optimal strategy needs to satisfy both properties 1 and 2.

We consider a realization of a strategy $(p_1, \ldots, p_T) \in \{0,1\}^T$ that violates property 1, i.e. let

$$\sum_{t=1}^{T} p_t = Tp^*_\infty + K \text{ for some integer } K \geq 1.$$

We also note the following recursion on the quantity $\widehat{P}_t$:

$$\widehat{P}_t = \frac{(t-1)\widehat{P}_{t-1} + p_t}{t}.$$

Unravelling this recursion backwards, we get

$$\widehat{P}_{t-1} = \frac{t\widehat{P}_t - p_t}{t-1}, \tag{4.70}$$

Let $l_0 = \min\{l \geq 1 : \widehat{P}_{T-l} \leq p^*_\infty\}$. Applying this repeatedly and noting that $p_t \leq 1$ for all $t$, we observe that

$$\begin{aligned}\widehat{P}_{T-l} &\geq \frac{Tp^*_\infty + K - l}{T-l} \\ &> p^*_\infty \text{ if } l < \frac{K}{1-p^*_\infty}.\end{aligned}$$

Thus, we need $l_0 \geq \left\lceil \frac{K}{1-p^*_\infty} \right\rceil$. This implies that

$$\begin{aligned}\sum_{t=T-l_0}^{T} f_t(p_t) &= \sum_{t=T-l_0}^{T} f(p_t; 2) \\ &< \sum_{t=T-l_0}^{T} f(p^*_\infty; 1) = l_0 f(p^*_\infty; 1)\end{aligned}$$

where the last *strict* inequality follows by the obvious dominance property.
Further, by definition of $l_0$ we have $\widehat{P}_{T-l_0} \leq p^*_\infty$. Thus, by Equation (4.69), we have

$$\begin{aligned}\sum_{t=1}^{T-l_0} f_t(p_t) &\leq \sum_{t=1}^{T-l_0} f(p_t; 1) \\ &= (T-l_0)\widehat{P}_{T-l_0} \leq (T-l_0) f(p^*_\infty; 1).\end{aligned}$$

Combining the two equations, we have

$$f_{T,\mathsf{avg}}(p_1, \ldots, p_T) < \frac{1}{T}\left((T-l_0) f(p^*_\infty; 1) + l_0 f(p^*_\infty; 1)\right) = f^*_\infty,$$

showing that any strategy that violates property 1 is strictly sub-optimal.

□

**Proofs of auxiliary lemmas**

**Proof of Lemma 4.12.5:** Here, we prove Lemma 4.12.5. From Equation (4.65), we note the following recursion for the quantity $\widehat{P}_t$:

$$\widehat{P}_1 = 1 \tag{4.71}$$

$$\widehat{P}_t = \begin{cases} \frac{(t-1)\widehat{P}_{t-1}+1}{t} \text{ if } \widehat{P}_{t-1} \leq \frac{q}{r} \\ \frac{(t-1)\widehat{P}_{t-1}}{t} \text{ if } \widehat{P}_{t-1} > \frac{q}{r}, \end{cases} \tag{4.72}$$

and we claim that for all $t \geq 2$, we have

$$\widehat{P}_t = \frac{\left\lceil \frac{(t-1)q}{r} \right\rceil}{t}. \tag{4.73}$$

Note that, under this claim, we can write

$$\begin{aligned} \widehat{P}_r &= \frac{\left\lceil \frac{(r-1)q}{r} \right\rceil}{r} \\ &= \frac{\left\lceil q - \frac{q}{r} \right\rceil}{r} \\ &= \frac{q}{r} \end{aligned}$$

as we know that $\frac{q}{r} < 1$.

It thus suffices to prove the claim in Equation (4.73). We prove this claim by induction. Note that Equation (4.73) is trivially true for the base case $t = 2$. This is because we have $\widehat{P}_1 = 1 > \frac{q}{r}$, and so $\widehat{P}_2 = \frac{1}{2}\widehat{P}_1 = \frac{1}{2}$. Further, we have $\frac{\left\lceil \frac{(2-1)q}{r} \right\rceil}{2} = \frac{1}{2}$, and so the two quantities are equal.

Now, let Equation (4.73) be true for $t = k$, i.e. $\widehat{P}_k = \frac{\left\lceil \frac{(k-1)q}{r} \right\rceil}{k}$. We need to show that Equation (4.73) be true for $t = k+1$, i.e. $\widehat{P}_{k+1} = \frac{\left\lceil \frac{kq}{r} \right\rceil}{k+1}$. To evaluate $\widehat{P}_{k+1}$, we have two cases:

1. $\widehat{P}_k \leq \frac{q}{r}$, which implies that

$$\left\lceil \frac{(k-1)q}{r} \right\rceil \leq \frac{kq}{r}. \tag{4.74}$$

In this case, by Equation (4.71), we have $\widehat{P}_{k+1} = \frac{k\widehat{P}_k+1}{k+1}$. We note that

$$\begin{aligned} k\widehat{P}_k + 1 &= \left\lceil \frac{(k-1)q}{r} \right\rceil + 1 \\ &= \left\lceil \frac{kq}{r} \right\rceil \end{aligned}$$

where the last step follows as a consequence of Equation (4.74) and noting that $\frac{q}{r} < 1$. Substituting this into the expression for $\widehat{P}_{k+1}$ completes the argument.

2. $\widehat{P}_k > \frac{q}{r}$, which implies that

$$\left\lceil \frac{(k-1)q}{r} \right\rceil > \frac{kq}{r}. \tag{4.75}$$

In this case, by Equation (4.71), we have $\widehat{P}_{k+1} = \frac{k\widehat{P}_k}{k+1}$. We note that

$$\begin{aligned} k\widehat{P}_k &= \left\lceil \frac{(k-1)q}{r} \right\rceil \\ &= \left\lceil \frac{kq}{r} \right\rceil \end{aligned}$$

where the last step follows as a consequence of Equation (4.75) and noting that $\frac{q}{r} < 1$. Substituting this into the expression for $\widehat{P}_{k+1}$ completes the argument, and the proof of the claim.

□

## Proof of Proposition 4.9.6

We start with the condition for a universally calibrated follower forecast in Equation (4.15). For any leader rule $I_t := i_t(\mathcal{H}_{t-1})$, we have

$$\frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} \|\widehat{\mathbf{X}}_T(\mathbf{r}) - \mathbf{r}\|_1 = \frac{o(T)}{T} \text{ almost surely.}$$

Let the realization of follower forecasts be $\mathbf{r}_1, \ldots, \mathbf{r}_T$. Then, the average leader payoff is given by

$$\begin{aligned} f_{T,\mathsf{calib}}(i_1, \ldots, i_T) &= \frac{1}{T} \sum_{t=1}^{T} A_{I_t, j^*(\mathbf{r}_t)} \\ &= \frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} \langle \widehat{\mathbf{X}}_T(\mathbf{r}),\, \mathbf{a}_{j^*(\mathbf{r})} \rangle \cdot N_T(\mathbf{r}) \\ &= \underbrace{\frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \langle \mathbf{r},\, \mathbf{a}_{j^*(\mathbf{r})} \rangle}_{A} + \underbrace{\frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \langle \widehat{\mathbf{X}}_T(\mathbf{r}) - \mathbf{r},\, \mathbf{a}_{j^*(\mathbf{r})} \rangle}_{B}, \end{aligned}$$

and we now proceed to bounding both terms. Note that $B$ is the term that arises from calibration error. To see this, we apply Holder's inequality to get

$$\begin{aligned} B &\leq \frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \|\mathbf{a}_{j^*(\mathbf{r})}\|_\infty \|\widehat{\mathbf{X}}_T(\mathbf{r}) - \mathbf{r}\|_1 \\ &\leq \frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \cdot f_{max} \cdot \|\widehat{\mathbf{X}}_T(\mathbf{r}) - \mathbf{r}\|_1 \\ &= f_{max} \frac{o(T)}{T}, \end{aligned}$$

where the last step follows from the definition of a universally calibrated forecast. Thus, we have

$$\begin{aligned} f_{T,\mathsf{calib}}(i_1, \ldots, i_T) &\leq \frac{1}{T} \sum_{\mathbf{r} \in \Delta_m} N_T(\mathbf{r}) \langle \mathbf{r},\, \mathbf{a}_{j^*(\mathbf{r})} \rangle + f_{max} \frac{o(T)}{T} \\ &\leq f^*_\infty + f_{max} \frac{o(T)}{T}, \end{aligned}$$

and this completes the proof of Proposition 4.9.6. □

## 4.13 Miscellaneous calculation for persuasion example

In this section, we collect detailed calculations that were used in the Bayesian persuasion game in Example 5.

First, we consider the time-averaged payoff expected by a prosecutor who uses the randomized rule in Equation (4.21) against judges who respond according to the rule in Equation (4.23). We observed in Section 4.9 that for any defendant sequence $(\Pi_1, \ldots, \Pi_T)$ this was given by

$$\mathbb{E}\left[f_{T,\mathsf{avg}}((P_1, \ldots, P_T))\right] \geq \underbrace{\frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0] p_{g,0}(t)\right)}_{A} - \underbrace{\frac{1}{T} \sum_{t=1}^{T} \Pr\left[\widehat{P}_{g,0}(t-1) > \frac{1}{2}\right]}_{B}$$

To lower bound $A$, we substitute the definition of the prosecutor rule from Equation (4.21), and recall our notation for number of innocent defendants as well as their arrival epochs $s_1, \ldots, s_j, \ldots$ We get

$$\begin{aligned} A = \sum_{t=1}^{T} \left(\mathbb{I}[\Pi_t = 1] + \mathbb{I}[\Pi_t = 0] p_{g,0}(t)\right) &\leq \frac{T - N_T}{T} + \sum_{j=1}^{N_T} \left(\frac{1}{2} - \frac{1}{j^\eta}\right) \\ &\geq \frac{T - N_T}{T} + \frac{N_T}{2} - \frac{2}{N_T^\eta}, \end{aligned}$$

where the last inequality is a consequence of Fact 4.14.1. To upper bound $B$, noting that $\widehat{P}_{g,0}(t-1)$ only changes at epochs $s_j$, together with a special case of Lemma 4.12.2 (in Section 4.12) gives us

$$B = \frac{1}{T}\sum_{t=1}^{T} \Pr\left[\widehat{P}_{g,0}(t-1) > \frac{1}{2}\right] \leq \frac{1}{T}\sum_{j=1}^{N_T}(s_j - s_{j-1})e^{-\frac{3j^{1-2\eta}}{25}}.$$

Finally, we take further expectations of the terms $A$ and $B$ over the defendant sequence $(\Pi_1, \ldots, \Pi_T)$. We first lower bound the expectation of $A$. Noting that $\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}[N_T] = \frac{2}{3}$, we get

$$\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}[A] = \frac{1}{3} + \frac{2}{2\cdot 3} - \mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left(\frac{2}{N_T^{\eta}}\right).$$

The number of innocent defendants is close to $\frac{2}{3}T$ with high probability. Formally, the Hoeffding bound gives us $\Pr[N_T < 0.5] \leq e^{-\frac{T(0.67-0.5)^2}{8}} < e^{-\frac{0.01T}{8}}$, we get $\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left(\frac{2}{N_T^{\eta}}\right) \leq \frac{2}{(0.5T)^{\eta}} + 2\Pr[N_T < 0.5] = \frac{2}{(0.5T)^{\eta}} + 2e^{-\frac{0.01T}{8}}$. Substituting this above, we get

$$A \geq \frac{2}{3} - \frac{2}{(0.5T)^{\eta}} - 2e^{-\frac{0.01T}{8}}.$$

To upper bound the expectation of $B$, note that the unconditional distribution of $s_j - s_{j-1}$ is $\text{Geom}(\frac{2}{3})$, and we get

$$\begin{aligned} T\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}[B] &= \mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left[\sum_{j=1}^{N_T}(s_j - s_{j-1})e^{-\frac{3j^{1-2\eta}}{25}}\right] \\ &\leq \mathbb{E}_{(\Pi_1,\ldots,\Pi_T,\ldots)}\left[\sum_{j=1}^{\infty}(s_j - s_{j-1})e^{-\frac{3j^{1-2\eta}}{25}}\right] \\ &= \sum_{j=1}^{\infty}\frac{3}{2}\cdot e^{-\frac{3j^{1-2\eta}}{25}} \leq C < \infty, \end{aligned}$$

where the last step follows from Fact 4.14.2. Thus we have $\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}[B] \leq \frac{C}{T}$ for some constant $C > 0$. Putting these together, we get

$$\mathbb{E}_{(\Pi_1,\ldots,\Pi_T)}\left[\mathbb{E}\left[f_{T,\mathsf{avg}}((P_1,\ldots,P_T))\right]\right] \geq \frac{2}{3} - \frac{2}{(0.5T)^{\eta}} - e^{-\frac{0.01T^2}{8}} - \frac{C}{T}.$$

## 4.14 Mathematical facts

In this appendix, we collect miscellaneous mathematical facts that were useful for various proofs.

**Fact 4.14.1.** *For any $T > 2$ and any $0 \leq \eta < 1/2$, we have*

$$T^{1-\eta} \leq \sum_{t=1}^{T} \frac{1}{t^\eta} < \frac{T^{1-\eta}}{1-\eta} \leq 2T^{1-\eta}. \tag{4.76}$$

*Proof.* The LHS inequality is trivial as for any $\eta \geq 0$, we have $\sum_{t=1}^{T} \frac{1}{t^\eta} \geq \sum_{t=1}^{T} \frac{1}{T^\eta} = \frac{T}{\eta} = T^{1-\eta}$. For the RHS inequality, we use the Euler-Maclaurin summation. Let $f(x) = \frac{1}{x^\eta}$ and let $f^{(r)}(\cdot)$ denote the $r^{th}$ derivative of $f$ with respect to $x$. Then, for any integer $m \geq 0$ we have

$$\sum_{t=1}^{T} \frac{1}{t^\eta} = \int_1^T \frac{dx}{x^\eta} + \sum_{r=0}^{m} \frac{(-1)^{r+1} B_{r+1}}{(r+1)!} \left(f^{(r)}(T) - f^{(r)}(1)\right) + R_m \text{ , where}$$
$$R_m = \frac{(-1)^m}{(m+1)!} \int_1^T B_{m+1}(x) f^{(m+1)}(x) dx.$$

Here, $B_m$ is the $m^{th}$ Bernoulli number and $B_m(x)$ is a periodic function of period 1 that coincides with the $m^{th}$ Bernoulli polynomial on $[0, 1)$.

We apply this equality for $m = 0$. Note that $B_1 = -1/2$ and $B_1(x) = \{x\} - 1/2$, where $\{x\}$ denotes the fractional part of $x$. Then, we observe that

$$\int_1^T \frac{dx}{x^\eta} = \left[\frac{x^{1-\eta}}{1-\eta}\right]_1^T = \frac{T^{1-\eta} - 1}{1-\eta}.$$

We also have

$$\begin{aligned}\sum_{r=0}^{m} \frac{(-1)^{r+1} B_{r+1}}{(r+1)!} \left(f^{(r)}(T) - f^{(r)}(1)\right) &= -B_1 \left(\frac{1}{T^\eta} - 1\right) \\ &= -\frac{1}{2}\left(1 - \frac{1}{T^\eta}\right),\end{aligned}$$

and finally, noting that $B_1(x) = \{x\} - \frac{1}{2} \leq 1 - \frac{1}{2} = \frac{1}{2}$, we have

$$\begin{aligned}R_0 &= \int_1^T B_1(x) f^{(1)}(x) dx \\ &\leq \frac{1}{2} \int_1^T f^{(1)}(x) dx \\ &= \frac{1}{2} \left[f(x)\right]_1^T = \frac{1}{2}\left[\frac{1}{x^\eta}\right]_1^T \\ &= -\frac{1}{2}\left(1 - \frac{1}{T^\eta}\right).\end{aligned}$$

Therefore, we get

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{t^\eta} &\leq \frac{T^{1-\eta}}{1-\eta} - \frac{1}{1-\eta} - 1 + \frac{1}{T^\eta} \\ &= \frac{T^{1-\eta}}{1-\eta} - \frac{2+\eta}{1-\eta} + \frac{1}{T^\eta} \\ &\leq \frac{T^{1-\eta}}{1-\eta} - 2 + \frac{1}{T^\eta} \\ &\leq \frac{T^{1-\eta}}{1-\eta} - 1 \text{ for all } T > 1 \text{ and } \eta > 0\ .\end{aligned}$$

Noting that $\frac{T^{1-\eta}}{1-\eta} - 1 < \frac{T^{1-\eta}}{1-\eta} \leq 2T^{1-\eta}$ for all $\eta < 1/2$ completes the proof.

□

**Fact 4.14.2.** *For any $C \geq 0$ and any $q > 0$, the series $\sum_{t=1}^{\infty} e^{-C\cdot t^q}$ is convergent.*

*Proof.* It suffices to show that there exists some $C' < \infty$ such that $\sum_{t=1}^{\infty} e^{-C\cdot t^q} \leq C'$. We use the Euler-Maclaurin summation to get

$$\sum_{t=1}^{\infty} e^{-C\cdot t^q} \leq \int_{u=1}^{\infty} e^{-C\cdot u^q}.$$

Substituting $v = Cu^q$, we get $dv = Cq \cdot u^{q-1}du$. Noting that $u = \left(\frac{v}{C}\right)^{\frac{1}{q}}$, we have

$$\begin{aligned}\int_{u=1}^{\infty} e^{-C\cdot u^q} &= \int_{v=C}^{\infty} e^{-v}\frac{C^{\frac{q-1}{q}}dv}{Cq\cdot v^{\frac{q-1}{q}}} \\ &= \frac{C''}{q}\int_{v=C}^{\infty} e^{-v}v^{\frac{1}{q}-1}dv \\ &\leq \frac{C''}{q}\int_{v=0}^{\infty} e^{-v}v^{\frac{1}{q}-1}dv \\ &= \frac{C''}{q}\Gamma\left(\frac{1}{q}\right)\end{aligned}$$

where the last step follows from the standard definition of the gamma function. Denoting $C' := \frac{C''}{q}\Gamma\left(\frac{1}{q}\right) < \infty$ for $q > 0$ completes the proof. □

# Chapter 5

# Two-sided, no-regret learning

In Chapter 4, we studied candidates for *natural rules* for a two-player repeated game with one-sided learning, i.e. only one of the players was learning from the other. We saw that the designated *follower* (the player doing the learning) would follow an adaptive online learning strategy, and the designated *leader* (the player who was responding to the learning) would follow a remarkably simple randomized rule (details in Proposition 4.9.2), according to which she would choose her strategies independently across rounds. A direct consequence is that the leader and follower's day-to-day behavior approaches Stackelberg equilibrium at the rate $\mathcal{O}(1/\sqrt{\text{number of rounds}})$.

In this chapter, we ask what the corresponding natural rules might look like for situations in which both agents are learning from one another. We uncover some surprising properties that arise when the ubiquitous no-regret learning strategies are used. While the *time-average* of the strategies is classically known to converge to various equilibrium concepts in simultaneous game theory (the Nash equilibrium for zero-sum games, and the set of correlated equilibria for non-zero-sum games), the *day-to-day behavior* is starting to be investigated in more recent research. In what follows, we will show that no-regret learning strategies could result in chaotic day-to-day behavior when deployed against one another. This is in sharp contrast to the one-sided setting and raises several interesting questions for what may constitute natural two-sided learning dynamics.

## 5.1 Introduction

The mixed strategy Nash equilibrium (NE) is one of the oldest solution concepts central to game theory. A finer understanding of how the NE arises as an outcome of learning behavior in a repeated game setting remains a somewhat elusive goal as well as an active area of research. Classical research in economics [236, 237] (see also [238]) as well as some recent work in computer science [211] has taught us that when both the players in a two-player zero-sum game use strategies based on no-regret learning dynamics [80, 239], then the time-average of

their strategies will converge[1] (almost surely) to a Nash equilibrium [132, 211]. However, the convergence of the *time-averaged* mixed actions to a NE does not necessarily imply that the *day-to-day behavior* of these players converges. That is, the sequence of the mixed strategies used by the players need not converge. In the asymptotic sense, the quantity that is of interest is the tuple of the limiting mixed strategies of both players, also referred to as the last-iterate in recent literature [242]. The following surprising property of the last-iterate was discovered recently by Bailey and Piliouras [243]: When the players in a two-player zero-sum game compete against each other with the popular multiplicative weights update with certain learning rates, which constitutes a popular no-regret algorithm, then their resulting mixed strategies drift away from any interior NE — in fact, they drift towards the boundary of the strategy space[2]. This intriguing result is derived in an environment where players can play what we call *telepathic strategies*, i.e. player 1 can observe the exact mixed strategies used by player 2, and vice versa.

The natural question that arises is whether this *last-iterate divergence* is a specific property of this family of algorithms in particular or a fundamental consequence of the property of the no-regret property itself. This chapter provides substantial evidence that it is the latter, by proving lack of last-iterate convergence for a broad class of generic, asymptotically optimal no-regret algorithms. Here, we study the traditional repeated game setting in which players can only observe the realizations of the opponent's mixed strategies; thus, the strategies cannot be telepathic. In this non-telepathic scenario, we show that the ensuing stochasticity in realizations is one of the critical ingredients (in addition to others) underlying the last-iterate divergence. Our results suggest that no-regret learning strategies possess certain intrinsic properties by which the two notions—no-regret and convergence of the limiting mixed strategies—could inherently conflict with one another.

**Our contributions:** We consider the setup of a repeated $2 \times 2$ zero-sum game, i.e. a two player, zero-sum game repeatedly played infinitely many times at steps $t = 1, 2, \ldots$, where both the players can play one of two pure strategies. The repeated game strategy for a player outlines the rule by which she picks her mixed action at step $t$ based on the history up to and including step $(t-1)$. We will make three natural assumptions on each player's repeated game strategy, that are ubiquitous to several popular learning dynamics:

1. We assume that a player's strategy is *self-agnostic*, i.e. it does not use the actual realizations of her own mixed actions to update her strategy. In other words, the player picks her mixed action at step $t$ only based on the action realizations of the other player up to, and including, step $(t-1)$.

[1] A related line of research considers strategies based on internal no-regret [240] or calibrated forecasting [241] and show that the sub-sequential limits of the empirical average of the action play converge to a correlated equilibrium [85, 208, 209].

[2] Bailey and Piliouras [243] consider a deterministic dynamic system comprised of the pair of mixed actions evolving according to the multiplicative weights updates on the time-average of the opponents mixed actions. In contrast, we are interested in the stochastic dynamic system of the pair of mixed actions whose evolution depends on the past realizations of the mixed actions.

2. We assume the player's strategy to be an *optimal no-regret* strategy, that is, she has an expected average regret of $\mathcal{O}(t^{1/2})$ irrespective of the strategy employed by the other player[3]. See Definition 5.2.3 for formal definitions of no-regret algorithms, optimal or otherwise. Note that no-regret is purely a property of the strategy of each individual player, unlike the solution concept of Nash equilibrium which intrinsically depends on the behavior of all the players.

3. Finally, we assume the player's strategy to be *mean-based*, i.e. the player uses only the empirical average of the actions of the other player at step $(t-1)$ as a sufficient statistic to decide her mixed action at step $t$. In general, the player is aware of the step $t$, and we accordingly allow her rule that maps empirical averages to mixed strategies to depend[4] on the step $t$.

Observe that the most popular learning dynamics, such as *Online-Mirror-Descent* and *Follow-the-Regularized Leader* strategies, satisfy all three of these assumptions. The natural question arises whether the resulting game play would be stable, that is, would the mixed actions of the players to converge to an equilibrium? We answer this question in the negative for the set of the games which only have purely mixed NE, designated as *competitive games* by Calvo [244]. Recently, Phade and Anantharam [245] showed that all competitive games have a *unique* strictly mixed NE which is also the unique correlated equilibrium of the game. We denote this unique NE by the tuple $(p^*, q^*)$, where $0 < p^*, q^* < 1$ denote the equilibrium strategies of playing action 1 by players 1 and 2 respectively. In Theorem 5.3.3, we prove the following statement for any competitive game (described here informally):

*If player* 1 *uses a self-agnostic, mean-based repeated game strategy that satisfies no-regret with the optimal regret rate, and player* 2 *plays the mixed action* $q^*$ *at all steps, then with a constant positive probability the mixed actions of player* 1 *do not converge to* $p^*$.

Theorem 5.3.3 suggests that the mixed strategies will also diverge when *both* players are using self-agnostic, mean-based, optimal no-regret strategies. We conjecture this last-iterate divergence in Conjecture 5.3.4. The intuition for this conjecture holding is a *proof-by-contradiction*: if the pair of mixed strategies for both players were to converge almost surely, then, player 2's mixed strategies would converge almost surely. Thus, the mixed strategies of player 2 would remain arbitrarily close to his NE strategy $q^*$ with arbitrarily high probability after enough steps. As per Theorem 5.3.3, the ensuing stochasticity in player 2's realizations would then necessitate player 1 to diverge. For technical reasons related to possible implicit dependencies across the realizations of both players, this intuition is difficult to formalize in the stochastic dynamical system ensuing from both players making updates on their strategies. However, we do show that player 1's mixed strategies will necessarily diverge when she is facing any *fixed-convergent* player 2, i.e. player 2 uses any *fixed* sequence

[3]A self-agnostic repeated game strategy has the following useful property: if it is a no-regret strategy with respect to an oblivious opponent, then it is a no-regret strategy with respect to a non-oblivious opponent. See Chapter 4, [81] for definitions of oblivious and non-oblivious opponents.

[4]In fact, this flexibility is in a certain sense *required* to design a mean-based, no-regret strategy.

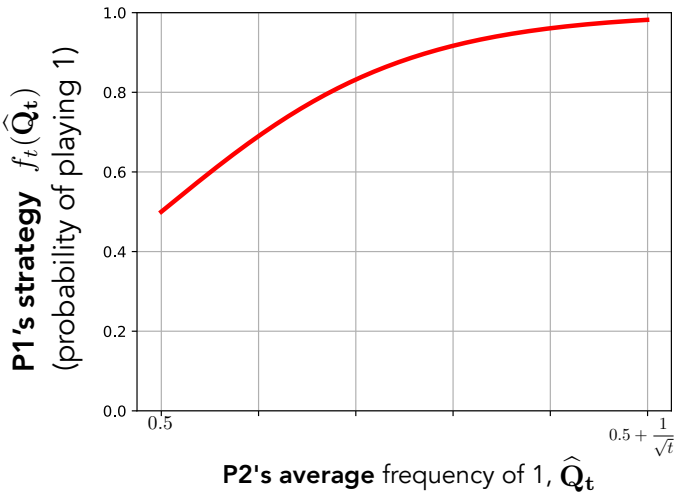



Figure 5.1: Depiction of sensitivity of the no-regret strategy multiplicative weights, $f_t(\widehat{\boldsymbol{Q}}_t)$, as a function of $\widehat{\boldsymbol{Q}}_t$ for $t = 10^6$. Observe that a deviation of $1/\sqrt{t}$ from $1/2$ causes a *constant* deviation in the function value. Figure from [92].

of mixed strategies $\{q_t\}_{t\geq 1}$ that converges to the NE $q^*$. The precise statement for this divergence is contained in Theorem 5.3.7. Moreover, Section 5.2 provides ample evidence for even stronger forms of divergence than we have conjectured in the stochastic dynamical system where both players update their strategies as a function of the past.

**Our techniques in a nutshell:** Consider a self-agnostic repeated game strategy for a player. Such a strategy is completely characterized by the mappings $\{f_t\}$ at each step from the opponents action history up to that step to the mixed action played by the player in the next step. Our first observation is concerning a fundamental *fluctuation-sensitivity* of the mappings of any self-agnostic, no-regret algorithm. We show (in Proposition 5.3.1) that any self-agnostic, optimal no-regret repeated game strategy used by player 1uses mappings at certain rounds, $f_t(\cdot)$, that deviate by at least a constant value, say $\delta$, from the NE strategy $p^*$ when player 2 deviates from his NE strategy on-average by an infinitesimal factor on the order of $t^{-1/2}$. Moreover, this deviation property is shown to be present *infinitely often*, i.e. for a sub-sequence $\{t_k\}_{k\geq 1}$. A depiction of this strong sensitivity is in Figure 5.1 for the example of the multiplicative weights algorithm in the matching pennies; we show that it is fundamental to any no-regret strategy for any competitive game.

Now, our second observation is that if player 2 is playing the mixed NE $q^*$ at all his steps, then the time-averages of his realized actions will fluctuate on the order of $t^{-1/2}$ infinitely often as well. In fact, this happens with a fixed positive probability. Figure 5.2 depicts[5] the

[5]This figure was inspired by Dean P. Foster's illustration of the law of the iterated logarithm: `https://en.wikipedia.org/wiki/Law_of_the_iterated_logarithm`

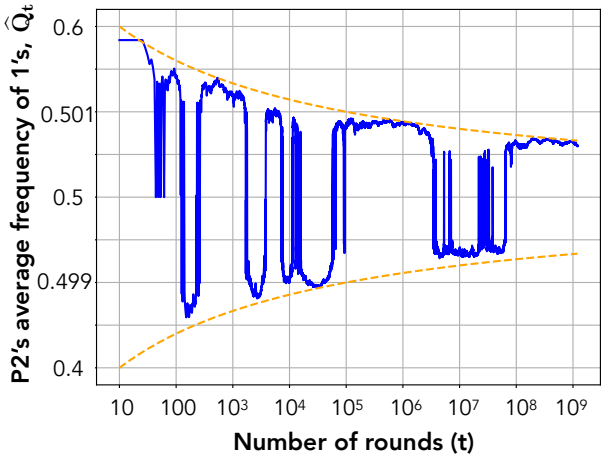



Figure 5.2: Depiction of constant fluctuations of player 2's time-averages, i.e. $\widehat{\boldsymbol{Q}}_{\boldsymbol{t}}$, as a function of $t$, when player 2 plays $\boldsymbol{Q_t}$ i.i.d $\sim q^*$. Figure from [92], inspired by Dean P. Foster's illustration of the law of the iterated logarithm.

recurring fluctuations on the order of $t^{-1/2}$ of player 2's time-averages for a typical realization of player 2's strategies $q^* = 1/2$ in the matching pennies game, and should remind the reader of the fluctuations of a symmetric random walk.

Putting these observations together, we see that optimal no-regret strategies *need* to be sensitive to the fluctuations of player 2 infinitely often. Moreover, these fluctuations happen infinitely often as well, as a consequence of the stochasticity in player 2's realizations. These constitute the key phenomena that underlie last-iterate divergence for such a broad class of optimal no-regret strategies.

**Related work:** While the evolution of the *time-averages* of players' strategies as a consequence of multiple players using no-regret dynamics has been an active topic of study for several decades [85, 132, 208, 209, 211, 236–238], the properties of the limiting mixed strategies, or the last-iterates, have only been examined more recently. This topic has also seen substantial attention in the related setup of *min-max optimization* [246–250], where the primary goal is to attain a pure-strategy NE of a game with a continuous-pure-strategy set through the use of first-order optimization algorithms, e.g. gradient descent-ascent. This problem has been primarily studied in the *deterministic setting*, corresponding to the aforementioned telepathic dynamics in the game-theoretic setup. Recently, Daskalakis and Panageas [242] show that a modification on the mean-based strategy of multiplicative weights that incorporates *recency bias* succeeds in last-iterate convergence in the game-theoretic setup with telepathic dynamics. This type of recency bias, commonly called *optimism*, has also been shown to successfully converge in min-max optimization when applied to the gradient descent/ascent algorithms [246–250]. Moreover, optimistic algorithms have other notable properties, such as leading to faster convergence rates of the time-average in zero-sum as well as non-zero-sum

games [251, 252]. However, we show in Section 5.3 that when stochastic, realization-based feedback is considered, optimistic variants on mean-based strategies *do not* resolve the last-iterate divergence issue. In other words, the phenomena we outlined above, that lead to last-iterate divergence in the traditional repeated game setting, manifest in recency-bias-based strategies as well. This illustrates that the issues of last-iterate divergence run deeper in the traditional repeated-game setting. We briefly discuss alternative (non-constructive) strategies that could satisfy the last-iterate-convergent property in Section 5.4 — these strategies are *not* no-regret, but satisfy a weaker property of "smoothly calibrated forecasting".

## 5.2 Setup

We consider a simple setting of a $2 \times 2$ zero-sum game in which both the players have two pure strategies, namely, *action* 0 *and action* 1. The payoffs for player 1 are given by the following matrix:

$$G := \begin{bmatrix} G(0,0) & G(0,1) \\ G(1,0) & G(1,1) \end{bmatrix}.$$

Thus for $i, j \in \{0, 1\}$, if player 1 plays action $i$ and player 2 plays action $j$, then the payoff to player 1 is given by $G(i, j)$, the $(i, j)$-th element of the matrix $G$, and the payoff of player 2 is given by its negation, viz. $-G(i, j)$.

We consider all entries of the payoff matrix to be finite and bounded, i.e. $|G(i, j)| \leq B$ for some finite $B > 0$. We denote by the indicator random variables $\boldsymbol{I}$ and $\boldsymbol{J}$, the mixed strategies of player 1 and player 2, respectively. We follow the convention of denoting random variables by the bold versions of their corresponding deterministic variables. Let $p := \mathbb{E}[\boldsymbol{I}]$ and $q := \mathbb{E}[\boldsymbol{J}]$ be the probabilities with which the two players play action 1, respectively. In general, since we will be considering *uncoupled dynamics*, the randomness in the strategies $\boldsymbol{I}$ and $\boldsymbol{J}$ will be independent. Therefore, the expected payoff for player 1 corresponding to the choice of mixed strategies $(p, q)$ is given by:

$$G(p, q) := (1 - p)(1 - q)G(0, 0) + (1 - p)qG(0, 1) + p(1 - q)G(1, 0) + pqG(1, 1),$$

and the expected payoff for player 2 is given by $-G(p, q)$.

We now consider a repeated game setting where $\{\boldsymbol{I_t}\}_{t\geq 1}$ and $\{\boldsymbol{J_t}\}_{t\geq 1}$ are the action sequences of the two players. Let $\boldsymbol{(I)^t} := \{\boldsymbol{I_s}\}_{s=1}^{t}$ and $\boldsymbol{(J)^t} := \{\boldsymbol{J_s}\}_{s=1}^{t}$. Let the empirical averages of the actions of the two players be given by $\widehat{\boldsymbol{P}}_{\boldsymbol{t}} := \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{I_t}$, and $\widehat{\boldsymbol{Q}}_{\boldsymbol{t}} := \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{J_t}$ respectively. General repeated game strategies for player 1 and player 2 are given by sequences of functions $\{f_t\}_{t\geq 1}$ and $\{g_t\}_{t\geq 1}$, where $f_t, g_t : \{0, 1\}^{2(t-1)} \to [0, 1], \forall t$ map the history up to step $t$, i.e. $(\boldsymbol{(I)^{t-1}}, \boldsymbol{(J)^{t-1}})$ to mixed strategies given by $\boldsymbol{P_t} = f_t(\boldsymbol{(I)^{t-1}}, \boldsymbol{(J)^{t-1}})$ and $Q_t = f_t(\boldsymbol{(I)^{t-1}}, \boldsymbol{(J)^{t-1}})$ for players 1 and 2 respectively. We will refer to these functions $f_t$ and $g_t$ as the strategy functions for players 1 and 2 respectively at step $t$. Critically, observe that we are not allowing for *telepathy* in the updates, i.e. the history used by player 1 at

step $t$ does not include $\{\boldsymbol{Q_s}\}_{s=1}^{t-1}$, and the history used by player 2 at step $t$ does not include $\{\boldsymbol{P_s}\}_{s=1}^{t-1}$. This is in agreement with the information structure of the traditional repeated game environment. Over and above this traditional information structure, we will make some further assumptions on the repeated game strategies as detailed below.

We first assume that player 1's repeated game strategy is *self-agnostic*, as defined below.

**Definition 5.2.1.** *We say that a repeated game strategy for player* 1 *is* self-agnostic *if player* 1 *uses only the action sequence of player* 2 *to decide her mixed strategy* $P_t$ *at step* $t$*. With an abuse of notation, the strategy function can be replaced by a function* $f_t : \{0,1\}^{t-1} \to [0,1]$*, such that the mixed strategy for player* 1 *at step* $t$ *is given by* $\boldsymbol{P_t} = f_t(\boldsymbol{(J)^{t-1}})$*.*

Note that player 1 is actually aware of her mixed strategies $\boldsymbol{P_1}, \boldsymbol{P_2}, \ldots, \boldsymbol{P_{t-1}}$ at step $t$ since she is aware of her strategy functions $f_1, f_2, \ldots, f_{t-1}$ and, for $1 \leq s \leq t-1$, $\boldsymbol{P_s}$ can be determined from $f_s$ and $\boldsymbol{(J)^{s-1}}$. Thus, by self-agnostic, we only mean that player 1 is agnostic to the actual realizations of her actions up to that step in order to chose the next mixed strategy.

From the point of view of player 1, we now define *no-regret* strategies, as well as *uniformly no-regret* strategies against an *oblivious opponent*. The former is precisely Hannan's classical definition of consistency [80], while the latter is a strictly stronger condition, requiring an effective non-asymptotic guarantee on regret. We will use the stronger uniform-no-regret condition to derive our results. Also note that, in accordance with the self-agnostic assumption we made on strategies, both definitions of no-regret are the weaker notion of *external*[6].

**Definition 5.2.2.** *A self-agnostic repeated game strategy strategy* $\{f_t\}_{t\geq 1}$ *is said to be no-regret if*

$$\limsup_{T\to\infty} \frac{1}{T}\left[\max_{i\in\{0,1\}} \sum_{t=1}^{T} G(i, J_t) - \sum_{t=1}^{T} G(f_t((J)^{t-1}), J_t)\right] \leq 0,$$

*for all opponent sequences* $\{J_t\}_{t\geq 1}$*.*

Note that we have defined the no-regret property in expectation over all the randomization in the strategy (only of player 1). Sometimes, even stronger definitions of no-regret are used [] that require the no-regret property to hold *almost surely* for all realizations of player 1's strategy. However, this definition will suffice for our purposes.

**Definition 5.2.3.** *A self-agnostic repeated game strategy* $\{f_t\}_{t\geq 1}$ *is said to be uniformly no-regret if*

$$\limsup_{T\to\infty} \max_{\{J_t\}_{t=1}^{T}} \frac{1}{T}\left[\max_{i\in\{0,1\}} \sum_{t=1}^{T} G(i, J_t) - \sum_{t=1}^{T} G(f_t((J)^{t-1}), J_t)\right] \leq 0.$$

[6]The stronger notion of *internal* no-regret can be derived from a given external no-regret algorithm, but the self-agnostic property is then violated. Moreover, internal no-regret algorithms are primarily of interest in non-zero-sum game environments, and a full study of their dynamics is of substantial interest, but outside the scope of this chapter.

*In particular, such a strategy is said to satisfy a no-regret rate of* $(r, c)$ *if*

$$\limsup_{T\to\infty} \max_{\{J_t\}_{t=1}^T} \frac{1}{T^r}\left[\max_{i\in\{0,1\}} \sum_{t=1}^{T} G(i, J_t) - \sum_{t=1}^{T} G(f_t((J)^{t-1}), J_t)\right] \leq c.$$

Observe that all uniformly no-regret strategies are also no-regret. The converse need not hold — however, algorithms that are no-regret but not uniformly no-regret are necessarily contrived examples, and unlikely to be practically used.

The following properties of uniformly no-regret strategies $\{f_t\}$ can easily be verified:

1. If a strategy $\{f_t\}$ satisfies a no-regret rate of $(r, c)$, then it satisfies a no-regret rate of $(r, c')$ for all $c' \geq c$.

2. If a strategy $\{f_t\}$ satisfies a no-regret rate of $(r, c)$, then it satisfies a no-regret rate of $(r', 0)$ for all $r' > r$.

3. Since the payoff matrix entries $G(\cdot, \cdot)$ are bounded, any uniformly-no-regret strategy $\{f_t\}$ satisfies a no-regret rate of $(r, 0)$ with $r > 1$.

4. Conversely, if a strategy $\{f_t\}$ satisfies a no-regret rate of $(r, c)$ where $r < 1$ or $r = 1$ and $c = 0$, then it is a uniformly no-regret strategy.

It is well-known (e.g. see [81, Chapter 3]) that, for any finite constant $0 < c < \infty$, the best possible no-regret rate is $r = 1/2$. Moreover, several commonly used algorithms, like multiplicative/exponential weights/Follow-the-Perturbed-Leader, typically match the optimal no-regret rate for appropriately chosen constant $c$.

Finally, we define the *mean-based* property of any repeated game strategy as below.

**Definition 5.2.4.** *A repeated game strategy for player* 1 *is* mean-based *if player* 1 *uses only the empirical averages of player* 2 *as a sufficient statistic to determine her mixed strategy* $\boldsymbol{P_t}$ *at round* $t$*. In this case, with an abuse of notation, the strategy function can be replaced by a function* $f_t : [0, 1] \to [0, 1]$*, such that the mixed strategy for player* 1 *at step* $t$ *is given by* $\boldsymbol{P_t} = f_t(\widehat{\boldsymbol{Q}}_{\boldsymbol{t-1}})$*.*

For example, all algorithms in the popular Online-Mirror-Descent framework satisfy the mean-based property while also being self-agnostic and uniformly no-regret. We also discuss variants on the mean-based property that incorporate a recency bias in Section 5.3.

In addition to the above assumptions on the repeated game strategy used by player 1, we need to make a few further assumptions on the payoff structure of the $2 \times 2$ game, which are detailed below:

**Assumption 5.2.5.** *We assume that the game matrix* $G$ *is chosen such that the unique Nash equilibrium* $(p^*, q^*)$ *is strictly in the interior, i.e.* $0 < p^*, q^* < 1$*;* $G(0, j) \neq G(1, j)$ *for at least one* $j \in \{0, 1\}$*, and* $G(i, 0) \neq G(i, 1)$ *for at least one* $i \in \{0, 1\}$*. The reasons for each of these assumptions are detailed below:*

1. *The assumption of* $0 < p^* < 1$ *implies that player* 1 *is agnostic between choosing between action* 0 *and* 1 *when player* 2 *is playing his Nash equilibrium strategy,* $q^*$*. This is useful for establishing the existence of a randomized opponent sequence against which all choices of strategy would result in the same expected payoff for player* 1*.*

2. *The assumption of* $G(0, j) \neq G(1, j)$ *for at least one value of* $j \in \{0, 1\}$ *and* $G(i, 0) \neq G(i, 1)$ *for at least one value of* $i \in \{0, 1\}$ *is necessary to make the definition of no-regret non-trivial for both players — for example, if* $G(0, 0) = G(1, 0)$ *and* $G(0, 1) = G(1, 1)$*, any strategy deployed by player* 1 *would trivially satisfy no-regret as she is always agnostic between actions* 0 *and* 1*, regardless of what player* 2 *chooses to do.*

3. *Finally, the assumption of* $0 < q^* < 1$ *means that player* 2 *actually has stochasticity in his realizations, which is fundamentally important to show the lack of last-iterate convergence for all no-regret algorithms.*

Observe that a consequence of having $0 < q^* < 1$ be the unique NE strategy for player 2 is as follows: if $G(0, 0) > G(0, 1)$, we need $G(1, 0) < G(1, 1)$. In particular, the inequalities between pure strategies 0 and 1 need to be in opposite directions for player 2, as otherwise one of the strategies would be strictly dominated and $q^* \in (0, 1)$ could not be an equilibrium strategy. Similarly, we will use the convention that $G(0, 1) < G(1, 1)$ and $G(0, 0) > G(1, 0)$ with the same reasoning applied to player 1. Note that the direction of these inequalities is without loss of generality (as we can just re-index the pure strategies); the relative direction of the inequalities is what is crucial.

Under the above assumptions, we can define $R^* := G(p^*, q^*) = G(1, q^*) = G(0, q^*)$ (where the chain of equalities follows because $p^*$ is in the interior and by the definition of a Nash equilibrium).

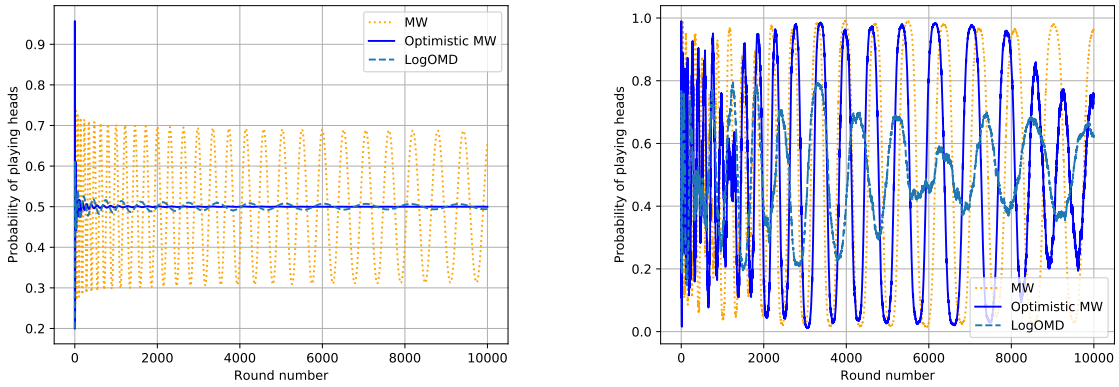
In the next section, we use the above assumptions to prove, in a series of steps, that the limiting mixed strategies diverge when arising as an outcome of both players using self-agnostic, mean-based, optimal no regret learning. Before stating and proving our theoretical results, we provide compelling empirical evidence for the phenomenon of last-iterate divergence.

## Empirical evidence for last-iterate divergence

To illustrate the last-iterate divergence that arises, we evaluate three commonly used no-regret algorithms:

1. The standard multiplicative weights update, which is known to lead to last-iterate divergence even in the deterministic setting [243].

2. The optimistic multiplicative weights update, which converges in the last-iterate in the deterministic setting [242].




(a) Deterministic case.

(b) Stochastic case (typical realization).

Figure 5.3: Evolution of the iterates of multiplicative weights in the matching pennies game (for player 1) when the optimal-no-regret rate $r = 1/2$ is used. Figures from [92].

3. The online mirror descent algorithm with the log function as regularizer, often called "log barrier" [253]. This regularizer has been successfully used to establish robustness of fast *time-average* convergence guarantees in limited-information feedback settings [166], and is thus naturally interesting to evaluate.

Note that all of the above algorithms fall under the online-mirror-descent framework, and employ fixed learning rates $\{\eta_t\}_{t\geq 1}$. The (asymptotic) rate of decay of $\eta_t$ with $t$ dictates the no-regret rate in all three cases: if $\eta_t = 1/t^r$, the no-regret rate is equal to $(r, c)$ for a suitable positive constant $c$. We will evaluate these algorithms with two learning rate choices: $\eta_t = 1/\sqrt{t}$ (optimal), and $\eta_t = 1/t^{0.7}$ (sub-optimal rate $r = 0.7$).

Furthermore, we will consider the simplest $2 \times 2$ game: the matching pennies game, for which $G(0,0) = G(1,1) = 1$ and $G(0,1) = G(1,0) = 0$ (without loss of generality, player 1 is the player who wants the coins to match). Note that the unique mixed-strategy equilibrium of this game is $p^* = q^* = 1/2$. We will plot the evolution of the mixed strategies of player 1 with time — since the matching pennies game is symmetric, player 2 has similar behavior.

Figure 5.3 studies the optimal-no-regret case, and shows the striking difference between the evolution of the mixed strategies when the players use opponents' mixtures (the deterministic case) as opposed to their realizations (the stochastic case, studied in this chapter). Notably, in Figure 5.3a, we see that while multiplicative weights converges to a limit cycle, optimistic multiplicative weights converges quite quickly. The third algorithm, log-barrier online mirror descent, also diverges in the last iterate, but the amplitude of the cycles is much smaller[7] than for multiplicative weights. On the other hand, we see in Figure 5.3b that all three of these algorithms diverge in the last iterate. In fact, they are very rarely close to the equilibrium strategy $p^* = 0.5$! All in all, Figure 5.3b provides strong empirical

[7]This likely reflects the increased entropy of the strategies used in the log-barrier algorithm.

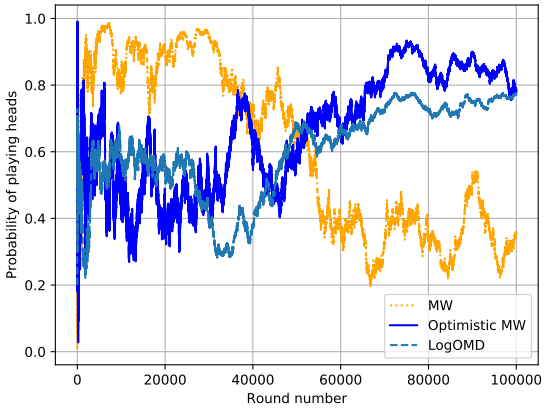


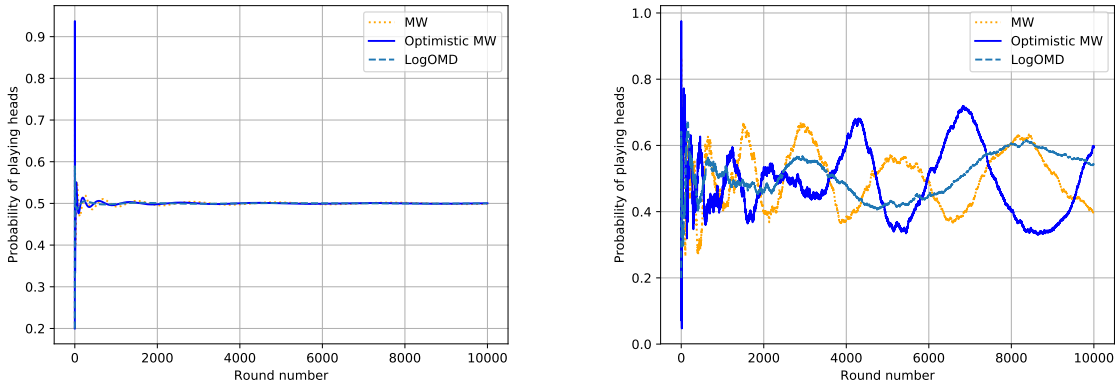

Figure 5.4: Evolution of the iterates of multiplicative weights in the matching pennies game (for player 1) when the optimal-no-regret rate $r = 1/2$ is used, and player 2 is already at NE. Figure from [92].

evidence for last-iterate-divergence for optimal-no-regret algorithms (Conjecture 5.3.4), and also corroborates our evidence (Theorem 5.3.9) that introducing optimism into no-regret strategies does not fix the last-iterate divergence issue.

It is also worth examining the differential impact on player 1 as a result of player 2 using an optimal no-regret strategy, as opposed to player 2 playing his fixed NE strategy. In the case of the matching pennies game, the latter case corresponds to player 2 playing $q^* = 0.5$ at every round. Figure 5.4 depicts the evolution of the mixed strategies of player 1 in this latter case. Comparing the evolution to Figure 5.3b, it is evident that the mixed strategies continue to diverge. While the "period" of limiting cycles, if any, seems to be larger in the fixed-strategy case, the amplitude of divergence is similar in both cases. Thus, the simplifying case that we studied in Theorem 5.3.3 successfully identifies at least some of the phenomena underlying last-iterate divergence.

Finally, while our forthcoming theory only hold for optimal-no-regret algorithms, Figure 5.5 provides preliminary empirical evidence even using sub-optimal no-regret algorithms may not resolve the last-iterate divergence issue. The smaller amplitude of the limit cycles makes them less visible, but Figure 5.5a shows that in the scenario of telepathic dynamics, multiplicative weights and log-barrier online-mirror-descent with sub-optimal learning rates continue to result in divergence of the last iterate, and optimistic multiplicative weights continues to result in convergence. More importantly, Figure 5.5b shows that all three algorithms continue to lead to last-iterate divergence under realization-based feedback, although the amplitude of the divergence does appear to be reduced.

In sum, the above simulations provide compelling evidence for a fundamental tension between the property of no-regret and the property of last-iterate convergence. We now proceed to show mathematically that self-agnostic, mean-based, no-regret strategies imply




(a) Deterministic case.

(b) Stochastic case (typical realization).

Figure 5.5: Evolution of the iterates of multiplicative weights in the matching pennies game (for player 1) when the sub-optimal-no-regret rate $r = 0.7$ is used. Figures from [92].

last-iterate divergence through a series of steps.

## 5.3 Main results

We start by proving a condition that any self-agnostic, optimal no-regret strategy necessarily satisfies (regardless of whether it is mean-based or not).

### A necessary condition for optimal no-regret algorithms

Figure 5.6 highlights, on linear scale, an interesting sensitivity of a popular no-regret algorithm to a deviation away from NE in the matching pennies game. The algorithm that is considered is multiplicative weights with learning rate $\eta_s = 1/\sqrt{s}$ for $s = 1, \ldots, t$, played over $t = 10^6$ rounds against a "matching pennies" opponent whose time-averages deviate from his NE strategy, $1/2$ by a factor on the order of $1/\sqrt{t}$. This tiny deviation (so small that it is not even visible in the figure!) causes the iterates of player 1, i.e. $\boldsymbol{P_t}$, to deviate all the way from 0.5 to 0.9. The details of this experiment are as follows:

1. Player 1 uses the multiplicative weights algorithm with learning rate $\eta_s = 1/\sqrt{s}$. This is a mean-based strategy, and for this particular choice of learning rate the strategy functions are given by $\boldsymbol{P_s} = f_s(\widehat{\boldsymbol{Q}}_{\boldsymbol{s}}) = \frac{e^{\sqrt{s}\cdot\widehat{\boldsymbol{Q}}_{\boldsymbol{s}}}}{e^{\sqrt{s}\cdot\widehat{\boldsymbol{Q}}_{\boldsymbol{s}}}+e^{\sqrt{s}\cdot(s-\widehat{\boldsymbol{Q}}_{\boldsymbol{s}})}}$.

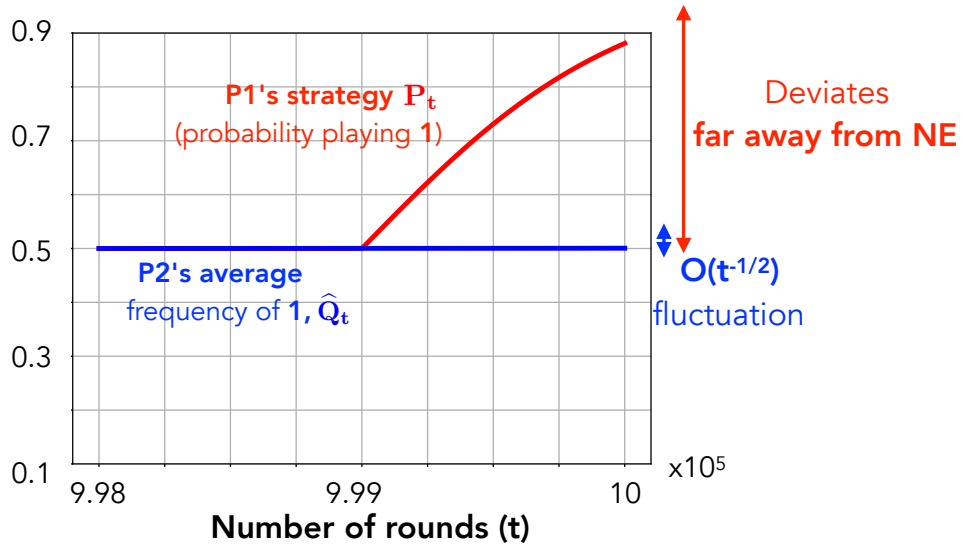



Figure 5.6: Response of player 1, who is using the multiplicative weights algorithm with learning rate $\eta = 1/\sqrt{t}$, against player 2 who is playing the sequence of alternating 0's and 1's up-to round number $s_0 := t - \sqrt{t}$, and 1 there-after until round $t$. Here, we set $t = 10^6$. The blue line plots the time-averages of player 2, and the red line plots the iterates of player 1. Note that the $y$-axis of the figure is on linear scale, so the fluctuation in the time-averages of player 2 is not visible. Figure from [92].

2. Player 2 plays the *fixed* sequence for $t$ rounds,

$$\boldsymbol{J_s} = \begin{cases} 1 \text{ if } s \leq t - \sqrt{t} \text{ and } s \text{ odd.} \\ 0 \text{ if } s \leq t - \sqrt{t} \text{ and } s \text{ even.} \\ 1 \text{ if } t - \sqrt{t} < s \leq t. \end{cases}$$

This leads to the following evolution of the time-averages of player 2:

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{s}} = \begin{cases} \frac{1}{2} \text{ if } s \leq t - \sqrt{t} \text{ and } s \text{ even.} \\ \frac{1}{2} + \frac{1}{s} \text{ if } s \leq t - \sqrt{t} \text{ and } s \text{ odd.} \\ \frac{1}{2} + \frac{s - (t - \sqrt{t})}{2s} \text{ if } t - \sqrt{t} < s \leq t. \end{cases}$$

Note in particular that $\widehat{\boldsymbol{Q}}_{\boldsymbol{t}} = 1/2 + 1/2\sqrt{t}$.

Putting these details together, we see that Figure 5.6 illustrates the strong sensitivity of the multiplicative weights algorithm to the $\mathcal{O}(1/\sqrt{t})$ fluctuation that is caused by player 2

deviating from his sequence of alternating 0's and 1's close to the final round $t$. Remarkably, we can generalize the above idea and show that, in a certain sense, this strong sensitivity to a deviation away from NE is a property of *any* no-regret algorithm! The following proposition describes precisely the nature of this *fluctuation-sensitivity*.

**Proposition 5.3.1.** *Assume that the $2 \times 2$ game satisfies all the conditions in Assumption 5.2.5. Then, for any self-agnostic repeated game strategy $\{f_t\}_{t\geq 1}$ that is uniformly no-regret with a rate of $(r, c)$ (where $1/2 \leq r < 1$) , and any $0 < \delta < (1 - p^*)/3$, there exists a positive constant $\alpha$ and an infinite sequence of integer tuples $\{(t_k, s_k)\}_{k\geq 1}$ such that*

$$0 < s_k \leq \alpha(t_k)^r, \text{ for all } k \geq 1, \tag{5.1}$$

*and*

$$\mathbb{E}\left[f_{t_k}\left((\boldsymbol{J'(k)})^{\boldsymbol{t_k}}\right)\right] \geq p^* + 2\delta, \text{ for all } k \geq 1, \tag{5.2}$$

*where the expectation is over the random sequence $(\boldsymbol{J'(k)})^{\boldsymbol{t_k}} := \{\boldsymbol{J'_s(k)}\}_{s=1}^{t_k}$ defined as below:*

$$\boldsymbol{J'_s(k)} = \begin{cases} \boldsymbol{J^*_s} \text{ i.i.d. } \sim \text{Bernoulli}(q^*), \text{ if } 1 \leq s \leq t_k - s_k, \\ 1 \text{ otherwise.} \end{cases}$$

In particular, if the self-agnostic repeated game strategy $\{f_t\}_{t\geq 1}$ is *optimally* uniformly no-regret i.e. $r = 1/2$, then condition 5.1 would be

$$0 < s_k \leq \alpha\sqrt{t_k}, \text{ for all } k \geq 1.$$

This case is important because, even if player 2 constantly plays her equilibrium mixed strategy $q^*$, there is a non-trivial probability of player 2's empirical average deviating from $q^*$ by a number on the order of $1/\sqrt{t}$ at step $t$. The sensitivity in an optimal no-regret strategy to pick deviations of this order will allow us to show the non-convergence of player 1's mixed strategies in the subsequent Sections 5.3, 5.3 and 5.3.

*Proof.* Recall our convention in Assumption 5.2.5 was $G(0, 1) < G(1, 1)$, and so we will denote $R_1^* := G(1, 1)$ as shorthand. Observe that for any $0 \leq p < 1$, we have $G(p, 1) < R_1^*$. Consider $\{f_t\}_{t\geq 1}$ to be any self-agnostic uniformly no-regret strategy with a no-regret rate of $(r, c)$. We note that for any $t$, and any sequence $\{J_s\}_{s=1}^t$, we have

$$\max_{i\in\{0,1\}} \sum_{s=1}^{t} G(i, J_s) = t \cdot \max\{G(0, \widehat{Q}_t), G(1, \widehat{Q}_t)\}.$$

Thus, for $c' > c$, there exists a sufficiently large $t_0$ such that for all $t \geq t_0$, we have

$$\frac{1}{t^r}\left[t \cdot \max\{G(0, \widehat{Q}_t), G(1, \widehat{Q}_t)\} - \sum_{s=1}^{t} G(f_s((J)^{s-1}), J_s))\right] \leq c', \tag{5.3}$$

for any sequence $\{J_s\}_{s=1}^t$.

Now let $0 < \delta < \frac{(1-p^*)}{3} = \frac{(R_1^* - R^*)}{3(G(1,1)-G(0,1))}$, and let $\delta' := \delta(G(1,1) - G(0,1))$. Note that $0 < \delta' < \frac{(R_1^*-R^*)}{3}$. Let $\alpha := \frac{c'}{(R_1^*-R^*-3\delta')} > 0$. For any $t > t_1 := \max\{t_0, \alpha^{1/r}, (\alpha\delta'/R_1^*)^{-1/r}\}$, let $t^*(t) := t - \lfloor \alpha t^r \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. Note that since $t > \alpha^{1/r}$, we have $t^*(t) \geq 1$. Let $\{\boldsymbol{J_s^*}\}$ be an i.i.d. sequence of Bernoulli($q^*$) random variables for $1 \leq s \leq t$. Let $\widehat{\boldsymbol{Q}}_{\boldsymbol{s}}^* := \frac{1}{s}\sum_{s'=1}^s \boldsymbol{J_{s'}^*}$ denote the empirical average of this sequence at round $s$. We state the following useful lemma. Recall that $R^*$ denotes the Nash equilibrium payoff of player 1.

**Lemma 5.3.2.** *For the sequence $\{\boldsymbol{J_s^*}\}_{s\geq 1}$ defined above, and any $t \geq 1$, we have*

$$\mathbb{E}\left[\sum_{s=1}^t G(f_s((\boldsymbol{J^*})^{\boldsymbol{s-1}}), \boldsymbol{J_s^*}))\right] = tR^* \text{ and} \tag{5.4a}$$

$$\mathbb{E}\left[\sum_{s=1}^t \boldsymbol{J_s^*}\right] = tq^*. \tag{5.4b}$$

See Appendix 5.5 for the proof of this lemma. We define the sequence $\{\boldsymbol{J'_s}\}_{s=1}^t$ as specified in the statement of Proposition 5.3.1. In other words, we define $\boldsymbol{J'_s} = \boldsymbol{J_s^*}$ for $1 \leq s \leq t^*(t)$, and $\boldsymbol{J'_s} = 1$ for $t^*(t) < s \leq t$. Then, we can denote the empirical average of this sequence as $\widehat{\boldsymbol{Q}}'_{\boldsymbol{s}} := \frac{1}{s}\sum_{s'=1}^s \boldsymbol{J'_{s'}}$ for $1 \leq s \leq t$. We denote

$$M := \frac{1}{\lfloor \alpha t^r \rfloor} \cdot \mathbb{E}\left[\sum_{s=1}^{\lfloor \alpha t^r \rfloor} G\left(f_{t^*(t)+s}\left((\boldsymbol{J'})^{\boldsymbol{t^*(t)+s-1}}\right), 1\right)\right]. \tag{5.5}$$

From the definition of uniform no-regret, i.e. Equation (5.3), we have

$$\begin{aligned}
c' &\geq \frac{1}{t^r} \cdot \mathbb{E}\left[T\max\{G(0, \widehat{\boldsymbol{Q}}'_{\boldsymbol{t}}), G(1, \widehat{\boldsymbol{Q}}'_{\boldsymbol{t}})\} - \sum_{s=1}^t G(f_s((\boldsymbol{J'})^{\boldsymbol{s-1}}), \boldsymbol{J'_s}))\right] \\
&\geq \frac{1}{t^r} \cdot \mathbb{E}\left[t \cdot G(1, \widehat{\boldsymbol{Q}}'_{\boldsymbol{t}}) - \sum_{s=1}^t G(f_s((\boldsymbol{J'})^{\boldsymbol{s-1}}), \boldsymbol{J'_s}))\right] \\
&= \frac{1}{t^r} \cdot [t^*(t) \cdot R^* + \lfloor \alpha t^r \rfloor \cdot R_1^* - t^*(t) \cdot R^* - \lfloor \alpha t^r \rfloor \cdot M] \\
&\geq (R_1^* - M)\alpha - R_1^* t^{-r}.
\end{aligned}$$

Using the fact that $\alpha := \frac{c'}{(R_1^*-R^*-3\delta')}$ and $t > \left(\frac{\alpha\delta'}{R_1^*}\right)^{-1/r}$, we get

$$M \geq R_1^* - \frac{c'}{\alpha} - \frac{R_1^* \cdot t^{-r}}{\alpha} = R^* + 3\delta' - \frac{R_1^* \cdot t^{-r}}{\alpha} \geq R^* + 2\delta'.$$

Now, using linearity of expectation, and linearity of the payoff function $G(p, 1)$ in the argument $p$, we get

$$M = G(\overline{f}, 1) \text{ where}$$
$$\overline{f} := \frac{1}{\lfloor \alpha t^r \rfloor} \sum_{s=1}^{\lfloor \alpha t^r \rfloor} \mathbb{E}\left[ f_{t^*(t)+s}\left( (\boldsymbol{J'})^{\boldsymbol{t^*(t)+s-1}} \right) \right].$$

Now, since $G(1, 1) > G(0, 1)$, we note that $G(p, 1) = G(0, 1) + (G(1, 1) - G(0, 1))p$ is an increasing function in $p$ and so we get

$$\overline{f} \geq p^* + \frac{2\delta'}{G(1,1) - G(0,1)} = p^* + 2\delta,$$

from the above inequality on $M$. Thus, there exists $s(t)$ such that $1 \leq s(t) \leq \lfloor \alpha t^r \rfloor$ and

$$\mathbb{E}\left[ f_{t^*(t)+s(t)}\left( (\boldsymbol{J'})^{\boldsymbol{t^*(t)+s(t)-1}} \right) \right] \geq p^* + 2\delta. \tag{5.6}$$

To write this in the language of Equation (5.2), we observe that

$$0 < \frac{s(t)}{t^*(t) + s(t)} \leq \frac{\lfloor \alpha t^r \rfloor}{t^*(t) + \lfloor \alpha t^r \rfloor} \leq \frac{\alpha t^r}{t} = \alpha t^{r-1} \leq \alpha (t^*(t) + s(t))^{r-1},$$

and therefore we get

$$s(t) \leq \alpha (t^*(t) + s(t))^r. \tag{5.7}$$

We note that $t^*(t) \to \infty$ as $t \to \infty$, and hence we can define an infinite sequence of integer tuples $\{t_k, s_k\}_{k \geq 1}$ (where we have defined $t_k = t^*(t_1 + k) + s_k$ and $s_k = s(t_1 + k)$ as above) such that $0 < s_k \leq \alpha (t_k)^r$ and

$$\mathbb{E}\left[ f_{t_k}\left( (\boldsymbol{J'(k)})^{\boldsymbol{t_k}} \right) \right] \geq p^* + 2\delta \text{ for all } k = 1, 2, \ldots \tag{5.8}$$

This is precisely the statement in Equation (5.2), and completes the proof of Proposition 5.3.1. □

## Warm-up: Last-iterate divergence when opponent is already at equilibrium

Equation (5.2) highlights a critical property of any self-agnostic uniformly no-regret algorithm with a regret rate of $(r, c)$: it needs to be sufficiently sensitive to small perturbations on the order of $t^r$ in the opponent's strategy. We can concretize this property to show last-iterate divergence when both players use self-agnostic, *optimal no-regret strategies*, i.e. $r = 1/2$, and use mean-based repeated game strategies as detailed in Definition 5.2.4. Recall that under the mean-based assumption, player 1's strategy functions are

$$f_t((J)^{t-1}) := f_t(\widehat{\boldsymbol{Q}}_{\boldsymbol{t-1}}) \text{ for all } t \geq 1. \tag{5.9}$$

The mean-based assumption underlies the broad family of Online-Mirror-Descent algorithms that satisfy the external-no-regret property. More generally, strategies that use an appropriate mean of the past history of outcomes are among the earliest algorithms satisfying related properties like Blackwell approachability [82], internal-no-regret [240] and calibration [241]. Moreover, in reality, for our techniques to work we only need strategies to be mean-based in an approximate sense. In Section 5.3, we will show that essentially the same results hold when the players use variants of the above class of strategies that incorporate a *recency bias*.

The way we will show last-iterate-divergence for player 1 is in the style of *proof-by-contradiction*: we show that if player 2 were able to converge in the last-iterate, player 1 *must* diverge. The central idea is that the stochasticity in the realizations of player 2, itself, cause player 1 to diverge. This result in its full generality is in Section 5.3, and its proof is contained in Appendix 5.3. Here, we state and prove a warm-up result that contains all of the key ideas underlying last-iterate divergence. This considers the special case where player 2 is playing his equilibrium strategy at all steps, i.e. $\{\boldsymbol{J}_t\}_{t\geq 1}$ i.i.d $\sim$ Bernoulli($q^*$). Remarkably, we show that even this simple case necessitates the limiting mixed strategy of player 1 to diverge! (This is in stark contrast to the setting of telepathic dynamics, in which a simple algorithm like multiplicative weights would lead player 1 to converge to the equilibrium strategy $p^*$ in games like matching pennies.)

**Theorem 5.3.3.** *Assume that player 2's strategy $\{\boldsymbol{J_t}\}_{t\geq 1}$ is an i.i.d. sequence of Bernoulli($q^*$) random variables. Then, any mean-based repeated game strategy $\{f_t\}_{t\geq 1}$ that has a regret rate of $(1/2, c)$ and satisfies the empirical averages condition in Equation (5.9) causes player 1's last iterate to diverge, i.e. there exist positive constants $(\delta, \epsilon)$ such that*

$$\limsup_{t\to\infty} \mathbb{P}\left[|\boldsymbol{P_t} - p^*| \geq \delta\right] \geq \epsilon. \tag{5.10}$$

The proof of Theorem 5.3.3 constitutes an elementary application of Markov's inequality, and a change-of-measure argument on the probability mass functions of two binomial random variables.

*Proof.* We start by defining some notation pertinent to mean-based strategies. Let $\boldsymbol{Z_t} \sim$ Binomial$(t, q^*)$, for $t \geq 1$. Let $\boldsymbol{Z''_{t,s}} \sim \boldsymbol{Z_{t-s}} + s$ and $\boldsymbol{Q''_{t,s}} = \boldsymbol{Z''_{t,s}}/t$, for $0 \leq s \leq t$, $t \geq 1$. Let $\{(t_k, s_k)\}_{k\geq 1}$ be an infinite sequence and $0 < \delta < (1 - p^*)/3$ as in Proposition 5.3.1. Thus, for every $k \geq 1$, we have

$$\boldsymbol{Q''_{t_k,s_k}} \stackrel{d}{=} \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{J'_t(k)},$$

where $\stackrel{d}{=}$ denotes that the two random variables are identical in distribution, and $\boldsymbol{J}'_t(\boldsymbol{k})$ are random variables as defined in Proposition 5.3.1. We thus have

$$0 < s_k \leq \alpha(t_k)^{1/2}, \text{ for all } k \geq 1, \tag{5.11}$$

and

$$\mathbb{E}\left[f_{t_k}\left(\boldsymbol{Q}''_{\boldsymbol{t_k},\boldsymbol{s_k}}\right)\right] \geq p^* + 2\delta, \text{ for all } k \geq 1, \tag{5.12}$$

Consider the random variable $\boldsymbol{Y} := 1 - f_{t_k}(\widehat{\boldsymbol{Q}}''_{\boldsymbol{t_k},\boldsymbol{s_k}})$. Since the range of $f_t$ is always $[0,1]$, we have $\boldsymbol{Y} \geq 0$. Thus, we have $\mathbb{E}[\boldsymbol{Y}] \leq 1-(p^*+2\delta) = (1-p^*)-2\delta$. By Markov's inequality, we have

$$\mathbb{P}\left(\boldsymbol{Y} \geq (1-p^*) - \delta\right) \leq \frac{(1-p^*)-2\delta}{(1-p^*)-\delta}.$$

Thus, we get

$$\mathbb{P}\left(f_{t_k}\left(\widehat{\boldsymbol{Q}}''_{\boldsymbol{t_k},\boldsymbol{s_k}}\right) > p^* + \delta\right) \geq \epsilon_0, \tag{5.13}$$

where we define $\epsilon_0 := \delta/((1-p^*)-\delta)$. Note that $0 < \epsilon_0 < 1/2$ (because we had defined $0 < \delta < (1-p^*)/3$). By the central limit theorem, we know that

$$\lim_{t\to\infty} \mathbb{P}\left(\widehat{\boldsymbol{Q}}_{\boldsymbol{t}} > q^* + \frac{\gamma}{\sqrt{t}}\right) = 1 - \Phi\left(\frac{\gamma}{\sqrt{q^*(1-q^*)}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Since the function $\Phi(\cdot) : \mathbb{R} \to (0,1)$ in continuous, strictly increasing, $\Phi(0) = 1/2$, and $\lim_{x\to\infty}\Phi(x) = 1$, we know that there exists $\gamma_0 > 0$ such that $1 - \Phi(\gamma_0/\sqrt{q^*(1-q^*)}) = \epsilon_0/4$ (note that $\epsilon_0/4 < 1/2$). Hence there exists $T'_1(\epsilon_0) > 1$ such that

$$\mathbb{P}\left(\widehat{\boldsymbol{Q}}_{\boldsymbol{t}} > q^* + \frac{\gamma_0}{\sqrt{t}}\right) \leq (1-\phi(\gamma_0)) + \frac{\epsilon_0}{4} \leq \frac{\epsilon_0}{2},$$

for all $t \geq T'_1(\epsilon_0)$.

Now, observe that

$$t_k \widehat{\boldsymbol{Q}}''_{\boldsymbol{t_k},\boldsymbol{s_k}} \stackrel{d}{=} (t_k - s_k)\widehat{\boldsymbol{Q}}_{\boldsymbol{t_k}-\boldsymbol{s_k}} + s_k.$$

Since $t_k - s_k \to \infty$ as $k \to \infty$, there exists a $k_1 > 1$ such that

$$\mathbb{P}\left(t_k \widehat{\boldsymbol{Q}}''_{\boldsymbol{t_k},\boldsymbol{s_k}} > q^*(t_k - s_k) + s_k + \gamma_0\sqrt{t_k - s_k}\right) \leq \frac{\epsilon_0}{2},$$

for all $k \geq k_1$. Using the bound $s_k \leq \alpha\sqrt{t_k}$, we get

$$\mathbb{P}\left(\boldsymbol{Z}''_{\boldsymbol{t_k},\boldsymbol{s_k}} > q^* \cdot t_k + \beta\sqrt{t_k}\right) \leq \frac{\epsilon_0}{2}, \tag{5.14}$$

for all $k \geq k_1$, where $\beta := \alpha + \gamma_0$. From the union bound and Equations (5.13) and (5.14), we get

$$\mathbb{P}\left(f_{t_k}\left(\frac{\boldsymbol{Z''_{t_k,s_k}}}{t_k}\right) \geq p^* + \delta, \boldsymbol{Z''_{t_k,s_k}} \leq q^* \cdot t_k + \beta\sqrt{t_k}\right) \geq \frac{\epsilon_0}{2}. \tag{5.15}$$

Recall that we defined the opponent sequence $\{\boldsymbol{J_t}\}_{t\geq 1}$ to be an i.i.d. sequence of Bernoulli($q^*$) random variables. Note that $\boldsymbol{Z_t} \stackrel{d}{=} t\widehat{\boldsymbol{Q}}_{\boldsymbol{t}}$, and by definition $\boldsymbol{Z''_{t,s}} \geq s$ point-wise. We denote $\beta_0 := \beta/q^*$. Note that $q^*t + \beta\sqrt{t} = q^*(t + \beta_0\sqrt{t})$ for any $t$. Now, we show that

$$\min_{s\leq z\leq q^*(t+\beta_0\sqrt{t})} \frac{\mathbb{P}(\boldsymbol{Z_t} = z)}{\mathbb{P}(\boldsymbol{Z''_{t,s}} = z)} \geq (1+\beta_0)^{-\alpha}, \tag{5.16}$$

for all $0 < s \leq \alpha\sqrt{t}$, $t \geq 1$. Indeed, we have,

$$\begin{aligned}
\frac{\mathbb{P}(\boldsymbol{Z_t} = z)}{\mathbb{P}(\boldsymbol{Z''_{t,s}} = z)} &= \frac{\binom{t}{z}(q^*)^z(1-q^*)^{t-z}}{\binom{t-s}{z-s}(q^*)^{z-s}(1-q^*)^{t-z}} = \frac{t}{z}\cdot\frac{t-1}{z-1}\cdots\frac{t-s+1}{z-s+1}\cdot(q^*)^s \\
&\geq \left(\frac{q^*t}{z}\right)^s \geq \left(\frac{t}{t+\beta_0\sqrt{t}}\right)^{\alpha\sqrt{t}} \\
&\geq (1+\beta_0)^{-\alpha} > 0,
\end{aligned}$$

where the first inequality follows from $z \leq t$ and therefore $\frac{t-\ell}{z-\ell}$ is increasing in $\ell$, and the last inequality follows from the fact that

$$\left(\frac{t+\beta_0\sqrt{t}}{t}\right)^{\alpha\sqrt{t}} = \left(1+\frac{\beta_0}{\sqrt{t}}\right)^{\alpha\sqrt{t}} \leq (1+\beta_0)^{\alpha}.$$

(Note that the function $(1+\beta_0/x)^{\alpha x}$ is decreasing in $x$ for $x \geq 1$.)

We are now ready to complete our proof via a simple "change-of-measure" argument and the above lower bound on the ratio of the probability mass functions. From Equations (5.15) and (5.16) and the law of total probability, we get

$$\begin{aligned}
&\mathbb{P}\left(f_{t_k}\left(\frac{\boldsymbol{Z_{t_k}}}{t_k}\right) \geq p^* + \delta, \boldsymbol{Z_{t_k}} \leq q^* \cdot t_k + \beta\sqrt{t_k}\right) \\
&\geq \sum_{z=s_k}^{q^*\cdot t_k+\beta\sqrt{t_k}} \mathbb{P}\left(\boldsymbol{Z_{t_k}} = z\right) \cdot \mathbb{I}\left[f_{t_k}\left(\frac{z}{t_k}\right) \geq p^* + \delta\right] \\
&\geq (1+\beta_0)^{-\alpha} \cdot \sum_{z=s_k}^{q^*\cdot t_k+\beta\sqrt{t_k}} \mathbb{P}\left(\boldsymbol{Z''_{t_k,s_k}} = z\right) \cdot \mathbb{I}\left[f_{t_k}\left(\frac{z}{t_k}\right) \geq p^* + \delta\right] \\
&= (1+\beta_0)^{-\alpha} \cdot \mathbb{P}\left(f_{t_k}\left(\frac{\boldsymbol{Z''_{t_k}}}{t_k}\right) \geq p^* + \delta, \boldsymbol{Z''_{t_k}} \leq q^* \cdot t_k + \beta\sqrt{t_k}\right) \\
&\geq \frac{\epsilon_0}{2}(1+\beta_0)^{-\alpha},
\end{aligned}$$

and hence

$$\mathbb{P}\left(f_{t_k}\left(\frac{\boldsymbol{Z_{t_k}}}{t_k}\right) \geq p^* + \delta\right) \geq \frac{\epsilon_0}{2}(1+\beta_0)^{-\alpha}$$

for $k \geq k_1$. Since $\boldsymbol{P_{t_k}} \stackrel{d}{=} f_{t_k}(\widehat{\boldsymbol{Q}}_{\boldsymbol{t_k}})$, taking $\epsilon := (\epsilon_0/2)(1+\beta_0)^{-\alpha}$, we get

$$\mathbb{P}\left[\boldsymbol{P_{t_k}} \geq p^* + \delta\right] \geq \epsilon,$$

for all $k \geq k_1$. This implies Equation (5.10) and completes the proof of Theorem 5.3.3. □

## Last-iterate divergence when *both* players use optimal no-regret

Now, we use the intuition from the proof of Theorem 5.3.3 to conjecture last-iterate divergence when player 2 is, himself, using a no-regret algorithm.

**Conjecture 5.3.4.** *Assume that both players* 1 *and* 2 *use self-agnostic, mean-based repeated game strategies* $\{f_t\}_{t\geq 1}$ *and* $\{g_t\}_{t\geq 1}$*, respectively, that are uniformly no-regret and each have a regret rate of* $(1/2, c)$*. Then, the pair of mixed strategies of both the players* $(\boldsymbol{P_t}, \boldsymbol{Q_t})$ *diverges with a positive probability.*

Observe that it is sufficient to show divergence of the pair $(\boldsymbol{P_t}, \boldsymbol{Q_t})$ from only the equilibrium $(p^*, q^*)$ with a positive probability[8] In particular, it suffices to show that there exist positive constants $(\delta, \epsilon)$ such that

$$\limsup_{t\to\infty} \mathbb{P}\left[|\boldsymbol{P_t} - p^*| \geq \delta\right] \geq \epsilon \text{ or} \tag{5.17a}$$

$$\limsup_{t\to\infty} \mathbb{P}\left[|\boldsymbol{Q_t} - q^*| \geq \delta\right] \geq \epsilon. \tag{5.17b}$$

While we do not prove this conjecture, we provide strong evidence for it. We show that if player 2 used any a-priori *fixed* sequence of mixed strategies $\{q_t\}_{t\geq 1}$ satisfying last-iterate convergent properties, player 1's mixed strategies would necessarily diverge. The idea is that the realizations of player 2 arising from any such convergent sub-sequence are quite "similar" (in a sense we will shortly define) to the realizations that would arise if player 2 had already converged to equilibrium. We define *fixed-convergent* sequences that player 2 can follow below.

**Definition 5.3.5.** *An* fixed-convergent *strategy for player* 2*, parameterized by positive constants* $(\delta, t_0, C)$*, is any sub-sequence* $\{\boldsymbol{Q_t} := q_t\}_{t\geq 1}$ *satisfying the following two properties:*

[8]This is because if a sequence $(\boldsymbol{P_t}, \boldsymbol{Q_t})_{t\geq 1}$ did converge to some other point $(p, q) \neq (p^*, q^*)$, then the time-averages $(\widehat{\boldsymbol{P}}_{\boldsymbol{t}}, \widehat{\boldsymbol{Q}}_{\boldsymbol{t}})_{t\geq 1}$ would have to converge to $(p, q)$ as well. However, we know that the time-averages have to converge to the equilibrium $(p^*, q^*)$ almost surely; thus the event that the sequence $(\boldsymbol{P_t}, \boldsymbol{Q_t})_{t\geq 1}$ converges to some point other than the equilibrium has zero probability.

$$|q_t - q^*| \leq \delta/2 \text{ for all } t \geq t_0 \tag{5.18a}$$

$$|\overline{q}_t - q^*| \leq \frac{C}{\sqrt{t}} \text{ for all } t \geq 1, \text{ where} \tag{5.18b}$$

$$\overline{q}_t := \frac{1}{t}\sum_{s=1}^{t} q_s.$$

*We also denote the set of such fixed-convergent strategies by $\mathcal{Q}_{\delta,t_0,C}$ and denote their truncation to round $t$ by $\mathcal{Q}_{\delta,t_0,C}(t)$.*

Note that a fixed-convergent player 2 would not be adapting her repeated-game strategy in response to feedback from player 1; nevertheless, the properties of her repeated-game strategy resemble the properties of a last-iterate convergent no-regret strategy in a *marginal* sense. To see this, we first note that Equation (5.18a) is a *necessary* condition for convergence of the mixed strategies of player 2. Next, the stronger time-averaged convergence property of Equation (5.18b) arises from the following classical lemma, which uses the optimal-no-regret property of player 2's strategy to establish the evolution of $\overline{\boldsymbol{Q}}_{\boldsymbol{t}}$ (which is a random variable) as a function of $t$. This lemma was first proved by Freund and Schapire [211] for the case of multiplicative weights, but can be easily shown to hold for all optimal no-regret strategies. We provide the proof in Appendix 5.3 for completeness.

**Lemma 5.3.6.** *If player 1 and 2 are both playing optimal no-regret strategies each with rate $(1/2, c)$, the time-average of player 2's mixed strategies evolves as*

$$|\overline{\boldsymbol{Q}}_{\boldsymbol{t}} - q^*| \leq \frac{C_a}{\sqrt{t}}, \tag{5.19}$$

*where $C_a$ is some positive constant that depends on the parameters of the game.*

Now that we have justified Equations (5.18a) and (5.18b) as central to the definition of a fixed-convergent sequence of player 2, we state our result on player 1's last-iterate divergence against such a sequence below.

**Theorem 5.3.7.** *Assume, as before, that player 1 uses a mean-based repeated game strategy $\{f_t\}_{t\geq 1}$ that is uniformly no-regret and has a regret rate of $(1/2, c)$. Then, for any fixed-convergent choice of strategies of player 2, player 1's limiting mixed strategy diverges with a positive probability. In other words, there exist positive constants $(\delta, \epsilon)$ such that for any $\{q_t\}_{t\geq 1} \in \mathcal{Q}_{\delta,t_0,C}$, we have*

$$\limsup_{t\to\infty} \mathbb{P}\left[|\boldsymbol{P}_{\boldsymbol{t}} - p^*| \geq \delta\right] \geq \epsilon \tag{5.20a}$$

The proof of Theorem 5.3.7 builds on the ideas in the proof of Theorem 5.3.3, but is more technical, primarily owing to the need to study the probability mass function of the Poisson-binomial random variable from the sum of independent Bernoulli($q_t$) random variables, and is therefore contained in Appendix 5.3.

### A comment on adaptive strategies

Note that Theorem 5.3.7 does not, strictly speaking, tell us exactly what happens when player 2 is using his own no-regret strategy. The difficulties that arise when studying this case are primarily technical. While the iterates $\{\boldsymbol{Q_t}\}_{t\geq 1}$ do *marginally* satisfy the conditions in Equations (5.18a) and (5.18b), the actual realizations $\{\boldsymbol{J^t}\}_{t\geq 1}$ are not necessarily mutually independent due to the adaptivity of player 2. This causes especial technical difficulty in the steps of the proof that relate various probability mass functions of the sums of player 2's realizations to one another. In fact, we considered the case of fixed-convergent sequences to ensure that the realizations of player 2's strategies are mutually independent across rounds.

More generally, dependencies across realizations are allowed as long as the probability mass function of player 2's sum of realizations at any step sufficiently resembles the probability mass function of a sum of mutually independent coin tosses. Thus, we do believe that Theorem 5.3.7 implies last-iterate divergence for the full stochastic dynamical system. However, the generality of our algorithmic framework will necessitate new mathematical techniques to formally prove this result. This is an intriguing question for future work.

## Last-iterate divergence beyond mean-based strategies

In this section, we examine the fidelity of our results on last-iterate divergence beyond the exact mean-based assumption in Definition 5.2.4. While the mean-based assumption is fairly strong, it is worth noting that mean-based strategies underlie the design of almost all no-regret algorithms in practice. Moreover, as we will now show, the essence of our results continues to hold even for algorithmic variants of mean-based strategies that are ubiquitous in the online learning and games literature.

One of the most common such variants incorporates a form of recency bias, colloquially called "optimism". We define the broad class of recency-bias strategies below.

**Definition 5.3.8.** *The class of $k$-recency bias strategies is defined as below:*

$$f_t(J^{t-1}) := f_t(\widehat{\boldsymbol{Q}}^{\boldsymbol{\ell}}_{\boldsymbol{t-1}}) \text{ where}$$
$$\widehat{\boldsymbol{Q}}^{\boldsymbol{\ell}}_{\boldsymbol{t}} := \frac{1}{t}\left(\sum_{s=1}^{t} \boldsymbol{J_s} + \sum_{j=1}^{\ell} r_j \boldsymbol{J_{t-j+1}}\right),$$

*and $\{r_j\}_{j=1}^{\ell}$ are positive integers taking values in $\{1, \ldots, \ell\}$.*

Note that the class of 0-recency bias strategies essentially constitutes mean-based strategies, and 1-recency bias strategies with $r_1 = 1$ constitutes the class of *optimistic* mean-based strategies, since they are using $\sum_{s=1}^{t} \boldsymbol{J_s} + \boldsymbol{J_t}$ as the summary statistic.

As mentioned in the introduction, the study of the last iterate of optimism-based strategies has generated a lot of interest in the optimization literature [246–250]; more-over, these strategies are known to cause faster time-averaged convergence [251, 252]. Most recently,

it was shown that the last iterate of the players' strategies in the setting of telepathic dynamics (that arises when players use each others' mixtures to update their strategies) will converge [242]. In the more realistic realization model, the following result shows that the ensuing stochasticity *alone* causes recency-bias-based strategies to diverge.

**Theorem 5.3.9.** *Assume that player 2's strategy $\{\boldsymbol{J_t}\}_{t\geq 1}$ is an i.i.d. sequence of Bernoulli($q^*$) random variables. Then, any $\ell$-recency-bias strategy $\{f_t\}_{t\geq 1}$, as defined in Definition 5.3.8, that has a regret rate of $(1/2, c)$ causes player 1's last iterate to diverge, i.e. there exist positive constants $(\delta, \epsilon)$ such that*

$$\limsup_{t\to\infty} \mathbb{P}\left[|\boldsymbol{P_t} - p^*| \geq \delta\right] \geq \epsilon. \tag{5.21}$$

The proof of Theorem 5.3.9 is essentially a relatively simple variant of the proof of Theorem 5.3.3 with slightly more involved algebraic calculations of the involved probability mass functions owing to the recency bias. The full proof is contained in Appendix 5.3. It is worth noting that the bounded memory of the recency bias, as well as bounded increments $\{r_j\}_{j=1}^{\ell}$ are critical to the essence of the argument. It is plausible that stronger recency biases that grow with the number of rounds $t$ could lead to different last-iterate behavior; however, stronger recency biases would also conceivably break the no-regret property.

## 5.4 Conclusion and future work

In this chapter, we have shown partial but compelling evidence for a fundamental tension between the guarantees of no-regret and last-iterate convergence on uncoupled dynamics that use the opponents' realizations alone as feedback. Perhaps the most important immediate question to address is whether Conjecture 5.3.4 formally holds — while the techniques introduced in this chapter provide strong evidence, non-trivial technical work remains to prove the conjecture. Additionally, we can ask whether the mean-based nature of the strategies is truly needed for our impossibility result. While we did show robustness of our results to this assumption (through Theorem 5.3.9 for recency-bias-based strategies), whether the same properties hold for strategies that are very different from mean-based strategies is an intriguing question. Owing to the mean-based nature of the offline benchmark in regret-minimization, it may even be that all no-regret strategies are in a certain approximate sense mean-based.

Section 5.2 provided some preliminary empirical evidence that in practice, last-iterate divergence can occur even when no-regret strategies with sub-optimal rates are used. Our techniques break down in the face of sub-optimal no-regret rates, so it is interesting to ponder whether last-iterate divergence happens even for sub-optimal no-regret algorithms more generally, or if it is just a property of the particular algorithms that were simulated.

Overall, the results presented in this chapter, while preliminary, merit a possible re-examination of choices of dynamics players should use to learn from one another. An alternate choice of dynamics that is of possible interest is the recently proposed (non-constructive)

*smoothly calibrated* strategies [254], which have been shown to converge in the last-iterate to NE *even for non-zero-sum games!* These strategies constitute randomized responses to deterministic forecasting, and are conceptually quite different from strategies satisfying the no-regret property. Whether these strategies can be studied more constructively, and from a behavioral game theory standpoint, is an important and intriguing question for future work.

## 5.5 Technical portions of proofs

### Notation conventions for proofs

Before proving the full proofs of statements, we state our convention for notation. We designate constants that take a value in $(0, 1)$ by $\epsilon$, and strictly positive, finite constants by $C \in (0, \infty)$. Moreover, we will designate $t_0$ as a lower bound on $t$ above which all our statements apply.

In general, these constants can depend on the parameters of the game, either directly, or just on the equilibrium strategies $(p^*, q^*)$.

For ease of exposition, we will also sub-script these constants by alphabets $\{a, b, \ldots\}$, corresponding to the lemmas in which they appear and are used. Thus, for example, in the first lemma the constants will be denoted as $\{\epsilon_a, C_a\}$ and the lower bound on $t$ will be denoted as $t_{0,a}$. While in general we will overload notation within a lemma for our choice of constants, we will be explicit about manipulations when possible.

### Proof of Lemma 5.3.2

We consider the distribution of *mutually independent* coin tosses,

$$\boldsymbol{J_t} \text{ i.i.d. } \sim \text{Bernoulli}(q^*), \tag{5.22}$$

and denote the expectation of quantities under this probability distribution by $\mathbb{E}[\cdot]$. Recall that $q^*$ is the Nash equilibrium strategy of player 2. By linearity of expectation, it is trivial to show the second statement, i.e.

$$\mathbb{E}\left[\sum_{t=1}^{T} \boldsymbol{J_t}\right] = Tq^*.$$

To show the first statement, recall that $\boldsymbol{J_t} \perp (\boldsymbol{J})^{\boldsymbol{t-1}}$ for all $t \in \{1, \dots, T\}$ due to mutual independence. Thus, we use the law of iterated expectations to get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} G(f_t((\boldsymbol{J})^{\boldsymbol{t-1}}), \boldsymbol{J_t})\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[f_t((\boldsymbol{J})^{\boldsymbol{t-1}}) \cdot G(1, \boldsymbol{J}_t) + (1 - f_t((\boldsymbol{J})^{\boldsymbol{t-1}}))G(0, \boldsymbol{J}_t)\middle|(\boldsymbol{J})^{\boldsymbol{t-1}}\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} f_t((\boldsymbol{J})^{\boldsymbol{t-1}}) \cdot G(1, q^*) + (1 - f_t((\boldsymbol{J})^{\boldsymbol{t-1}})) \cdot G(0, q^*)\right] \\
&= TR^*,
\end{aligned}$$

where the last statement follows by statement 1 in Assumption 5.2.5, noting that $G(0, q^*) = G(1, q^*) = G(p^*, q^*) = R^*$. This completes the proof. □

Proving Lemma 5.3.2 completes the proof of Proposition 5.3.1.

## Proof of Lemma 5.3.6

The argument resembles the statement presented in Section 6.1 of [211] for which, if both players are using the multiplicative weights update, we have

$$\max_{p \in [0,1]} G(p, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) \leq g^* + \frac{2c}{\sqrt{t}},$$

where $c$ is the no-regret parameter corresponding to the multiplicative weights algorithm. Here, we show this argument for all no-regret strategies of rate $(1/2, c)$. Note that, since player 1's strategy is no-regret, we have

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t} G(\boldsymbol{P_s}, \boldsymbol{Q_s}) &\geq \max_{p \in [0,1]} G(p, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) - \frac{c}{\sqrt{t}} \\
\implies \max_{p \in [0,1]} G(p, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) &\leq \frac{1}{t}\sum_{s=1}^{t} G(\boldsymbol{P_s}, \boldsymbol{Q_s}) + \frac{c}{\sqrt{t}}.
\end{aligned}$$

On the other hand, since player 2 is also using a no-regret strategy of rate $(1/2, c)$, we have

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t} G(\boldsymbol{P_s}, \boldsymbol{Q_s}) &\leq \min_{q \in [0,1]} G(\overline{\boldsymbol{P}}_{\boldsymbol{t}}, q) + \frac{c}{\sqrt{t}} \\
&\leq \max_{p \in [0,1]} \min_{q \in [0,1]} G(p, q) + \frac{c}{\sqrt{t}} \\
&= g^* + \frac{c}{\sqrt{t}}.
\end{aligned}$$

Putting the two inequalities together, we get

$$\max_{p\in[0,1]} G(p, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) \leq g^* + \frac{2c}{\sqrt{t}}.$$

Recall that we chose the convention that $G(0,1) < G(1,1)$ and $G(1,0) < G(1,1)$. Further, to satisfy the requirement that neither player has a strictly dominated strategy, we then require $G(0,0) > G(0,1)$. We will use this last fact. Also, recall that $G(0,q^*) = G(1,q^*) = G(p^*,q^*)$. Now, we observe that $\max_{p\in[0,1]} G(p, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) = \max\{G(0, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}), G(1, \overline{\boldsymbol{Q}}_{\boldsymbol{t}})\}$ and, thus, there are two cases:

1. Case 1: $\overline{\boldsymbol{Q}}_{\boldsymbol{t}} > q^*$, in which case we get

$$\begin{aligned} G(1, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) &\leq G(1,q^*) + \frac{2c}{\sqrt{t}} \\ \implies (G(1,1) - G(1,0))(\overline{\boldsymbol{Q}}_{\boldsymbol{t}} - q^*) &\leq \frac{2c}{\sqrt{t}} \\ \implies (q^* - \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) &\leq \frac{C'_a}{\sqrt{t}}, \end{aligned}$$

   where $C'_a := 2c/(G(1,1) - G(1,0))$.

2. Case 2: $\overline{\boldsymbol{Q}}_{\boldsymbol{t}} \leq q^*$, in which case we get

$$\begin{aligned} G(0, \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) &\leq G(0,q^*) + \frac{2c}{\sqrt{t}} \\ \implies (G(0,0) - G(0,1))(q^* - \overline{\boldsymbol{Q}}_{\boldsymbol{t}}) &\leq \frac{2c}{\sqrt{t}} \\ \implies (\overline{\boldsymbol{Q}}_{\boldsymbol{t}} - q^*) &\leq \frac{C''_a}{\sqrt{t}}, \end{aligned}$$

   where $C''_a := 2c/(G(0,0) - G(0,1))$.

Taking $C_a := \max\{C'_a, C''_a\}$ and combining the cases above completes the proof.

## Proof of Theorem 5.3.7

Recall that we had earlier defined random variables $\boldsymbol{Z}_{\boldsymbol{t}} \sim \text{Binomial}(t, q^*)$, for $t \geq 1$, and $\boldsymbol{Z}''_{\boldsymbol{t,s}} \sim \boldsymbol{Z}_{\boldsymbol{t-s}} + s$. Our main proof strategy is to show that the ensuing random variables from any fixed-convergent sub-sequence highly resemble the ensuing random variables from a sequence that is already exactly at equilibrium.

We consider the sequence $\{\widetilde{\boldsymbol{J}}_{\boldsymbol{t}}\}_{t\geq 1}$ generated by independent random draws from the fixed-convergent strategy of player 2, i.e. we have $\widetilde{\boldsymbol{J}}_{\boldsymbol{t}} \sim \text{Bernoulli}(q_t)$ for all $t \geq 1$. Define a new random variable, $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}} := \sum_{s=1}^{t} \widetilde{\boldsymbol{J}}_{\boldsymbol{s}}$.

To be able to prove last iterate divergence when the opponent's sequence is independent and Bernoulli($q_t$) at round $t$, we need to lower bound the ratio of pmfs of the random variables $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}}$ and $\boldsymbol{Z}''_{\boldsymbol{t}}$ for all $z \in [tq^*, tq^* + \beta\sqrt{t}]$. First, note that by Equation (5.16), it suffices to show that the ratio of the pmfs of random variables $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}}$ and $\boldsymbol{Z}_{\boldsymbol{t}}$ is lower bounded by a universal positive constant. This constitutes the main technical effort of our proof, which we do through a series of key lemmas.

First, we show that the pmf of $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}}$ is very close to the pmf of a Binomial$(t, q^*)$ for any fixed-convergent sequence $\{q_t\}_{t\geq 1}$ satisfying Equations (5.18a) and (5.18b). This is encapsulated in the following technical lemma.

**Lemma 5.5.1.** *Consider any fixed-convergent sequence $\{q_t\}_{t\geq 1}$, i.e. such that Equations* (5.18a) *and* (5.18b) *both hold. Let $\widetilde{\boldsymbol{Z}}'_{\boldsymbol{t}} :=$ Binomial$(t, \overline{q}_t)$. Then, there exists positive constant $\epsilon_b$ and integer $t_{0,b}$ such that for all $t \geq t_{0,b}$, we have*

$$\frac{\mathbb{P}(\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}} = z)}{\mathbb{P}(\widetilde{\boldsymbol{Z}}'_{\boldsymbol{t}} = z)} \geq \epsilon_b \text{ for all } z \in [tq^*, tq^* + \beta\sqrt{t}].$$

*Proof.* We consider the constant round index $t_{0,b} := \max\{t_0, 4C^2/\delta^2\}$, and note that by the triangle inequality, we have

$$\begin{aligned} |q_t - \overline{q}_t| &\leq |q_t - q^*| + |q^* - \overline{q}_t| \\ &\leq \frac{\delta}{2} + \frac{C}{\sqrt{t}} \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} = \delta, \end{aligned}$$

where the last inequality holds for all $t \geq t_{0,b}$. Thus, we have $|q_t - \overline{q}_t| \leq \delta$, which is useful for comparing the pmfs of the random variables $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}}$ and $\widetilde{\boldsymbol{Z}}'_{\boldsymbol{t}}$.

Consider a fixed $t \geq t_{0,b}$. In general, relating the probability mass function of $\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}}$, which is the Poisson binomial random variable, directly to the binomial distribution, is challenging. The following technical lemma characterizes the sequence $\{q_s\}_{s=1}^t$ that *minimizes* the probability mass function $\mathbb{P}(\widetilde{\boldsymbol{Z}}_{\boldsymbol{t}} = z)$ for a fixed choice of $z$. This minimizing sequence takes values $q_s \in \{\overline{q}_t - \delta, \overline{q}_t, \overline{q}_t + \delta\}$, which turns out to be a much simpler form to analyze.

**Lemma 5.5.2.** *Consider any round index $t \geq 1$. Then, for every $z \in \{1, \ldots, t\}$, there exists an even integer $0 \leq n_t(z) \leq t$ such that*

$$\mathbb{P}(\widetilde{Z}_t = z) \geq \mathbb{P}(\widetilde{\widetilde{Z}}_t = z),$$

*where $\widetilde{\widetilde{Z}}_t :=$ Binomial$\left(\frac{n_t(z)}{2}, q^* + \delta\right)$ + Binomial$\left(\frac{n_t(z)}{2}, q^* - \delta\right)$ + Binomial$(t - n_t(z), q^*)$.*

*Proof.* Let $\eta_s := q_s - \overline{q}_t$, for $1 \leq s \leq t$, denote the deviation of $q_s$ from the average at time $t$, $\overline{q}_t$. Thus we have $\sum_{s=1}^t \eta_s = 0$, and $\eta_s \in [-\delta, \delta]$ for all $s \in \{1, \ldots, t\}$.

Let $e^t := \{e_1, e_2, \ldots, e_t\}$ where $e_s \in \{-1, +1\}$ for all $1 \leq s \leq t$ represent the unique encoding of the output sequence $J^t \in \{0,1\}^t$. Let $|e^t| := |\{e_s = +1 : 1 \leq s \leq t\}|$ denote the number of positive ones in the vector $e^t$. Now, we consider $1 \leq z \leq t$. We have,

$$\begin{aligned}\mathbb{P}(\widetilde{Z}_t = z) &= \sum_{|e^t|=z} \prod_{s=1}^{t} q_s(\mathsf{i})\{e_s = +1\} + (1-q_s)(\mathsf{i})\{e_s = -1\}) \\ &= \sum_{|e^t|=z} \prod_{s=1}^{t} (\overline{q}_t + \eta_s)(\mathsf{i})\{e_s = +1\} + (1-\overline{q}_t - \eta_s)(\mathsf{i})\{e_s = -1\}).\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathbb{P}(\widetilde{Z}'_t = z) &= \sum_{|e^t|=z} \prod_{s=1}^{t} (\overline{q}_t)(\mathsf{i})\{e_s = +1\} + (1-\overline{q}_t)(\mathsf{i})\{e_s = -1\}) \\ &= \binom{t}{z} (\overline{q}_t)^z (1-\overline{q}_t)^{(t-z)}.\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\mathbb{P}(\widetilde{Z}_t = z)}{\mathbb{P}(\widetilde{Z}'_t = z)} &= \binom{t}{z}^{-1} \sum_{|e^t|=z} \prod_{s=1}^{t} \frac{\overline{q}_t + \eta_s}{\overline{q}_t}(\mathsf{i})\{e_s = +1\} + \frac{1-\overline{q}_t - \eta_s}{1-\overline{q}_t}(\mathsf{i})\{e_s = -1\}) \\ &= \binom{t}{z}^{-1} \sum_{|e^t|=z} \prod_{s=1}^{t} \left(1 + \frac{\eta_s}{\overline{q}_t}\right)(\mathsf{i})\{e_s = +1\} + \left(1 - \frac{\eta_s}{1-\overline{q}_t}\right)(\mathsf{i})\{e_s = -1\}).\end{aligned}$$

Let $\widehat{e}_s = \frac{+1}{\overline{q}_t}$ if $e_s = +1$ and $\widehat{e}_s = \frac{-1}{1-\overline{q}_t}$ if $e_s = -1$. Let $\widehat{e}^t = \{\widehat{e}_1, \ldots, \widehat{e}_t\}$ and let $|\widehat{e}^t| := |\{\widehat{e}_s = +1/\overline{q}_t : 1 \leq s \leq t\}|$. Then, we get

$$\frac{\mathbb{P}(\widetilde{Z}_t = z)}{\mathbb{P}(\widetilde{Z}'_t = z)} = \binom{t}{z}^{-1} \sum_{|\widehat{e}^t|=z} \prod_{s=1}^{t} (1 + \widehat{e}_s \eta_s). \tag{5.23}$$

We will now try to lower bound the ratio $\frac{\mathbb{P}(\widetilde{Z}_t=z)}{\mathbb{P}(\widetilde{Z}'_t=z)}$ over $\eta^t := \{\eta_1, \ldots, \eta_t\}$ such that $\eta_s \in [-\delta, \delta]$ for all $1 \leq s \leq t$ and $\sum_{s=1}^{t} \eta_s = 0$. Let $F$ denote the set of all such vectors $\eta^t$. Let

$$P(\eta^t) = \sum_{|\widehat{e}^t|=z} \prod_{s=1}^{t} (1 + \widehat{e}_s \eta_s),$$

for $\eta \in F$, and let

$$\widetilde{\eta}^t \in \arg\min_{\eta^t \in F} P(\eta).$$

Note that $P(\eta)$ is a multinomial in $\eta_1, \ldots, \eta_t$. We now show that $\widetilde{\eta}^t$ satisfies: $\widetilde{\eta}_s \in \{-\delta, 0, \delta\}, \forall 1 \leq s \leq t$. First note that if $\widetilde{\eta}_s \in \{-\delta, \delta\}$ for all $1 \leq s \leq t$ then we are done. If this does not hold, then without loss of generality let $\widetilde{\eta}_t \in (-\delta, \delta)$. Since $\sum_{s=1}^{t} \widetilde{\eta}_s = 0$, let us substitute $\widetilde{\eta}_t = -\sum_{s=1}^{t-1} \widetilde{\eta}_s$. We now argue that $\widetilde{\eta}_1 \in \{-\delta, \delta\}$. We have

$$\sum_{|\widehat{e}^t|=z} \prod_{s=1}^{t}(1 + \widehat{e}_s\widetilde{\eta}_s) = \sum_{|\widehat{e}^t|=z} (1 + \widehat{e}_1\widetilde{\eta}_1)(1 - \widehat{e}_t(\widetilde{\eta}_1 + \cdots + \widetilde{\eta}_{t-1}) \prod_{s=2}^{t-1}(1 + \widehat{e}_s\widetilde{\eta}_s)$$

$$= \sum_{|\widehat{e}^t|=z} (1 + \widehat{e}_1\widetilde{\eta}_1 - \widehat{e}_t\widetilde{\eta}_1 - \widehat{e}_t(\widetilde{\eta}_2 + \cdots + \widetilde{\eta}_{t-1}) - \widehat{e}_1\widehat{e}_t\widetilde{\eta}_1^2 - \widehat{e}_1\widehat{e}_2(\widetilde{\eta}_2 + \cdots + \widetilde{\eta}))H(\widehat{e}_2, \ldots, \widehat{e}_{t-1}),$$

where

$$H(\widehat{e}_2, \ldots, \widehat{e}_{t-1}) = \prod_{s=2}^{t-1}(1 + \widehat{e}_s\widetilde{\eta}_s).$$

Note that the above is a quadratic expression in $\widetilde{\eta}_1$. We now observe that the coefficient of $\widetilde{\eta}_1$ in this expression is zero. Indeed, the coefficient of $\widetilde{\eta}_1$ is given by

$$\sum_{|\widehat{e}|=z} H(\widehat{e}_2, \ldots, \widehat{e}_{t-1})(\widehat{e}_1 - \widehat{e}_t) = 0,$$

because of the symmetry in $\widehat{e}_1$ and $\widehat{e}_t$ in the above expression. A quadratic of the form $ax^2 + b$ attains its minimum on an interval $[l, h]$ either at $x = l, h$ or $x = 0$.

This establishes that $\widetilde{\eta}_1 \in \{-\delta, 0, \delta\}$, and indeed the same argument works for all $t \in \{1, \ldots, (T-1)\}$. Moreover, we get $\widetilde{\eta}_t \in \{-\delta, 0, \delta\}$, as these are the only choices that can allow $\sum_{s=1}^{t} \widetilde{\eta}_s = 0$.

Thus, we have established that $\widetilde{\eta}_s \in \{-\delta, 0, \delta\}$ for all $s \in \{1, \ldots, t\}$.

Thus, there must be exact $n_t(z)/2$ values of $s$ corresponding to $\widetilde{\eta}_s = \delta$, $n_t(z)/2$ values of $s$ corresponding to $\widetilde{\eta}_s = -\delta$, and $(t - n_t(z))$ values of $s$ corresponding to $\widetilde{\eta}_s = 0$. Thus, we have shown that

$$\mathbb{P}(\widetilde{Z}_t = z) \geq \mathbb{P}(\widetilde{\widetilde{Z}}_t = z),$$

where $\widetilde{\widetilde{Z}}_t := \text{Binomial}\left(\frac{n_t(z)}{2}, q^* + \delta\right) + \text{Binomial}\left(\frac{n_t(z)}{2}, q^* - \delta\right) + \text{Binomial}\left(t - n_t(z), q^*\right)$. $\square$

We now state and prove a final technical lemma relating the random variables $\widetilde{\widetilde{Z}}_t$ and $Y_t = \text{Binomial}(t, \overline{q}_t)$.

**Lemma 5.5.3.** *Let $Z_t(n) := Binomial\left(\frac{n}{2}, \overline{q_t} + \delta\right) + Binomial\left(\frac{n}{2}, \overline{q_t} - \delta\right) + Binomial\left(t - n, \overline{q_t}\right)$ for any even $n \in \{1, \ldots, t\}$. Then, there exists universal constant $\epsilon_c > 0$ such that for every $z \in [q^*t, q^*t + \beta\sqrt{t}]$, we have*

$$\frac{\mathbb{P}(Z_t(n) = z)}{\mathbb{P}(Y_t = z)} \geq \epsilon_c > 0.$$

*Here, $Y_t = Binomial(t, \overline{q}_t)$.*

Note that Lemma 5.5.3 immediately implies that

$$\frac{\mathbb{P}(\widetilde{\widetilde{Z}}_t = z)}{\mathbb{P}(Y_t = z)} = \frac{\mathbb{P}(Z_t(n_t(z)) = z)}{\mathbb{P}(Y_t = z)} \geq \epsilon_c > 0,$$

and we have thus related the original random variable $\widetilde{Z}_t$ to the Binomial random variable $Y_t$ through the constant $\epsilon_b := \epsilon_c$, which completes our proof argument. Thus, it only remains to prove Lemma 5.5.3, which we do below.

*Proof.* Our proof will critically use the classical de-Moivre-Laplace theorem, stated below.

**Theorem 5.5.4** (de-Moivre and Laplace, statement from [255])**.** *We denote the pdf of the normal distribution $\mathcal{N}(\mu, \sigma^2)$ by $p(x; \mu, \sigma^2)$. Further, let $X \sim$ Binomial$(t, q)$ for any $0 < q < 1$, and consider any sequence $\{k_t\}_{t\geq 1}$ such that $k_t^3/t^2 \to 0$ as $t \to \infty$. Then, for every $0 < \epsilon < 1$, there exists a $t_0$ sufficiently large such that for all $t > t_0$, we have*

$$1 - \epsilon < \frac{\mathbb{P}(X = z)}{p(z; tq, tq(1-q))} \leq 1 + \epsilon, \text{ for all integers } qt - k_t \leq z \leq qt + k_t.$$

Theorem 5.5.4 is a much sharper form of asymptotic normality than the typically stated Central Limit Theorem, as it obtains direct control on the probability mass function itself. To see how we can apply the de-Moivre-Laplace theorem to the denominator $\mathbb{P}(Y_t = z)$, we fix $q := \overline{q}_t$. Then, note that since $z \in [tq^*, tq^* + \beta\sqrt{t}]$ and (from Lemma 5.3.6), we have $\overline{q}_t \in \{q^* - C/\sqrt{t}, q^* + C/\sqrt{t}\}$, we have $z \in [t\overline{q}_t - C\sqrt{t}, t\overline{q}_t + (C+\beta)\sqrt{t}]$. Thus, designating $C_c := (\beta + C)$, we consider the choice of sequence $\{k_t = C_c\sqrt{t}\}_{t\geq 1}$. This sequence clearly satisfies $k_t^3/t^2 \to 0$, and so we can directly apply the statement of the DeMoivre-Laplace theorem to get

$$(1-\epsilon_c)\cdot p\left(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t})\right) \leq \mathbb{P}(Y_t = z) \leq (1+\epsilon_c)\cdot p\left(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t})\right) \text{ for } t \geq t_{0,c}.$$

Further, we will adjust $t_{0,c}$ such that $t > t_{0,c} := t_{0,c}/q^*$.

There are two cases to study depending on the value that $n$ takes. The first one considers $n \leq t_{0,c}$. Noting that $t_{0,c}$ is a constant, in this case we can directly bound the ratio of pmfs. First, we very crudely lower bound the numerator to get

$$\begin{aligned}
\mathbb{P}(\widetilde{\widetilde{Z}}_t = z) = & \sum_{0\leq k_1,k_2\leq n/2, 0\leq k_3\leq (t-n), k_1+k_2+k_3=z} \binom{n/2}{k_1}\binom{n/2}{k_2}\binom{t-n}{k_3} \\
& (\overline{q_t}+\delta)^{k_1}\cdot(1-\overline{q_t}-\delta)^{n/2-k_1}\cdot(\overline{q_t}-\delta)^{k_2}\cdot(1-\overline{q_t}+\delta)^{n/2-k_2}\cdot(\overline{q_t})^{k_3}(1-\overline{q_t})^{t-n-k_3} \\
& > \binom{t-n}{z-n}(\overline{q_t}+\delta)^{n/2}(\overline{q_t}-\delta)^{n/2}(\overline{q_t})^{z-n}(1-\overline{q_t})^{t-z},
\end{aligned}$$

where in the last inequality we considered only the point $k_1 = k_2 = n/2, k_3 = z - n$ in the sum. (Note that this is a valid point as $z \leq t$ and $z - n \geq q^*t - t_{0,c} > 0$. The latter inequality

follows because we assumed that $q^*T > t_{0,c}$.) On the other hand, for the denominator we have

$$\mathbb{P}(Y_t = z) = \binom{t}{z}(\overline{q_t})^z(1-\overline{q_t})^{t-z},$$

and so we get, after some algebraic simplification,

$$\begin{aligned}\frac{\mathbb{P}(\widetilde{\widetilde{Z}}_t = z)}{\mathbb{P}(Y_t = z)} &> \frac{\binom{t-n}{z-n}\left(1-\frac{\delta^2}{\overline{q_t}^2}\right)^{n/2}}{\binom{t}{z}} \\ &\geq \epsilon_c > 0,\end{aligned}$$

where the constant $\epsilon_c$ will depend on $\overline{q_t}, \delta, t_{0,c}$, but not on $t$. Here, we have critically used $n \leq t_{0,c}$ to lower bound the term $\left(1-\frac{\delta^2}{\overline{q_t}^2}\right)$ by such a constant, as well as noting that

$$\begin{aligned}\frac{\binom{t-n}{z-n}}{\binom{t}{z}} &= \frac{z}{t}\cdot\frac{z-1}{t-1}\cdots\frac{z-n+1}{t-n+1} \\ &\geq \left(\frac{z-n+1}{t-n+1}\right)^n \\ &\geq (\epsilon_c)^n \geq (\epsilon_c)^{t_{0,c}},\end{aligned}$$

for some constant $\epsilon_c$ that is close to $q^*$.

Notice that the above crude argument does not work for the case where $n > t_{0,c}$, in particular, if it can grow indefinitely as a function of $t$, is less trivial. For this case, we make the following claim using the de-Moivre-Laplace theorem, under which it suffices to prove the lemma.

**Claim 5.5.5.** *There exists a constant $\epsilon_c \in (0,1)$ that can depend on $C_c$, but is independent of $(t,n)$, such that for $t_{0,c} \leq n \leq t$, we have*

$$\mathbb{P}(Y_t = z) \leq (1+\epsilon_c)\cdot p\left(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t})\right) \tag{5.24a}$$

$$\mathbb{P}(Z_t(n) = z) \geq (1-\epsilon_c)\cdot p\left(z; t\overline{q_t}, t\sigma^2(\overline{q_t}, n, t)\right) \text{ norm where} \tag{5.24b}$$

$$\sigma^2(\overline{q_t}, n, t) := \frac{1}{t}\left(\frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta) + \frac{n}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + (t-n)\overline{q_t}(1-\overline{q_t})\right)$$

*for any $z \in [tq^* - C_c\sqrt{t}, tq^* + C_c\sqrt{t}]$.*

First, notice that Claim 5.5.5 directly gives us our proof for the case where $n \geq t_{0,c}$. To see this, consider the second case where $\overline{q_t} > 1/2$. This gives us

$$\begin{aligned}\frac{\mathbb{P}(Z_t(n) = z)}{\mathbb{P}(Y_t = z)} &\geq \frac{(1-\epsilon_c)}{(1+\epsilon_c)}\cdot\frac{p(z; t\overline{q_t}, t\sigma^2(\overline{q_t}, n, t))}{p\left(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t})\right)} \\ &= \frac{(1-\epsilon_c)}{(1+\epsilon_c)}\cdot\frac{\sqrt{2\pi\cdot t(\overline{q}_t)(1-\overline{q_t})}}{\sqrt{2\pi\cdot t\sigma^2(\overline{q_t}, n, t)}}\cdot\frac{e^{-\frac{(z-t\overline{q_t})^2}{2t\sigma^2(\overline{q_t}, n, t)}}}{e^{-\frac{(z-t\overline{q_t})^2}{2t\overline{q_t}(1-\overline{q_t})}}}.\end{aligned}$$

First, we note that $\sigma^2(\overline{q_t}, n, t) \leq \frac{1}{4}$. Moreover, we know that $\overline{q_t} \in [q^* - \delta, q^* + \delta]$, and so we have

$$\frac{\sqrt{2\pi \cdot t\overline{q_t}(1-\overline{q_t})}}{\sqrt{2\pi \cdot t\sigma^2(\overline{q_t}, n, t)}} \geq \epsilon_c > 0,$$

where $\epsilon_c$ is a constant that depends only on $\delta$. Thus, we get

$$\frac{\mathbb{P}(Z_t(n) = z)}{\mathbb{P}(Y_t = z)} \geq \epsilon_c \cdot \frac{e^{-\frac{(z-t\overline{q_t})^2}{2t\sigma^2(\overline{q_t}, n, t)}}}{e^{-\frac{(z-t\overline{q_t})^2}{2t\overline{q_t}(1-\overline{q_t})}}}.$$

Finally, we note that $z \in [tq^* - C_c\sqrt{t}, tq^* + C_c\sqrt{t}]$. Thus, to lower bound the numerator we get

$$\begin{aligned} e^{-\frac{(z-t\overline{q_t})^2}{2t\sigma^2(\overline{q_t}, n, t)}} &\geq e^{-\frac{4C_c^2 \cdot t}{2t\sigma^2(\overline{q_t}, n, t)}} \\ &\geq e^{-\frac{4C_c^2}{2\sigma^2(\overline{q_t}, n, t)}} \geq \epsilon_c > 0, \end{aligned}$$

where we now use the fact that $\sigma^2(\overline{q_t}, n, t) \geq (\overline{q_t} + \delta)(1 - \overline{q_t} - \delta)$ (this is a consequence of $\overline{q_t} > 1/2$). [9] Thus, we get $\overline{q_t} + \delta \leq q^* + 2\delta < 1$. Note that this constant $\epsilon_c$ will depend on $(q^*, \delta, C_c)$, but is independent of $t$.

For the denominator, we trivially have $e^{-\frac{(z-t\overline{q_t})^2}{2t\overline{q_t}(1-\overline{q_t})}} \leq 1$. Putting all of these together, we get

$$\frac{\mathbb{P}(Z_t(n) = z)}{\mathbb{P}(Y_t = z)} \geq \epsilon_c > 0,$$

where $\epsilon_c$ is the product of all the above constants and thus depends on $(t_{0,c}, q^*, C_c, \delta)$, but is independent of $t$. Thus, given Claim 5.5.5, we have proved Lemma 5.5.3. (A symmetric argument, which we omit, also works for the case $\overline{q_t} \leq 1/2$.)

It only remains to prove this claim, which we do below using the DeMoivre-Laplace theorem.

*Proof of Claim 5.5.5.* As we noted above, Equation (5.24a) follows immediately from the statement of Theorem 5.5.4. To prove Equation (5.24b), we need to do a little more work, but essentially we can exploit the mixture-of-binomials structure in the random variable $Z_t(n) := \text{Binomial}\left(\frac{n}{2}, \overline{q_t} + \delta\right) + \text{Binomial}\left(\frac{n}{2}, \overline{q_t} - \delta\right) + \text{Binomial}\left(t - n, \overline{q_t}\right)$ for any *even* $n \in \{1, \ldots, t\}$.

First, we consider the extreme case where the distribution is "most different" from $Y_t$, i.e. $n = t$. In this case, note that $Z_t(t) := Z_{t,1} + Z_{t,2}$ where $Z_{t,1} \sim \text{Binomial}\left(\frac{t}{2}, \overline{q_t} + \delta\right)$ and

[9] It seems to me that this would hold more generally and we do not need the assumption $\overline{q}_T > 1/2$.

$Z_{t,2} \sim \text{Binomial}\left(\frac{t}{2}, \overline{q_t} - \delta\right)$, and the random variables $Z_{t,1}$ and $Z_{t,2}$ are independent. Thus, we get

$$\begin{aligned}\mathbb{P}(Z_t(t) = z) &= \sum_{y=z-t\overline{q_t}+C_c\sqrt{t}}^{t\overline{q_t}-C_c\sqrt{t}} \mathbb{P}(Z_{t,1} = y)\mathbb{P}(Z_{t,2} = (z-y)) \\ &\geq \sum_{y=\frac{t}{2}\cdot(\overline{q_t}+\delta)-C_ct^{5/9}}^{\frac{t}{2}\cdot(\overline{q_t}+\delta)+C_ct^{5/9}} \mathbb{P}(Z_{t,1} = y)\mathbb{P}(Z_{t,2} = (z-y)).\end{aligned}$$

Now, observe that $y \in [t/2 \cdot (\overline{q_t} + \delta) - C_c t^{5/9}, t/2 \cdot (\overline{q_t} + \delta) + C_c t^{5/9}]$, and because we have assumed that $z \in [\overline{q_t} - C_c\sqrt{t}, \overline{q_t} + C_c\sqrt{t}]$, we also have $(z-y) \in [t/2 \cdot (\overline{q_t} - \delta) - C_c t^{5/9}, t/2 \cdot (\overline{q_t} - \delta) + C_c t^{5/9}]$ for slightly adjusted constant $C_c$. Moreover, it is easy to verify that the sequence $\{k_t := C_c t^{5/9}\}_{t\geq 1}$ satisfies the conditions required for application of de-Moivre-Laplace theorem.

Therefore, for large enough $t \geq t_{0,c}$ (where $t_{0,c}$ will depend on $(\epsilon_c, \delta, q^*, C_c)$, and appropriately chosen constant $\epsilon_c \in (0,1)$, and the specified ranges of $(y,z)$, we get

$$\begin{aligned}\mathbb{P}(Z_{t,1} = y) &\geq (1-\epsilon_c) \cdot p\left(y; \frac{t}{2}(\overline{q_t}+\delta), \frac{t}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta)\right) \\ \mathbb{P}(Z_{t,2} = (z-y)) &\geq (1-\epsilon_c) \cdot p\left((z-y); \frac{t}{2}(\overline{q_t}-\delta), \frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta))\right),\end{aligned}$$

and so we get

$$\begin{aligned}&\mathbb{P}(Z_t(t) = z) \\ &\geq (1-\epsilon_c)^2 \sum_{y=\frac{t}{2}\cdot(\overline{q_t}+\delta)-C_ct^{5/9}}^{\frac{t}{2}\cdot(\overline{q_t}+\delta)+C_ct^{5/9}} p\left(y; \frac{t}{2}(\overline{q_t}+\delta), \frac{t}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta)\right) \cdot \\ &p\left((z-y); \frac{t}{2}(\overline{q_t}-\delta), \frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right) \\ &\overset{\text{(i)}}{\geq} \cdot(1-\epsilon_c)^2 \cdot p\left(z; \frac{t}{2}(\overline{q_t}+\delta), \frac{t}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta)\right) \star p\left(z; \frac{t}{2}(\overline{q_t}-\delta), \frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right) \\ &- 2(1-\epsilon_c)^2 \cdot e^{-C_c t^{1/9}} \\ &= (1-\epsilon_c)^2 \cdot p\left(z; t\overline{q_t}, \frac{t}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + \frac{t}{2}(\overline{q_t}-\delta), \frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right) \\ &- 2(1-\epsilon_c)^2 \cdot e^{-C_c t^{1/9}} \\ &\overset{\text{(ii)}}{\geq} \epsilon_c(1-\epsilon_c) p\left(z; t\overline{q_t}, \frac{t}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + \frac{t}{2}(\overline{q_t}-\delta), \frac{t}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right).\end{aligned}$$

Here, inequality (ii) follows for large enough $t \geq t_{0,c}$ noting that for the specified range of $z$, we have $p\left(z; t\overline{q}_t, \frac{t}{2}(\overline{q}_t+\delta)(1-\overline{q}_t-\delta)+\frac{t}{2}(\overline{q}_t-\delta), \frac{t}{2}(\overline{q}_t-\delta)(1-\overline{q}_t+\delta)\right) \geq \epsilon_c > 0$; and also that $e^{-C_c t^{1/9}}$ goes to 0 as $t \to \infty$. Inequality (i) follows by noting that

$$\sum_{y > \frac{t}{2}\cdot(\overline{q}_t+\delta)+C_c t^{5/9}} p\left(y; \frac{t}{2}(\overline{q}_t+\delta), \frac{t}{2}(\overline{q}_t+\delta)(1-\overline{q}_t-\delta)\right) \leq \mathbb{P}\left(\boldsymbol{W} > C_c t^{1/18}\right) \leq e^{-C_c t^{1/9}},$$

where $\boldsymbol{W}$ denotes the standard normal random variable, and we have overloaded notation in choices of $C_c$. Similarly, we have

$$\sum_{y < \frac{t}{2}\cdot(\overline{q}_t-\delta)-C_c t^{5/9}} p\left(y; \frac{t}{2}(\overline{q}_t-\delta), \frac{t}{2}(\overline{q}_t-\delta)(1-\overline{q}_t+\delta)\right) \leq e^{-C_c t^{1/9}}.$$

From this, we get

$$\begin{aligned}
&\sum_{y=\frac{t}{2}\cdot(\overline{q}_t+\delta)-C_c t^{5/9}}^{\frac{t}{2}\cdot(\overline{q}_t+\delta)+C_c t^{5/9}} p\left(y; \frac{t}{2}(\overline{q}_t+\delta), \frac{t}{2}(\overline{q}_t+\delta)(1-\overline{q}_t-\delta)\right) \cdot \\
&p\left((z-y); \frac{t}{2}(\overline{q}_t-\delta), \frac{t}{2}(\overline{q}_t-\delta)(1-\overline{q}_t+\delta)\right) \\
&\geq p\left(z; \frac{t}{2}(\overline{q}_t+\delta), \frac{t}{2}(\overline{q}_t+\delta)(1-\overline{q}_t-\delta)\right) \star p\left(z; \frac{t}{2}(\overline{q}_t-\delta), \frac{t}{2}(\overline{q}_t-\delta)(1-\overline{q}_t+\delta)\right) \\
&- 2e^{-C_c t^{1/9}},
\end{aligned}$$

and plugging this in above gives inequality (i).

Clearly, we have proved Equation (5.24b) for this extreme case. Let us now extend this extreme case more generally. Recall that we assumed $n \geq t_{0,c}$. In this case, by an identical argument to the above, we get

$$\begin{aligned}
&\mathbb{P}(Z_{n,1} + Z_{n,2} = z') \geq \epsilon_c(1-\epsilon_c)^2 \cdot \\
&p\left(z; n\overline{q}_t, \frac{n}{2}(\overline{q}_t+\delta)(1-\overline{q}_t-\delta)+\frac{n}{2}(\overline{q}_t-\delta), \frac{n}{2}(\overline{q}_t-\delta)(1-\overline{q}_t+\delta)\right).
\end{aligned}$$

Thus, we can utilize a similar convolution argument as before to study $Z_t(n) := Z_{n,1} + Z_{n,2} + Z_{(t-n),3}$, where $Z_{(t-n),3} \sim \text{Binomial}((t-n), \overline{q}_t)$ and is independent from $\{Z_{n,1}, Z_{n,2}\}$. Thus, we get

$$\begin{aligned}
\mathbb{P}(Z_t(n) = z) &= \mathbb{P}(Z_{n,1} + Z_{n,2} + Z_{(t-n),3} = z) \\
&\geq \sum_{z'=n\overline{q}_t - C_c n^{5/9}}^{n\overline{q}_t + C_c n^{5/9}} \mathbb{P}(Z_{n,1} + Z_{n,2} = z')\mathbb{P}(Z_{(t-n),3} = (z-z')).
\end{aligned}$$

There are two cases depending on the value of $(t-n)$:

1. $(t-n) \leq t_{0,c}$. In this case, because $Z_{(t-n),3} \in \{0, \ldots, t_{0,c}\}$, we have

$$\begin{aligned}\mathbb{P}(Z_{(t-n),3} = (z - z')) &\geq (\min\{\overline{q_t}, 1 - \overline{q_t}\})^{(t-n)} \\ &\geq (\min\{\overline{q_t}, 1 - \overline{q_t}\})^{t_{0,c}}.\end{aligned}$$

   Similarly, it is easy to verify that the normal pdf $p((z - z'); (t-n)\overline{q_t}, (t-n)\overline{q_t}(1-\overline{q_t}))$ is also bounded above by a constant $c'$ as $(z - z') \leq t_{0,c}$. Thus, the ratio of the two is bounded by a universal constant $\epsilon_c$ *for all* $(z - z') \in \{0, \ldots, t_{0,c}\}$.

2. The second case is $(t-n) \geq t_{0,c}$. Now, we note that $(z - z') \in [(t-n)\overline{q_t} - C_c\sqrt{n} - C_c\sqrt{t}, (t-n)\overline{q_t} + C_c\sqrt{n} + C_c\sqrt{t}]$, therefore, $(z - z') \in [(t-n)\overline{q_t} - C_c\sqrt{(t-n)}, (t-n)\overline{q_t} + C_c\sqrt{(t-n)}]$ for slightly adjusted constant $C_c$. Then, since $(t-n) \geq t_{0,c}$ as well, we can apply the de-Moivre-Laplace theorem on the Binomial random variable $Z_{(t-n),3}$ and show that

$$\frac{\mathbb{P}(Z_{(t-n),3} = (z - z'))}{p((z - z'); (t-n)\overline{q_t}, (t-n)\overline{q_t}(1-\overline{q_t}))} \geq (1 - \epsilon_c)$$

   for the specified range on $(z - z')$.

Thus, in both cases, for appropriately chosen $\epsilon_c > 0$ we get

$$\begin{aligned}&\mathbb{P}(Z_t(n) = z) \geq \epsilon_c \sum_{z'=n\overline{q_t}-C_c n^{5/9}}^{n\overline{q_t}+C_c n^{5/9}} p\left(z'; n\overline{q_t}, \frac{n}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + \frac{n}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right) \cdot \\ &p\left((z - z'); (t-n)\overline{q_t}, (t-n)\overline{q_t}(1-\overline{q_t})\right) \\ &\geq \epsilon_c^2 \cdot p\left(z; n\overline{q_t}, \frac{n}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + \frac{n}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta)\right) \star \\ &p\left(z; (t-n)\overline{q_t}, (t-n)\overline{q_t}(1-\overline{q_t})\right) \\ &= \epsilon_c^2 \cdot p\left(z; t\overline{q_t}, \frac{n}{2}(\overline{q_t}+\delta)(1-\overline{q_t}-\delta) + \frac{n}{2}(\overline{q_t}-\delta)(1-\overline{q_t}+\delta) + (t-n)\overline{q_t}(1-\overline{q_t})\right),\end{aligned}$$

where the second inequality uses an identical argument as earlier again noting that $n \geq t_{0,c}$. This completes the statement of the claim, and thus completes the proof. □

Now that we have proved Claim 5.5.5, we have completed the proof of Lemma 5.5.3. □

Finally, we show that the pmfs of $Y_t = \text{Binomial}(t, \overline{q}_t)$ and $Z_t = \text{Binomial}(t, q^*)$ are sufficiently close. This follows from the time-averaged convergence property in Lemma 5.3.6 and the argument is detailed below.

**Lemma 5.5.6.** *There exists a positive constant $\epsilon_d > 0$ and a sufficiently large $t_{0,d}$ such that for all $t > t_{0,d}$, we have*

$$\frac{\mathbb{P}(Y_t = z)}{\mathbb{P}(Z_t = z)} \geq \epsilon_d, \forall z \in [q^*t, q^*t + C\sqrt{T}].$$

*Proof.* Again, from the DeMoivre-Laplace theorem, there exists positive constant $\epsilon_d \in (0,1)$ and a sufficiently large $t_{0,d}$, such that for all $t > t_{0,d}$, we have

$$\mathbb{P}(Y_t = z) \geq (1-\epsilon_d) \cdot p\left(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t})\right),$$
$$\mathbb{P}(Z_t = z) \leq (1+\epsilon_d) \cdot p\left(z; tq^*, tq^*(1-q^*)\right).$$

Hence, we have

$$\begin{aligned}\frac{\mathbb{P}(Y_t(n) = z)}{\mathbb{P}(Z_t = z)} &\geq \frac{(1-\epsilon_d)}{(1+\epsilon_d)} \cdot \frac{p(z; t\overline{q_t}, t\overline{q_t}(1-\overline{q_t}))}{p\left(z; tq^*, tq^*(1-q^*)\right)} \\ &= \frac{(1-\epsilon_d)}{(1+\epsilon_d)} \cdot \frac{\sqrt{2\pi \cdot t(\overline{q_t})(1-\overline{q_t})}}{\sqrt{2\pi \cdot tq_t^*(1-q^*)}} \cdot \frac{e^{-\frac{(z-t\overline{q_t})^2}{2t\overline{q_t}(1-\overline{q_t})}}}{e^{-\frac{(z-tq^*)^2}{2tq^*(1-q^*)}}}.\end{aligned}$$

Now, we have

$$\frac{\sqrt{2\pi \cdot t(\overline{q}_t)(1-\overline{q_t})}}{\sqrt{2\pi \cdot tq^*(1-q^*)}} \geq \max\left\{\frac{\sqrt{2\pi \cdot (q^*-\delta)(1-q^*+\delta)}}{\sqrt{2\pi \cdot q^*(1-q^*)}}, \frac{\sqrt{2\pi \cdot (q^*+\delta)(1-q^*-\delta)}}{\sqrt{2\pi \cdot q^*(1-q^*)}}\right\} > 0,$$

for all $t$. This is because $\overline{q_t} \in [q^*-\delta, q^*+\delta]$ and $\overline{q_t}(1-\overline{q_t})$ being concave over this interval attains its minimum on the boundary. Further, we get

$$e^{-\frac{(z-t\overline{q_t})^2}{2t\overline{q_t}(1-\overline{q_t})}} \geq \max\left\{e^{-\frac{2C_d^2}{2(q^*-\delta)(1-q^*+\delta)}}, e^{-\frac{2C_d^2}{2(q^*+\delta)(1-q^*-\delta)}}\right\} > 0.$$

Note that we have obtained bounds that do not depend on $t$. Thus there exists a positive constant $\epsilon_d$ such that the statement in the lemma holds. □

Putting all the equations together, we get

$$\frac{\mathbb{P}(\boldsymbol{Z_t'} = z)}{\mathbb{P}(\boldsymbol{Z_t} = z)} \geq \epsilon \text{ for all } z \in [tq^*, tq^* + C\sqrt{t}].$$

for universal constant $\epsilon > 0$. Thereafter, an identical argument to the proof of Theorem 5.3.3 completes the proof. □

## Proof of Theorem 5.3.9

*Proof.* We will essentially mimic the proof of Theorem 5.3.3. We first define notation pertinent to $\ell$-recency-bias strategies. We denote $\boldsymbol{Z_t^\ell} := \sum_{t'=1}^{t} \boldsymbol{J_{t'}} + \sum_{j=1}^{\ell} r_j \boldsymbol{J_{t-j+1}}$, where $J_s$i.i.d. $\sim$ Bernoulli($q^*$). Note that $\widehat{\boldsymbol{Q}}_{\boldsymbol{t}}^{\boldsymbol{\ell}} = \boldsymbol{Z_t^\ell}/t$. More conveniently, we can also write

$$\boldsymbol{Z_t^\ell} := \sum_{t'=1}^{t} (1 + r'_{t'}) \boldsymbol{J_{t'}},$$

where we designate $r'_{t'} = 0$ for $t' \leq (t - \ell)$, and $r'_{t'} = r_{t-t'+1}$ thereafter. Using this notation, we can then write, for any $0 \leq s \leq t$,

$$(\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} := \sum_{t'=1}^{t-s}(1 + r'_{t'})\boldsymbol{J_{t'}} + \sum_{t'=t-s+1}^{t}(1 + r'_{t'})$$
$$(\widehat{\boldsymbol{Q}}\boldsymbol{''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} := \frac{(\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}}}{t}.$$

From Proposition 5.3.1, we know that there exists a sequence $\{t_k, s_k\}_{k\geq 1}$ such that $0 \leq s_k \leq \alpha(t_k)^{1/2}$ for all $k \geq 1$, and

$$\mathbb{E}\left[f_{t_k}((\widehat{\boldsymbol{Q}}\boldsymbol{''})^{\boldsymbol{\ell}}_{\boldsymbol{t_k,s_k}})\right] \geq p^* + 2\delta \text{ for all } k \geq 1.$$

As in the proof of Theorem 5.3.3, we can use Markov's inequality to get

$$\mathbb{P}\left(f_{t_k}((\widehat{\boldsymbol{Q}}\boldsymbol{''})^{\boldsymbol{\ell}}_{\boldsymbol{t_k,s_k}}) > p^* + \delta\right) \geq \epsilon_0,$$

where $\epsilon_0 := \delta/((1 - p^*) - \delta)$. Note that $0 < \epsilon_0 < 1/2$, as in the proof of Theorem 5.3.3.

Next, we apply the central limit theorem on the recency-biased random variable $\widehat{\boldsymbol{Q}}^{\boldsymbol{\ell}}_{\boldsymbol{t}} = \boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}}/t$. Observe that $\boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}}$ is a sum of bounded random variables (as $r_j \leq \ell$ for all $j \in \{1, \ldots, \ell\}$). In fact, we have

$$\mathbb{E}\left[\boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}}\right] = \left(t + \sum_{j=1}^{\ell} r_j\right)\frac{1}{2}$$
$$\implies \mathbb{E}\left[\widehat{\boldsymbol{Q}}^{\boldsymbol{\ell}}_{\boldsymbol{t}}\right] = \frac{1}{2} + \frac{\sum_{j=1}^{\ell} r_j}{2t} \in \left[\frac{1}{2}, \frac{1}{2} + \frac{\ell^2}{2t}\right].$$

Moreover, we have

$$\text{var}\left[\boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}}\right] = \left((t - \ell) + \sum_{j=1}^{\ell}(1 + r_j)^2\right) \cdot \frac{1}{4} \in \left[\frac{t}{4}, \frac{(t + \ell^2(\ell + 2))}{4}\right]$$

Thus, by the central limit theorem, we have

$$\lim_{t\to\infty} \mathbb{P}\left(\frac{\boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}} - (t + \sum_{j=1}^{\ell} r_j)/2}{\sqrt{\left((t - \ell) + \sum_{j=1}^{\ell}(1 + r_j)^2\right)/4}} > \gamma'\right) = 1 - \phi\left(\gamma'\right),$$

where recall that $\phi$ is the CDF of the standard normal distribution. As before, substituting $\gamma := \frac{\gamma}{\sqrt{1/4}}$, and considering large enough $t \geq T_1(\epsilon_0)$ and suitable choice of $\gamma := \gamma_0$

(just as in the proof of Theorem 5.3.3, although the exact choices could be slightly different here), we get

$$\mathbb{P}\left(\frac{\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{\ell}} - (t+\sum_{j=1}^{\ell} r_j)/2}{\sqrt{\left((t-\ell)+\sum_{j=1}^{\ell}(1+r_j)^2\right)}} > \gamma\right) \leq \frac{\epsilon_0}{4} + \frac{\epsilon_0}{4} = \frac{\epsilon_0}{2}.$$

Thus, we have shown that

$$\frac{\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{\ell}} - (t+\sum_{j=1}^{\ell} r_j)/2}{\sqrt{\left((t-\ell)+\sum_{j=1}^{\ell}(1+r_j)^2\right)}} > \gamma_0$$

with probability at least $(1-\epsilon_0/2)$. Observe that this implies that

$$\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{\ell}} \leq (t+\sum_{j=1}^{\ell} r_j)/2 + \gamma_0\sqrt{\left((t-\ell)+\sum_{j=1}^{\ell}(1+r_j)^2\right)} \leq t/2 + \gamma_0'\sqrt{t}$$

for large enough $t$, where $\gamma_0'$ can depend on $\ell$. Thus, since $t_k - s_k \to \infty$ as $k \to \infty$, there exists a $k_1 > 1$ such that

$$\mathbb{P}\left(t_k \widehat{\boldsymbol{Q}}_{\boldsymbol{t_k},\boldsymbol{s_k}}^{\boldsymbol{\ell}} > 1/2(t_k - s_k) + s_k + \gamma_0'\sqrt{t_k - s_k}\right) \leq \frac{\epsilon_0}{2}$$

for all $k \geq k_1$. Using the bound $s_k \leq \alpha\sqrt{t_k}$, and the union bound we get

$$\mathbb{P}\left(f_{t_k}\left(\frac{(\boldsymbol{Z''})_{\boldsymbol{t_k},\boldsymbol{s_k}}^{\boldsymbol{\ell}}}{t_k}\right) \geq p^* + \delta, (\boldsymbol{Z''})_{\boldsymbol{t_k},\boldsymbol{s_k}}^{\boldsymbol{\ell}} \leq 1/2 \cdot t_k + \beta\sqrt{t_k}\right) \geq \frac{\epsilon_0}{2}. \tag{5.25}$$

Observe that $(\boldsymbol{Z''})_{\boldsymbol{t},\boldsymbol{s}}^{\boldsymbol{\ell}} \geq s'(\ell) := \sum_{j=1}^{\min\{s,\ell\}}(1+r_s) + \max\{(s-\ell),0\}$. We will now show that the pmfs of the random variables are within a constant fraction of each other. The argument resembles the one in the proof of Theorem 5.3.3 with slightly more involved calculations.

**Lemma 5.5.7.** *Denote $\beta_0 := 2\beta$. Then, for any value of $0 \leq s \leq \alpha\sqrt{t}$, and for all $t \geq 1$, we have:*

$$\min_{s'(\ell)\leq z\leq 1/2(t+\beta_0\sqrt{t})} \frac{\mathbb{P}(\boldsymbol{Z}_{\boldsymbol{t}}^{\boldsymbol{\ell}} = z)}{\mathbb{P}((\boldsymbol{Z''})_{\boldsymbol{t},\boldsymbol{s}}^{\boldsymbol{\ell}} = z)} \geq \left(\frac{1}{2}\right)^{\ell^2} \cdot (1+2\beta_0)^{-\alpha}.$$

Using this lemma, we can complete the proof of Theorem 5.3.9 using an exactly identical argument to the proof of Theorem 5.3.3. Since the argument is identical, we omit the details here.

To complete our proof, it thus suffices to prove Lemma 5.5.7, which we now do.

*Proof.* We will split the argument into two cases: $s \geq \ell$ and $s < \ell$. First, we note that for $s \geq \ell$, we have

$$\mathbb{P}((\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} = z) = \mathbb{P}\left(\boldsymbol{Z}_{t-s} = z - s - \sum_{j=1}^{\ell} r_j\right)$$

It is important to note that when $s \geq \ell$, we have $s'(\ell) = \sum_{j=1}^{\ell} r_j + s$, and so $(z - s - \sum_{j=1}^{\ell} r_j) = z - s'(\ell) > 0$. This ensures that the ensuing pmf is valid. Thus, we get

$$\mathbb{P}\left((\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} = z\right) = \binom{(t-s)}{(z - s - \sum_{j=1}^{\ell} r_j)} \left(\frac{1}{2}\right)^{(t-s)}$$

For the numerator, we get

$$\begin{aligned}
&\mathbb{P}\left(\boldsymbol{Z}^{\ell}_{t} = z\right) \\
&= \sum_{x_1,\ldots,x_\ell \in \{0,1\}^\ell} \mathbb{P}\left(\boldsymbol{J_t} = x_1, \boldsymbol{J_{t-1}} = x_2, \ldots, \boldsymbol{J_{t-\ell+1}} = x_\ell\right) \cdot \mathbb{P}\left(\boldsymbol{Z}_{t-\ell} = z - \sum_{j=1}^{\ell}(1 + r_j)x_j\right) \\
&= \sum_{x_1,\ldots,x_\ell \in \{0,1\}^\ell} \mathbb{P}\left(\boldsymbol{J_t} = x_1, \boldsymbol{J_{t-1}} = x_2, \ldots, \boldsymbol{J_{t-\ell+1}} = x_\ell\right) \cdot \binom{(t-\ell)}{(z - \sum_{j=1}^{\ell}(1 + r_j)x_j)} \left(\frac{1}{2}\right)^{t-l}.
\end{aligned}$$

Note that since $z \geq s'(\ell)$ and $z \leq t/2 + \beta\sqrt{t}$, we have for large enough $t$,

$$\begin{aligned}
(z - \sum_{j=1}^{\ell}(1 + r_j)x_j) \geq z - \ell - \sum_{j=1}^{\ell} r_j > z - s'(\ell) > 0, \text{ and} \\
(z - \sum_{j=1}^{\ell}(1 + r_j)x_j) \leq z \leq (t - \ell),
\end{aligned}$$

since for large enough $t$ we know that $t/2 + \beta\sqrt{t} < t - \ell$. Thus, the binomial coefficients above are valid for all values of $\boldsymbol{x} := (x_1, \ldots, x_\ell)$.

Clearly, it suffices to bound the term

$$R(\boldsymbol{x}) = \frac{\binom{(t-\ell)}{(z - \sum_{j=1}^{\ell}(1+r_j)x_j)} \left(\frac{1}{2}\right)^{t-l}}{\binom{(t-s)}{(z-s-\sum_{j=1}^{\ell} r_j)} \left(\frac{1}{2}\right)^{(t-s)}}$$

uniformly for every $\boldsymbol{x} := (x_1, \ldots, x_\ell) \in \{0,1\}^\ell$. We will now do this. Note that

$$R(\boldsymbol{x}) = \underbrace{\frac{\binom{(t-\ell)}{(z - \sum_{j=1}^{\ell}(1+r_j)x_j)}}{\binom{(t-s)}{(z-s-\sum_{j=1}^{\ell} r_j)}}}_{A(\boldsymbol{x})} \cdot \frac{1}{2^{s-l}},$$

and we will now aim to lower bound the term $A(\boldsymbol{x})$. Since the binomial coefficient is uni-modal, we know that

$$\begin{aligned} A(\boldsymbol{x}) &\geq \min\{A_1(\boldsymbol{x}), A_2(\boldsymbol{x})\} \text{ where} \\ A_1(\boldsymbol{x}) &:= \frac{\binom{(t-\ell)}{z}}{\binom{(t-s)}{(z-s-\sum_{j=1}^{\ell} r_j)}} \\ A_2(\boldsymbol{x}) &:= \frac{\binom{(t-\ell)}{(z-\ell-\sum_{j=1}^{\ell} r_j)}}{\binom{(t-s)}{(z-s-\sum_{j=1}^{\ell} r_j)}}. \end{aligned}$$

Thus, we only have to lower bound each of the two terms $A_1(\boldsymbol{x}), A_2(\boldsymbol{x})$. Denoting $z' := z - \sum_{j=1}^{\ell} r_j$ as shorthand, a simple calculation shows that

$$\begin{aligned} A_2(\boldsymbol{x}) &= \frac{(t-\ell)\dots(t-s+1)}{(z'-\ell)\dots(z'-s+1)} \\ &\geq \left(\frac{t-\ell}{z'-\ell}\right)^s \\ &\geq \left(\frac{t-\ell}{t/2+\beta_0\sqrt{t}-\sum_{j=1}^{\ell} r_j-\ell}\right)^{s-\ell}, \end{aligned}$$

where the first inequality used the property that for any $t \geq z'$, the function $\frac{t-i}{z'-i}$ is increasing in $i$.

Moreover, we can show that

$$\begin{aligned} \frac{A_1(\boldsymbol{x})}{A_2(\boldsymbol{x})} &= \frac{\binom{(t-\ell)}{z}}{\binom{(t-\ell)}{z-\ell-\sum_{j=1}^{\ell} r_j}} \\ &= \frac{(t-z+\sum_{j=1}^{\ell} r_j)!}{(t-z-\ell)!} \cdot \frac{(z-\ell-\sum_{j=1}^{\ell} r_j)!}{z!} \\ &= \frac{(t-z+\sum_{j=1}^{\ell} r_j)\dots(t-z-\ell+1)}{z\dots(z-\ell-\sum_{j=1}^{\ell} r_j+1)} \\ &\geq (0.8)^{\sum_{j=1}^{\ell} r_j+\ell} \\ &\geq (0.8)^{\ell^2}, \end{aligned}$$

where the second-to-last inequality holds for large enough $t$ noting that $z \leq t/2 + \beta\sqrt{t}$.

Putting all of this together, we get

$$\begin{aligned} R(\boldsymbol{x}) &\geq (0.8)^{\ell^2} \cdot \left( \frac{(t-\ell)/2}{t/2 + \beta_0\sqrt{t} - \sum_{j=1}^{\ell} r_j - \ell} \right)^{s} \\ &\geq (0.8)^{\ell^2} \cdot \left( \frac{(t-\ell)}{t + 2\beta_0\sqrt{t} - \sum_{j=1}^{\ell} r_j} - \ell \right)^{\alpha\sqrt{t}} \\ &\geq (0.8)^{\ell^2} \cdot (1+2\beta_0)^{-\alpha} \end{aligned}$$

for large enough $t$.

This completes the proof of the ratios for the case $s \geq \ell$. For the case $s < \ell$, we can use an even simpler argument. First, we note that

$$\mathbb{P}\left( (\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} = z \right) = \mathbb{P}\left( \sum_{t'=1}^{t-\ell} \boldsymbol{J_{t'}} + \sum_{j=t-\ell}^{t-s} (1+r_j)\boldsymbol{J_t} = (z-s) \right) > 0,$$

as we know that $(z-s) \leq (t-\ell)$.

Next, we note that we can very crudely lower bound

$$\begin{aligned} \mathbb{P}\left( \boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}} = z \right) &= \mathbb{P}\left( \sum_{t'=1}^{t-\ell} \boldsymbol{J_{t'}} + \sum_{j=t-\ell}^{t-s} (1+r_j)\boldsymbol{J_t} = (z-s) \right) \cdot \mathbb{P}\left( \boldsymbol{J_{t-s+1}} = 1, \ldots, \boldsymbol{J_t} = 1 \right) \\ &= \mathbb{P}\left( \sum_{t'=1}^{t-\ell} \boldsymbol{J_{t'}} + \sum_{j=t-\ell}^{t-s} (1+r_j)\boldsymbol{J_t} = (z-s) \right) \cdot \left( \frac{1}{2} \right)^{s} \\ &> \mathbb{P}\left( \sum_{t'=1}^{t-\ell} \boldsymbol{J_{t'}} + \sum_{j=t-\ell}^{t-s} (1+r_j)\boldsymbol{J_t} = (z-s) \right) \cdot \left( \frac{1}{2} \right)^{\ell}, \end{aligned}$$

where the last step follows because we are in the case where $s < \ell$. and so we get

$$\min_{s'(\ell) \leq z \leq 1/2(t+\beta_0\sqrt{t})} \frac{\mathbb{P}(\boldsymbol{Z}^{\boldsymbol{\ell}}_{\boldsymbol{t}} = z)}{\mathbb{P}((\boldsymbol{Z''})^{\boldsymbol{\ell}}_{\boldsymbol{t,s}} = z)} \geq \left( \frac{1}{2} \right)^{\ell} > \left( \frac{1}{2} \right)^{\ell^2} (1+2\beta_0)^{-\alpha}$$

where the last step follows because $\ell \geq 1$. Putting the two cases together completes the proof of Lemma 5.5.7. □

□

# Chapter 6

# Future Directions

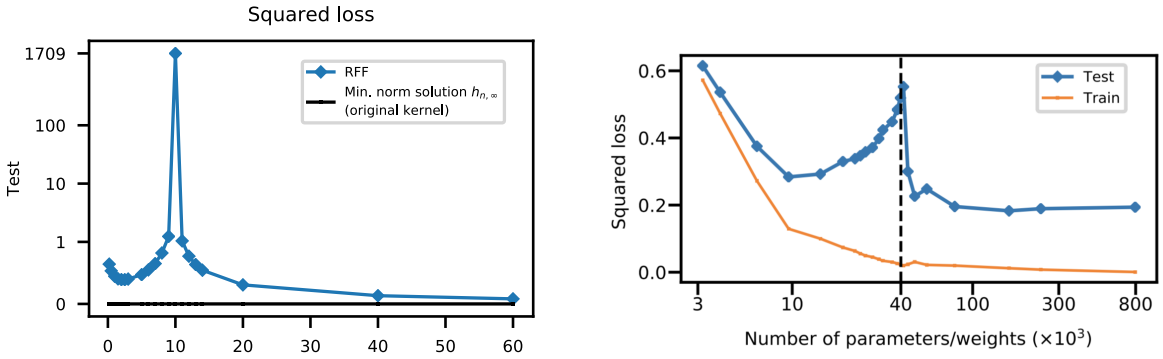
This thesis presented a fundamental perspective on intersecting learning and strategic behavior, by posing two key questions: *how should we learn from data provided by an unknown, possibly strategic agent?* And *how should a strategic agent generate her data in response to learning?* We partially answered these questions for three types of agents: stochastic, adversarial, and competitive; and with the simplest possible ML models like binary sequence prediction, $2 \times 2$ games and linear models. Listed below are key conclusions from each chapter.

1. In Chapters 2 and 3, we designed explicit learning algorithms that successfully adapt not only between stochastic and adversarial environments, but also between different *minimum-description-length* model orders that could describe the environment. The model selection problem highlighted the need to measure guarantees on adaptivity by the metric of overall reward rather than regret.

2. In Chapter 4, we showed that when only one agent is doing the learning, the data generator is incentivized to generate in her data in order to build up a commitment/reputation; thus achieving Stackelberg equilibrium. Implicitly, the data generator is incentivized to reveal her private information to the learner even though the game is non-cooperative!

3. In Chapter 5, we saw that the ubiquitous no-regret learning dynamics can lead to surprisingly divergent day-to-day behavior when deployed against one another, even in the simplest $2 \times 2$ games.

As we saw in each of the chapters, each of these results gives rise to several future questions that need to be addressed in order for a more complete understanding of learning intersected with strategic behavior, even in simple ML models and under purely stochastic, adversarial or competitive environments. To conclude, we briefly discuss two important topics that we did not address in this dissertation: a) intersecting the ideas developed here with the practice of modern machine learning, b) understanding and detecting cooperative environments.




(a) Random Fourier features — test error plotted as a function of number of features.

(b) 2-layer neural networks — test error plotted as a function of the width.

Figure 6.1: Experiments by Belkin, Hsu, Ma and Mandal [125] showing the "double descent" behavior of the test error as a function of model complexity in two popular modern machine learning models. Note that the second descent happens after the model complexity exceeds the sample size and models *interpolate* the training data.

## 6.1 Adaptivity in modern ML

In Chapters 2 and 3, we saw that data-driven model selection in online learning was a critically important problem. However, our perspective and approach was driven by classical statistical learning perspectives (see, for e.g. [2]), which state that the estimation error incurred by using a certain model class will increase, monotonically, with the complexity of the model class[1]. However, in today's practice of modern machine learning, the most successful model classes appear to be the most complex, and are in fact *over-parameterized* with respect to the number of training data points. Figure 6.1, attributed to Belkin, Hsu, Ma and Mandal [125], shows that the *test error* decreases as a function of the model complexity.

This phenomenon has more recently been called the "double descent" behavior, and hints of it were observed earlier in neural networks as well as simpler models [124, 256, 257]. The double descent experiments were unique in that they allowed for tractable theoretical models to study the impact of using over-parameterized models that interpolate noisy training data, at least for linear models. Several recent papers provide a fundamental theoretical understanding of this behavior for linear least-squares regression [133–136, 258, 259] as well as classification [260–263], including corroborating the double-descent behavior under special data generating mechanisms.

It is clear that for online model selection methodology, as we described in Chapters 2 and 3, to be practically applicable to modern ML, it needs to directly engage with this non-standard behavior of over-parameterized models. Clearly, the SRM-based approaches in their current form, i.e. penalizing model complexity, would not have the desired effect:

[1]In fact, this was our reason for designating these classes as *model orders*.

*they would rule out precisely the models that we wish to select!* In fact, the recent efforts to characterize the generalization behavior of over-parameterized models can be viewed as a modern remaking of the principles of SRM, and it would be fascinating to see whether a SRM-based approach could be adapted to work in this regime. In particular, test error could increase or decrease along axes that are a function of the model class that are significantly different from complexity, or number of parameters of the model. On the other hand, partly owing to the surprising success of overly complex models, the primary empirical methodology for data-driven model selection in machine learning is validation on a hold-out set. While we *did* analyze online validation using traditional complexity hierarchies, the algorithm itself did not explicitly penalize complex models in any way. As we remarked in Chapter 2, it would also be fascinating to investigate whether online validation automatically gives us model selection guarantees in this modern regime as well.

## 6.2 Understanding cooperation

In Section 1.4 of this thesis, we stipulated a goal of Alice, the learner, as learning the strategic nature of Bob, which could be stochastic, adversarial, competitive or cooperative. While we examined the first three categories of strategic behavior, we did not at all consider the case of *cooperatively generated* data. This is a possibility that is essential to detect to be able to successfully "learn-to-cooperate" as desired in DARPA's recent Spectrum Challenge, version 2.

Understanding cooperative behavior poses several challenges even when we consider the components of learning and game theory separately. On the side of game theory, the primary model for cooperation involves competition between teams of players, where cooperative behavior could be either externally enforced (as in contract law) or a-priori known. The framework of cooperative game theory then involves analyzing which coalitions will form and how groups of players will collectively act in some non-cooperative form of equilibrium. How these principles of cooperative game theory might manifest in an automated engineering setting is a question as yet under-explored.

On the side of learning, even assuming a *known* cooperative nature of all agents, challenges continue to remain in deploying decentralized schemes in both stochastic and adversarial multi-armed bandit environments [139–141], although promising recent progress has been made [264, 265]. To *learn to cooperate*, we will, at the very least, need to know how to distinguish between multiple possible types of cooperative agents. The seminal work of Goldreich, Juba and Sudan [266] takes a goal-oriented perspective on learning to communicate with an unknown agent; these ideas have been leveraged in modern wireless communication settings [70]. More generally, *learning to cooperate* will require distinguishing a cooperative agent from the other three categories, a problem that is wide open both theoretically and practically. Therefore, several open directions remain in better understanding cooperative (and not just decentralized) strategic behavior in the paradigm of learning and decision-making by automated agents.

# Bibliography


[1] R. W. Keener, *Theoretical Statistics: Topics for a Core Course*. Springer, 2011.

[2] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 10. Springer Series in Statistics New York, 2001, vol. 1.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] D. Kirk *et al.*, "NVIDIA CUDA software and GPU parallel computing architecture", in *ISMM*, vol. 7, 2007, pp. 103–104.

[5] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall Professional Technical Reference, 2002.

[6] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[8] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty", *Journal of Risk and uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[9] S. Mullainathan and R. H. Thaler, "Behavioral Economics", National Bureau of Economic Research, Tech. Rep., 2000.

[10] C. F. Camerer and G. Loewenstein, *Behavioral Economics: Past, Present, Future*. Princeton University Press, 2003.

[11] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.

[12] P. Milgrom and P. R. Milgrom, *Putting Auction Theory to Work*. Cambridge University Press, 2004.

[13] W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders", *The Journal of Finance*, vol. 16, no. 1, pp. 8–37, 1961.

[14] E. H. Clarke, "Multipart Pricing of Public Goods", *Public Choice*, vol. 11, no. 1, pp. 17–33, 1971.

[15] T. Groves, "Incentives in Teams", *Econometrica: Journal of the Econometric Society*, pp. 617–631, 1973.

[16] S. Li, "Obviously Strategy-Proof Mechanisms", *American Economic Review*, vol. 107, no. 11, pp. 3257–87, 2017.

[17] K. Leyton-Brown, P. Milgrom, and I. Segal, "Economics and Computer Science of a Radio Spectrum Reallocation", *Proceedings of the National Academy of Sciences*, vol. 114, no. 28, pp. 7202–7209, 2017.

[18] N. Newman, A. Fréchette, and K. Leyton-Brown, "Deep Optimization for Spectrum Repacking", *Communications of the ACM*, vol. 61, no. 1, pp. 97–104, 2017.

[19] *In the Matter of Expanding the Economic and Innovation Opportunities of Spectrum Through Incentive Auctions: Report and Order*, Jun. 2014. [Online]. Available: `https://apps.fcc.gov/edocs_public/attachmatch/FCC-14-50A1.pdf`.

[20] R. Johari, V. Kamble, and Y. Kanoria, "Matching While Learning", *arXiv preprint arXiv:1603.04549*, 2016.

[21] L. T. Liu, H. Mania, and M. Jordan, "Competing Bandits in Matching Markets", in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1618–1628.

[22] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[23] R. B. Myerson, "Optimal Auction Design", *Mathematics of Operations Research*, vol. 6, no. 1, pp. 58–73, 1981.

[24] B. Sivan, "Open Problems in the Display Ads Market Place", 2019. [Online]. Available: `https://simons.berkeley.edu/talks/tba-127`.

[25] J. Hartline, "Foundations of Non-truthful Mechanism Design", *21st ACM Conference on Economics and Computation (Tutorial)*, 2020. [Online]. Available: `https://sites.northwestern.edu/hartline/tutorial-foundations-of-non-truthful-mechanism-design/`.

[26] Y. Feng and J. D. Hartline, "An end-to-End Argument in Mechanism Design (Prior-independent Auctions for Budgeted Agents)", in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2018, pp. 404–415.

[27] M. Ostrovsky and M. Schwarz, "Reserve Prices in Internet Advertising Auctions: A Field Experiment", in *Proceedings of the 12th ACM Conference on Electronic Commerce*, 2011, pp. 59–60.

[28] S. R. Balseiro, O. Besbes, and G. Y. Weintraub, "Repeated Auctions with Budgets in Ad Exchanges: Approximations and Design", *Management Science*, vol. 61, no. 4, pp. 864–884, 2015.

[29] S. Balseiro, A. Kim, M. Mahdian, and V. Mirrokni, "Budget Management Strategies in Repeated Auctions", in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 15–23.

[30] S. R. Balseiro, V. S. Mirrokni, and R. P. Leme, "Dynamic Mechanisms with Martingale Utilities", *Management Science*, vol. 64, no. 11, pp. 5062–5082, 2018.

[31] R. Cole and T. Roughgarden, "The Sample Complexity of Revenue Maximization", in *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, 2014, pp. 243–252.

[32] J. H. Morgenstern and T. Roughgarden, "On the Pseudo-Dimension of Nearly Optimal Auctions", in *Advances in Neural Information Processing Systems*, 2015, pp. 136–144.

[33] M.-F. F. Balcan, T. Sandholm, and E. Vitercik, "Sample Complexity of Automated Mechanism Design", in *Advances in Neural Information Processing Systems*, 2016, pp. 2083–2091.

[34] M. Mohri and A. M. Medina, "Learning Algorithms for Second-Price Auctions with Reserve", *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2632–2656, 2016.

[35] J. Morgenstern and T. Roughgarden, "Learning Simple Auctions", in *Conference on Learning Theory*, 2016, pp. 1298–1318.

[36] V. Syrgkanis, "A Sample Complexity Measure with Applications to Learning Optimal Auctions", in *Advances in Neural Information Processing Systems*, 2017, pp. 5352–5359.

[37] Y. A. Gonczarowski and S. M. Weinberg, "The Sample Complexity of up-to-epsilon Multi-Dimensional Revenue Maximization", in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2018, pp. 416–426.

[38] Z. Huang, Y. Mansour, and T. Roughgarden, "Making the Most of your Samples", *SIAM Journal on Computing*, vol. 47, no. 3, pp. 651–674, 2018.

[39] M. Dudik, N. Haghtalab, H. Luo, R. E. Schapire, V. Syrgkanis, and J. W. Vaughan, "Oracle-efficient online learning and auction design", in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2017, pp. 528–539.

[40] A. Acquisti and H. R. Varian, "Conditioning Prices on Purchase History", *Marketing Science*, vol. 24, no. 3, pp. 367–381, 2005.

[41] D. Fudenberg and J. M. Villas-Boas, "Behavior-based Price Discrimination and Customer Recognition", *Handbook on Economics and Information Systems* , vol. 1, pp. 377–436, 2006.

[42] K. Amin, A. Rostamizadeh, and U. Syed, "Learning Prices for Repeated Auctions with Strategic Buyers", in *Advances in Neural Information Processing Systems*, 2013, pp. 1169–1177.

[43] N. Golrezaei, A. Javanmard, and V. Mirrokni, "Dynamic Incentive-aware Learning: Robust Pricing in Contextual Auctions", in *Advances in Neural Information Processing Systems*, 2019, pp. 9759–9769.

[44] E. Guerre, I. Perrigne, and Q. Vuong, "Optimal Nonparametric Estimation of First-Price Auctions", *Econometrica*, vol. 68, no. 3, pp. 525–574, 2000.

[45] S. Chawla, J. D. Hartline, and D. Nekipelov, "Mechanism Redesign", *arXiv preprint arXiv:1708.04699*, 2017.

[46] J. Hartline and S. Taggart, "Sample Complexity for Non-Truthful Mechanisms", in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019, pp. 399–416.

[47] S. Agrawal, C. Daskalakis, V. Mirrokni, and B. Sivan, "Robust Repeated Auctions under Heterogeneous Buyer Behavior", in *19th ACM Conference on Economics and Computation*, 2018.

[48] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal", *IEEE personal communications*, vol. 6, no. 4, pp. 13–18, 1999.

[49] "WSRD workshop VIII: Wireless spectrum sharing: Enforcement Frameworks, Technology and Research and Development", [Online]. Available: `https://www.nitrd.gov/nitrdgroups/index.php?title=WSRD_Workshop_VIII_Wireless_Spectrum_Sharing`.

[50] *Realizing the Full Potential of Government-Held Spectrum to Spur Economic Growth*. [Online]. Available: `https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast_spectrum_report_final_july_20_2012.pdf`.

[51] R. H. Coase, "The Federal Communications Commission", *Journal of Law and Economics*, vol. 56, no. 4, pp. 879–915, 1959.

[52] V. Muthukumar, A. Daruna, V. Kamble, K. Harrison, and A. Sahai, "Whitespaces after the usa's tv incentive auction: A spectrum reallocation case study", in *2015 IEEE International Conference on Communications (ICC)*, IEEE, 2015, pp. 7582–7588.

[53] E. Noam, "Spectrum auctions: Yesterday's heresy, today's orthodoxy, tomorrow's anachronism. Taking the next step to open spectrum access", *The Journal of Law and Economics*, vol. 41, no. S2, pp. 765–790, 1998.

[54] J. Huang, R. A. Berry, and M. L. Honig, "Auction-based spectrum sharing", *Mobile Networks and Applications*, vol. 11, no. 3, pp. 405–418, 2006.

[55] K. L. Harrison, "A quantitative approach to wireless spectrum regulation", PhD thesis, UC Berkeley, 2015.

[56] D. Niyato and E. Hossain, "Competitive pricing for spectrum sharing in cognitive radio networks: Dynamic game, inefficiency of Nash equilibrium, and collusion", *IEEE journal on selected areas in communications*, vol. 26, no. 1, 2008.

[57] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios", in *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, Ieee, vol. 1, 2004, pp. 772–776.

[58] D. Čabrić, S. M. Mishra, D. Willkomm, R. Brodersen, and A. Wolisz, "A cognitive radio approach for usage of virtual unlicensed spectrum", in *14th IST mobile and wireless communications summit*, Citeseer, 2005.

[59] S. M. Mishra, D. Cabric, C. Chang, D. Willkomm, B. Van Schewick, A. Wolisz, and R. W. Brodersen, "A real time cognitive radio testbed for physical and link layer experiments", in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005.*, IEEE, 2005, pp. 562–567.

[60] S. M. Mishra, A. Sahai, and R. W. Brodersen, "Cooperative sensing among cognitive radios", in *Communications, 2006. ICC'06. IEEE International conference on*, IEEE, vol. 4, 2006, pp. 1658–1663.

[61] R. Tandra and A. Sahai, "SNR walls for signal detection", *IEEE Journal of selected topics in Signal Processing*, vol. 2, no. 1, pp. 4–17, 2008.

[62] R. Tandra, S. M. Mishra, and A. Sahai, "What is a spectrum hole and what does it take to recognize one?", *Proceedings of the IEEE*, vol. 97, no. 5, pp. 824–848, 2009.

[63] R. Etkin, A. Parekh, and D. Tse, "Spectrum sharing for unlicensed bands", *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 517–528, 2007.

[64] K. A. Woyach, "Building trust into light-handed regulations for cognitive radio", PhD thesis, UC Berkeley, 2013.

[65] K. Woyach and A. Sahai, "The need for a new model of trust in spectrum and the case for spectrum jails", in *Dynamic Spectrum Access Networks (DYSPAN), 2014 IEEE International Symposium on*, IEEE, 2014, pp. 427–438.

[66] G. S. Becker, "Crime and Punishment: An Economic Approach", *Journal of Political Economy*, vol. 76, no. 2, pp. 169–217, 1968.

[67] R. A. Posner, *Economic analysis of law*. Little Brown and Company, 1973.

[68] ——, "An economic theory of the criminal law", *Columbia Law Review*, vol. 85, no. 6, pp. 1193–1231, 1985.

[69] V. Muthukumar and A. Sahai, "Fundamental limits on ex-post enforcement and implications for spectrum rights", *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 491–504, 2017.

[70] A. Sahai, J. Sanz, V. Subramanian, C. Tran, and K. Vodrahalli, "Blind interactive learning of modulation schemes: Multi-agent cooperation without co-design", *IEEE Access*, vol. 8, pp. 63 790–63 820, 2020.

[71] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*. Now Publishers Inc, 2004.

[72] L. Ljung, "System Identification", *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–19, 1999.

[73] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2002.

[74] B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

[75] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[76] R. B. Myerson, *Game Theory*. Harvard University Press, 2013.

[77] R. J. Aumann, M. Maschler, and R. E. Stearns, *Repeated Games with Incomplete Information*. MIT press, 1995.

[78] D. Fudenberg and D. K. Levine, *The theory of learning in games*. MIT Press, 1998, vol. 2.

[79] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Nonstochastic Multiarmed Bandit Problem", *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[80] J. Hannan, "Approximation to Bayes risk in repeated play", *Contributions to the Theory of Games*, vol. 3, pp. 97–139, 1957.

[81] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[82] D. Blackwell, "An analog of the minimax theorem for vector payoffs", *Pacific Journal of Mathematics*, vol. 6, no. 1, pp. 1–8, 1956.

[83] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm", *Information and computation*, vol. 108, no. 2, pp. 212–261, 1994.

[84] N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth, "How to use expert advice", *Journal of the ACM (JACM)*, vol. 44, no. 3, pp. 427–485, 1997.

[85] S. Hart, "Adaptive heuristics", *Econometrica*, vol. 73, no. 5, pp. 1401–1430, 2005.

[86] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium", *SIAM Journal on Computing*, vol. 39, no. 1, pp. 195–259, 2009.

[87] S. Hart and A. Mas-Colell, "Uncoupled dynamics do not lead to Nash equilibrium", *American Economic Review*, vol. 93, no. 5, pp. 1830–1836, 2003.

[88] V. Muthukumar, M. Ray, A. Sahai, and P. Bartlett, "Best of many worlds: Robust model selection for online supervised learning", in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 3177–3186.

[89] N. Chatterji, V. Muthukumar, and P. Bartlett, "OSOM: A simultaneously optimal algorithm for multi-armed and linear contextual bandits", in *International Conference on Artificial Intelligence and Statistics*, 2020, pp. 1844–1854.

[90] V. Muthukumar and A. Sahai, "Robust commitments and partial reputation", *arXiv preprint arXiv:1905.11555*, 2019.

[91] ——, "Robust commitments and partial reputation", in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019, pp. 637–638.

[92] V. Muthukumar, S.-R. Phade, and A. Sahai, "On the divergence of mixed strategies arising from no-regret learning", *preprint*, 2020.

[93] M. Hutter and J. Poland, "Adaptive online prediction by following the perturbed leader", *Journal of Machine Learning Research*, vol. 6, no. Apr, pp. 639–660, 2005.

[94] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz, "Improved second-order bounds for prediction with expert advice", *Machine Learning*, vol. 66, no. 2-3, pp. 321–352, 2007.

[95] T. V. Erven, W. M. Koolen, S. D. Rooij, and P. Grünwald, "Adaptive hedge", in *Advances in Neural Information Processing Systems*, 2011, pp. 1656–1664.

[96] N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz, "Mirror descent meets fixed share (and feels no regret)", in *Advances in Neural Information Processing Systems*, 2012, pp. 980–988.

[97] A. Rakhlin and K. Sridharan, "Online learning with predictable sequences", 2013.

[98] S. De Rooij, T. Van Erven, P. D. Grünwald, and W. M. Koolen, "Follow the leader if you can, hedge if you must.", *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1281–1316, 2014.

[99] F. Orabona, "Simultaneous model selection and optimization through parameter-free stochastic learning", in *Advances in Neural Information Processing Systems*, 2014, pp. 1116–1124.

[100] W. M. Koolen and T. Van Erven, "Second-order quantile methods for experts and combinatorial games", in *Conference on Learning Theory*, 2015, pp. 1155–1175.

[101] H. Luo and R. E. Schapire, "Achieving all with no parameters: Adanormalhedge", in *Conference on Learning Theory*, 2015, pp. 1286–1304.

[102] F. Orabona and D. Pál, "Coin betting and parameter-free online learning", in *Advances in Neural Information Processing Systems*, 2016, pp. 577–585.

[103] T. van Erven and W. M. Koolen, "Metagrad: Multiple learning rates in online learning", in *Advances in Neural Information Processing Systems*, 2016, pp. 3666–3674.

[104] D. J. Foster, S. Kale, M. Mohri, and K. Sridharan, "Parameter-free online learning via model selection", in *Advances in Neural Information Processing Systems*, 2017, pp. 6022–6032.

[105] Z. Mhammedi, W. M. Koolen, and T. Van Erven, "Lipschitz Adaptivity with Multiple Learning rates in Online Learning", in *Conference on Learning Theory*, 2019, pp. 2490–2511.

[106] I. Guyon, V. Vapnik, B. Boser, L. Bottou, and S. A. Solla, "Structural risk minimization for character recognition", in *Advances in Neural Information Processing Systems*, 1992, pp. 471–479.

[107] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation", *Machine Learning*, vol. 48, no. 1-3, pp. 85–113, 2002.

[108] P. Massart, *Concentration inequalities and model selection*. Springer, 2007, vol. 6.

[109] D. P. Helmbold and R. E. Schapire, "Predicting nearly as well as the best pruning of a decision tree", *Machine Learning*, vol. 27, no. 1, pp. 51–68, 1997.

[110] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences", *IEEE transactions on Information Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.

[111] A. B. Tsybakov *et al.*, "Optimal aggregation of classifiers in statistical learning", *The Annals of Statistics*, vol. 32, no. 1, pp. 135–166, 2004.

[112] W. M. Koolen, T. Van Erven, and P. Grünwald, "Learning the learning rate for prediction with expert advice", in *Advances in Neural Information Processing Systems*, 2014, pp. 2294–2302.

[113] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding", *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[114] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin", in *Proceedings of the Ninth Annual Conference on Computational learning theory*, 1996, pp. 89–98.

[115] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL", *IEEE Transactions on Information theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

[116] N. Gravin, Y. Peres, and B. Sivan, "Tight lower bounds for multiplicative weights algorithmic families", in *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[117] V. N. Vapnik, "An overview of statistical learning theory", *IEEE transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[118] P. L. Bartlett, O. Bousquet, S. Mendelson, *et al.*, "Local rademacher complexities", *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[119] H. Akaike, "Information theory and an extension of the maximum likelihood principle", in *Selected Papers of Hirotugu Akaike*, Springer, 1998, pp. 199–213.

[120] M. Stone, "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[121] S. Geisser, "The predictive sample reuse method with applications", *Journal of the American statistical Association*, vol. 70, no. 350, pp. 320–328, 1975.

[122] S. Shalev-Shwartz *et al.*, "Online learning and online convex optimization", *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[123] W. M. Koolen, P. Grünwald, and T. van Erven, "Combining adversarial guarantees and stochastic fast rates in online learning", in *Advances in Neural Information Processing Systems*, 2016, pp. 4457–4465.

[124] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization", *arXiv preprint arXiv:1611.03530*, 2016.

[125] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[126] S. Dudoit and M. J. van der Laan, "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment", *Statistical Methodology*, vol. 2, no. 2, pp. 131–154, 2005.

[127] E. LeDell, M. Petersen, and M. van der Laan, "Computationally efficient confidence intervals for cross-validated area under the roc curve estimates", *Electronic Journal of Statistics*, vol. 9, no. 1, p. 1583, 2015.

[128] M. Austern and W. Zhou, "Asymptotics of Cross-Validation", *arXiv:2001.11111*, 2020.

[129] P. Bayle, A. Bayle, L. Janson, and L. Mackey, "Cross-validation Confidence Intervals for Test Error", *arXiv preprint arXiv:2007.12671*, 2020.

[130] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties", *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 653–664, 1995.

[131] E. Hazan and T. Koren, "The computational power of optimization in online learning", in *Proceedings of the Forty-Eighth annual ACM Symposium on Theory of Computing*, 2016, pp. 128–141.

[132] A. Kalai and S. Vempala, "Efficient algorithms for online decision problems", *Journal of Computer and System Sciences*, vol. 71, no. 3, pp. 291–307, 2005.

[133] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression", *Proceedings of the National Academy of Sciences*, 2020. eprint: `https://www.pnas.org/content/early/2020/04/22/1907378117.full.pdf`. [Online]. Available: `https://www.pnas.org/content/early/2020/04/22/1907378117`.

[134] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features", *arXiv preprint arXiv:1903.07571*, 2019.

[135] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high dimensional ridgeless least squares interpolation", *arXiv preprint arXiv:1903.08560*, 2019.

[136] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression", *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[137] K. Marton *et al.*, "Bounding d-distance by informational divergence: A method to prove measure concentration", *The Annals of Probability*, vol. 24, no. 2, pp. 857–866, 1996.

[138] K. Marton, "A simple proof of the blowing-up lemma (corresp.)", *IEEE Transactions on Information Theory*, vol. 32, no. 3, pp. 445–446, 1986.

[139] L. Lai, H. Jiang, and H. V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework", in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, IEEE, 2008, pp. 98–102.

[140] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players", *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[141] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret", *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[142] P. Whittle, "Multi-armed bandits and the Gittins index", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 143–149, 1980.

[143] J. N. Tsitsiklis, "A short proof of the Gittins index theorem", *The Annals of Applied Probability*, pp. 194–199, 1994.

[144] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem", in *Conference on Learning Theory*, 2012, pp. 39–1.

[145] ——, "Thompson sampling for contextual bandits with linear payoffs", in *International Conference on Machine Learning*, 2013, pp. 127–135.

[146] ——, "Further optimal regret bounds for thompson sampling", in *Artificial Intelligence and Statistics*, 2013, pp. 99–107.

[147] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling", *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

[148] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples", *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[149] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules", *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[150] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem", *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[151] J.-Y. Audibert and S. Bubeck, "Regret bounds and minimax policies under partial monitoring", *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2785–2836, 2010.

[152] R. Degenne and V. Perchet, "Anytime optimal algorithms in stochastic multi-armed bandits", in *Proceedings of the International Conference on Machine Learning*, 2016.

[153] M. Woodroofe, "A one-armed bandit problem with a concomitant variable", *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 799–806, 1979.

[154] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation", in *Proceedings of the International conference on World Wide Web*, ACM, 2010.

[155] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, S. Sen, and A. Slivkins, "Making contextual decisions with low technical debt", *arXiv preprint arXiv:1606.03966*, 2016.

[156] A. Tewari and S. A. Murphy, "From ads to interventions: Contextual bandits in mobile health", in *Mobile Health*, Springer, 2017, pp. 495–517.

[157] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions", in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.

[158] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits", in *Proceedings of the Advances in Neural Information Processing Systems*, 2011.

[159] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information", in *Proceedings of the Advances in Neural Information Processing Systems*, 2008.

[160] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits", in *Proceedings of the International Conference on Machine Learning*, 2014.

[161] A. Rakhlin and K. Sridharan, "BISTRO: An efficient relaxation-based method for contextual bandits", in *Proceedings of the International Conference on Machine Learning*, 2016.

[162] V. Syrgkanis, A. Krishnamurthy, and R. Schapire, "Efficient algorithms for adversarial contextual learning", in *Proceedings of the International Conference on Machine Learning*, 2016.

[163] V. Syrgkanis, H. Luo, A. Krishnamurthy, and R. Schapire, "Improved regret bounds for oracle-based adversarial contextual bandits", in *Proceedings of the Advances in Neural Information Processing Systems*, 2016.

[164] D. Foster and A. Krishnamurthy, "Contextual bandits with surrogate losses: Margin bounds and efficient algorithms", in *Proceedings of the Advances in Neural Information Processing Systems*, 2018.

[165] A. Agarwal, H. Luo, B. Neyshabur, and R. Schapire, "Corralling a Band of Bandit Algorithms", in *Proceedings of the Conference on Learning Theory*, 2017.

[166] D. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos, "Learning in games: Robustness of fast convergence", in *Proceedings of the Advances in Neural Information Processing Systems*, 2016.

[167] T. Lattimore, "The pareto regret frontier for bandits", in *Proceedings of the Advances in Neural Information Processing Systems*, 2015.

[168] S. Bubeck and A. Slivkins, "The best of both worlds: Stochastic and adversarial bandits", in *Proceedings of the Conference on Learning Theory*, 2012.

[169] D. Foster, A. Krishnamurthy, and H. Luo, "Model selection for contextual bandits", *Proceedings of the Advances in Neural Information Processing Systems*, 2019.

[170] H. Bastani, M. Bayati, and K. Khosravi, "Mostly exploration-free algorithms for contextual bandits", *Management Science*, 2020.

[171] S. Kannan, J. Morgenstern, A. Roth, B. Waggoner, and Z. S. Wu, "A smoothed analysis of the greedy algorithm for the linear contextual bandit problem", in *Proceedings of the Advances in Neural Information Processing Systems*, 2018.

[172] M. Raghavan, A. Slivkins, J. V. Wortman, and Z. S. Wu, "The Externalities of Exploration and How Data Diversity Helps Exploitation", in *Proceedings of the Conference On Learning Theory*, 2018.

[173] D. J. Foster, A. Krishnamurthy, and H. Luo, "Open Problem: Model Selection for Contextual Bandits", in *Conference on Learning Theory*, 2020, pp. 3842–3846.

[174] A. Pacchiano, M. Phan, Y. Abbasi-Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari, "Model selection in contextual stochastic bandit problems", *arXiv preprint arXiv:2003.01704*, 2020.

[175] Y. Abbasi-Yadkori, A. Pacchiano, and M. Phan, "Regret Balancing for Bandit and RL Model Selection", *arXiv preprint arXiv:2006.05491*, 2020.

[176] N. Wanigasekara and C. L. Yu, "Nonparametric Contextual Bandits in an Unknown Metric Space", in *Proceedings of the Advances in Neural Information Processing Systems*, 2019.

[177] A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang, "Contextual Bandits with Continuous Actions: Smoothing, Zooming, and Adapting", in *Proceedings of the Conference On Learning Theory*, 2019.

[178] J. Zimmert and Y. Seldin, "An optimal algorithm for stochastic and adversarial bandits", in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 467–475.

[179] J. Tropp, "Freedman's inequality for matrix martingales", *Electronic Communications in Probability*, vol. 16, pp. 262–270, 2011.

[180] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2019.

[181] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus, "Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games", in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 895–902.

[182] D. M. Kreps and R. Wilson, "Reputation and imperfect information", *Journal of Economic Theory*, vol. 27, no. 2, pp. 253–279, 1982.

[183] P. Milgrom and J. Roberts, "Predation, reputation, and entry deterrence", *Journal of Economic Theory*, vol. 27, no. 2, pp. 280–312, 1982.

[184] D. Fudenberg and D. K. Levine, "Reputation and equilibrium selection in games with a patient player", *Econometrica: Journal of the Econometric Society*, pp. 759–778, 1989.

[185] ——, "Maintaining a reputation when strategies are imperfectly observed", *The Review of Economic Studies*, vol. 59, no. 3, pp. 561–579, 1992.

[186] O. Gossner, "Simple Bounds on the Value of a Reputation", *Econometrica*, vol. 79, no. 5, pp. 1627–1641, 2011.

[187] B. Von Stengel and S. Zamir, "Leadership games with convex strategy sets", *Games and Economic Behavior*, vol. 69, no. 2, pp. 446–457, 2010.

[188] E. Kamenica and M. Gentzkow, "Bayesian persuasion", *The American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.

[189] H. Von Stackelberg, *Marktform und gleichgewicht*. J. springer, 1934.

[190] T. C. Schelling, *The strategy of conflict*. Harvard university press, 1980.

[191] R. Selten, "The chain store paradox", *Theory and Decision*, vol. 9, no. 2, pp. 127–159, 1978.

[192] J. Ely, D. Fudenberg, and D. K. Levine, "When is reputation bad?", *Games and Economic Behavior*, vol. 63, no. 2, pp. 498–526, 2008.

[193] S. Sorin, "Merging, reputation, and repeated games with incomplete information", *Games and Economic Behavior*, vol. 29, no. 1-2, pp. 274–308, 1999.

[194] G. J. Mailath and L. Samuelson, "Reputations in repeated games", in *Handbook of Game Theory with Economic Applications*, vol. 4, Elsevier, 2015, pp. 165–238.

[195] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[196] R. Selten, "Reexamination of the perfectness concept for equilibrium points in extensive games", *International journal of game theory*, vol. 4, no. 1, pp. 25–55, 1975.

[197] E. Van Damme and S. Hurkens, "Games with imperfectly observable commitment", *Games and Economic Behavior*, vol. 21, no. 1-2, pp. 282–308, 1997.

[198] K. Bagwell, "Commitment and observability in games", *Games and Economic Behavior*, vol. 8, no. 2, pp. 271–280, 1995.

[199] J. C. Ely and J. Välimäki, "Bad reputation", *The Quarterly Journal of Economics*, vol. 118, no. 3, pp. 785–814, 2003.

[200] J. C. Harsanyi, "Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model", *Management science*, vol. 14, no. 3, pp. 159–182, 1967.

[201] A. Blum, N. Haghtalab, and A. D. Procaccia, "Lazy Defenders Are Almost Optimal against Diligent Attackers.", in *AAAI*, 2014, pp. 573–579.

[202] S. Morris, "The common prior assumption| n| economic theory", *Economics and Philosophy*, vol. 11, pp. 227–253, 1995.

[203] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards", *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[204] R. Agrawal, "Sample mean based index policies with O (log n) regret for the multi-armed bandit problem", *Advances in Applied Probability*, pp. 1054–1078, 1995.

[205] S. Bubeck, N. Cesa-Bianchi, *et al.*, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems", *Foundations and Trends® in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[206] P.-A. Coquelin and R. Munos, "Bandit algorithms for tree search", in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, 2007, pp. 67–74.

[207] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized markov decision processes", in *Conference on Learning Theory*, 2015, pp. 861–898.

[208] D. P. Foster and R. V. Vohra, "Calibrated learning and correlated equilibrium", *Games and Economic Behavior*, vol. 21, no. 1-2, p. 40, 1997.

[209] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium", *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.

[210] D. Blackwell *et al.*, "Controlled random walks", in *Proceedings of the international congress of mathematicians*, vol. 3, 1954, pp. 336–338.

[211] Y. Freund and R. E. Schapire, "Adaptive game playing using multiplicative weights", *Games and Economic Behavior*, vol. 29, no. 1-2, pp. 79–103, 1999.

[212] Y. Deng, J. Schneider, and B. Sivan, "Strategizing against no-regret learners", in *Advances in Neural Information Processing Systems*, 2019, pp. 1579–1587.

[213] B. Von Stengel and S. Zamir, "Leadership with commitment to mixed strategies", Citeseer, Tech. Rep., 2004.

[214] V. Conitzer and T. Sandholm, "Computing the optimal strategy to commit to", in *Proceedings of the 7th ACM Conference on Electronic Commerce*, ACM, 2006, pp. 82–90.

[215] B. An, D. Kempe, C. Kiekintveld, E. Shieh, S. Singh, M. Tambe, and Y. Vorobeychik, "Security games with limited surveillance", in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI Press, 2012, pp. 1241–1248.

[216] T. Roughgarden, "Stackelberg scheduling strategies", *SIAM Journal on Computing*, vol. 33, no. 2, pp. 332–350, 2004.

[217] V. P. Crawford and J. Sobel, "Strategic information transmission", *Econometrica: Journal of the Econometric Society*, pp. 1431–1451, 1982.

[218] J. C. Ely, "Beeps", *The American Economic Review*, vol. 107, no. 1, pp. 31–53, 2017.

[219] V. Conitzer, "On Stackelberg mixed strategies", *Synthese*, vol. 193, no. 3, pp. 689–703, 2016.

[220] E. Shieh, B. An, R. Yang, M. Tambe, C. Baldwin, J. DiRenzo, B. Maule, and G. Meyer, "Protect: A deployed game theoretic system to protect the ports of the United States", in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 13–20.

[221] S. Dughmi, D. Kempe, and R. Qiang, "Persuasion with limited communication", in *Proceedings of the 2016 ACM Conference on Economics and Computation*, ACM, 2016, pp. 663–680.

[222] S. Dughmi, N. Immorlica, and A. Roth, "Constrained signaling in auction design", in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics, 2014, pp. 1341–1357.

[223] D. Korzhyk, V. Conitzer, and R. Parr, "Solving Stackelberg games with uncertain observability", in *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 1013–1020.

[224] J. Pita, M. Jain, M. Tambe, F. Ordóñez, and S. Kraus, "Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition", *Artificial Intelligence*, vol. 174, no. 15, pp. 1142–1171, 2010.

[225] A. Blum, N. Haghtalab, and A. D. Procaccia, "Learning optimal commitment to overcome insecurity", in *Advances in Neural Information Processing Systems*, 2014, pp. 1826–1834.

[226] N. Haghtalab, F. Fang, T. H. Nguyen, A. Sinha, A. D. Procaccia, and M. Tambe, "Three strategies to success: Learning adversary models in security games", in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 308–314.

[227] M.-F. Balcan, A. Blum, N. Haghtalab, and A. D. Procaccia, "Commitment without regrets: Online learning in Stackelberg security games", in *Proceedings of the sixteenth ACM conference on economics and computation*, ACM, 2015, pp. 61–78.

[228] Y. Mansour, A. Slivkins, V. Syrgkanis, and Z. S. Wu, "Bayesian Exploration: Incentivizing Exploration in Bayesian Games", in *Proceedings of the 2016 ACM Conference on Economics and Computation*, ACM, 2016, pp. 661–661.

[229] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[230] A. C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variates", *Transactions of the american mathematical society*, vol. 49, no. 1, pp. 122–136, 1941.

[231] C.-G. Esseen, *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell, 1942.

[232] D. P. Bertsekas, *Dynamic programming and optimal control*, 2. Athena scientific Belmont, MA, 1995, vol. 1.

[233] D. Kreps, "A theory of Learning, Experimentation, and Equilibrium in Games", Technical Report, mimeo 1988. and,"Learning in extensive-form games I. Self ..., Tech. Rep., 2011.

[234] R. Kannan and H. Narayanan, "Random walks on polytopes and an affine interior point method for linear programming", *Mathematics of Operations Research*, vol. 37, no. 1, pp. 1–20, 2012.

[235] L. Devroye, "The equivalence of weak, strong and complete convergence in l1 for kernel density estimates", *The Annals of Statistics*, pp. 896–904, 1983.

[236] G. W. Brown, "Iterative solution of games by fictitious play", *Activity analysis of Production and Allocation*, vol. 13, no. 1, pp. 374–376, 1951.

[237] J. Robinson, "An iterative method of solving a game", *Annals of mathematics*, pp. 296–301, 1951.

[238] D. Fudenberg and D. Levine, "Learning in games", *European economic review*, vol. 42, no. 3-5, pp. 631–639, 1998.

[239] N. Littlestone, M. K. Warmuth, *et al.*, *The weighted majority algorithm*. University of California, Santa Cruz, Computer Research Laboratory, 1989.

[240] A. Blum and Y. Mansour, "From external to internal regret", *Journal of Machine Learning Research*, vol. 8, no. Jun, pp. 1307–1324, 2007.

[241] D. P. Foster and R. V. Vohra, "Asymptotic calibration", *Biometrika*, vol. 85, no. 2, pp. 379–390, 1998.

[242] C. Daskalakis and I. Panageas, "Last-iterate convergence: Zero-sum games and constrained min-max optimization", *arXiv preprint arXiv:1807.04252*, 2018.

[243] J. P. Bailey and G. Piliouras, "Multiplicative weights update in zero-sum games", in *Proceedings of the 2018 ACM Conference on Economics and Computation*, ACM, 2018, pp. 321–338.

[244] A. Calvó-Armengol, "The set of correlated equilibria of 2x2 games", *Working Paper 79, Barcelona Graduate School of Eco- nomics, Barcelona, Spain.*, 2006.

[245] S. R. Phade and V. Anantharam, "On the geometry of Nash and Correlated Equilibria with Cumulative Prospect Theoretic Preferences", *Decision Analysis*, vol. 16, no. 2, pp. 142–156, 2019.

[246] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training gans with optimism.", in *International Conference on Learning Representations (ICLR 2018)*, 2018.

[247] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras, "Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile", in *International Conference on Learning Representations*, 2018.

[248] T. Liang and J. Stokes, "Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks", in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 907–915.

[249] J. Abernethy, K. A. Lai, and A. Wibisono, "Last-iterate convergence rates for min-max optimization", *arXiv preprint arXiv:1906.02027*, 2019.

[250] Q. Lei, S. G. Nagarajan, I. Panageas, and X. Wang, "Last iterate convergence in no-regret learning: Constrained min-max optimization for convex-concave landscapes", *arXiv preprint arXiv:2002.06768*, 2020.

[251] S. Rakhlin and K. Sridharan, "Optimization, learning, and games with predictable sequences", in *Advances in Neural Information Processing Systems*, 2013, pp. 3066–3074.

[252] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, "Fast convergence of regularized learning in games", in *Advances in Neural Information Processing Systems*, 2015, pp. 2989–2997.

[253] A. S. Nemirovsky and D. B. Yudin, *Problem complexity and method efficiency in optimization.* Wiley, 1983.

[254] D. P. Foster and S. Hart, "Smooth calibration, leaky forecasts, finite recall, and nash dynamics", *Games and Economic Behavior*, vol. 109, pp. 271–293, 2018.

[255] W. Feller, "An introduction to probability and statistics, vol. 1", *New York*, vol. 6, no. 31, p. 138, 1957.

[256] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, *et al.*, "Boosting the margin: A new explanation for the effectiveness of voting methods", *The annals of statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[257] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning.", in *ICLR (Workshop)*, 2015.

[258] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and double descent curve", *arXiv preprint arXiv:1908.05355*, 2019.

[259] P. P. Mitra, "Understanding overfitting peaks in generalization error: Analytical risk curves for l2 and l1 penalized interpolation", *arXiv preprint arXiv:1906.03667*, 2019.

[260] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional binary linear classification", *arXiv preprint arXiv:1911.05822*, 2019.

[261] A. Montanari, F. Ruan, Y. Sohn, and J. Yan, "The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime", *arXiv preprint arXiv:1911.01544*, 2019.

[262] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. Hsu, and A. Sahai, "Classification vs regression in overparameterized regimes: Does the loss function matter?", *arXiv preprint arXiv:2005.08054*, 2020.

[263] N. S. Chatterji and P. M. Long, "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime", *arXiv preprint arXiv:2004.12019*, 2020.

[264] S. Bubeck, Y. Li, Y. Peres, and M. Sellke, "Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without", in *Conference on Learning Theory*, 2020, pp. 961–987.

[265] S. Bubeck and T. Budzinski, "Coordination without communication: Optimal regret in two players multi-armed bandits", *arXiv preprint arXiv:2002.07596*, 2020.

[266] O. Goldreich, B. Juba, and M. Sudan, "A Theory of Goal-Oriented Communication", *Journal of the ACM (JACM)*, vol. 59, no. 2, pp. 1–65, 2012.