Incentive Design in the Machine Learning Age

A dissertation presented

by

Tao Lin

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Dissertation Committee

Yiling Chen (Advisor)

Ariel D. Procaccia

David C. Parkes

Harvard University
Cambridge, Massachusetts
May 2025

Dissertation Advisor: Yiling Chen Tao Lin

Incentive Design in the Machine Learning Age

Abstract

This dissertation investigates the design of incentives in multi-agent systems where traditional assumptions – such as fully rational agents and omniscient principals – do not hold. As real-world systems increasingly involve learning-based decision-makers, either human or algorithmic, this work explores how learning alters the landscape of incentive design. The dissertation is organized into three parts.

The first part focuses on **incentive design by learning principals**, specifically in information and mechanism design settings. For information design, this part introduces novel algorithms that allow a principal to learn an agent's non-Bayesian belief updating process, such as a subjective prior or cognitive bias, via strategic interaction with the agent. For mechanism design, this part examines how a coordinator can learn to compute Bayes correlated equilibria in non-truthful auctions using limited samples of agent types.

The second part studies **incentive design for learning agents**, who are modeled as boundedly rational learners rather than best responders. This part first presents, for a general class of principal-agent problems, a reduction from no-regret learning agents to approximately best-responding agents, enabling a precise analysis of the principal's performance. It then characterizes the convergence properties of multi-agent learning in first-price auction games, identifying when convergence to equilibrium is possible.

The third part explores **incentive issues in deployed machine learning systems**, with a case study on recommender systems. It demonstrates that the strategic behaviors by content creators can exacerbate polarization, even under diversity-promoting algorithms, and proposes alternative algorithmic designs that mitigate these effects.

Collectively, this dissertation lays foundational insights for designing systems that are robust to the incentives of learning-based, data-driven participants.

Acknowledgments

First and foremost, I owe my deepest thank to my advisor, Yiling Chen. Your guidance has shaped not just my research, but my life. You helped me develop a sense of research taste, find direction amid uncertainty, and explore seemingly unrelated problems that ultimately converge to one point. Those countless conversations where we dove into big ideas and fine details will be my invaluable memories. I really appreciate your warning that *"every PhD student will be depressed at some time in five years"* and helping me overcome those times. Your support beyond research meant more than words can express. I feel incredibly lucky to have had you as my advisor.

I am also deeply grateful to Ariel Procaccia and David Parkes, who have been with me since my qualifying exam and supporting me throughout these years. Our annual meetings were more than just check-ins – they were moments of clarity and encouragement that helped me shape my research theme, navigate career decisions, and stay grounded in what mattered most. Your thoughtful guidance and generous advice were instrumental to my dissertation. And I really enjoyed the "tough" questions you asked in my defense.

I would also like to thank Milind Tambe, another member of my qualifying exam committee, who constantly reminded me of the importance of connecting theory to practice. His perspective encouraged me to think beyond abstraction and consider the social impact of theoretical research – an influence that will stay with me well beyond my PhD.

I cannot forget to thank Yannai Gonczarowski, whose course *A Computer Science Toolbox for Modern Economic Theory* inspired my first PhD publication. I highly recommend this course to any Boston-based student who happens to be reading this paragraph.

Let me use this chance to also express gratitude to my undergraduate advisor at Peking University, Xiaotie Deng, who introduced me to the field of algorithmic game theory. I was drawn to this field not just by the elegance of the theory, but by Xiaotie's passion in the classroom, which sparked my curiosity and eventually led me to pursue this journey.

I am also thankful to my mentors from industry research internships, Yang Liu at ByteDance and Christopher Liaw at Google. Their support for me pursuing theoretical research – something not typical in industry – meant a great deal. Their hands-on advising, thoughtful feedback, and openness to deep technical exploration made those experiences truly valuable.

Among my collaborators, I am especially grateful to Ron Lavi and Haifeng Xu. My long-term collaborations with Ron and Haifeng, with their continued mentorship and encouragement, have made them almost my co-advisors. I have also learned so much from the following list of wonderful researchers (in reverse alphabetical order): Weiqiang Zheng, Xiaoying Zhang, Jiahao Zhang, Brian Zhang, Xiang Yan, Tao Xiao, Weiwen Wang, Tonghan Wang, Yifeng Teng, Itai Shapira, Tuomas Sandholm, Manon Revel, Aaditya Ramdas, Qi Qi, Hongyi Ling, Ce Li, Kun Jin, Xinyan Hu, Safwan Hossain, Daniel Halpern, Yotam Gafni, Hu Fu, Zhe Feng, Andrew Estornell, Yang Cai, and Kshipra Bhawalkar. As said by more than one people above, *"research is often not about what you do, but who you work with."*

The friends I met at Harvard provided me an uncountably infinite amount of support, both inside and outside research. Thank you! Zhou, Sai, Jamie, Nicole, Gary, Hugh, Ke, Shirley, Sara, Jeff, Sadie, Shi, Shixuan, Jiayuan, Lily, Jessie, Xintong, Matheus, Fang-Yi, and Siddarth. Also, a special thank to Yiding Feng's help in my job search!

As expected, the final paragraph of acknowledgement goes to my parents, for their endless love and patience. Thank you for tolerating the long stretches when I didn't reply to your messages, for trusting me to follow my own path, and for supporting me quietly but steadfastly throughout this journey. I couldn't have done it without you.

– Tao Lin

May 15, 2025

Table of Contents

List of Figures

Chapter 1

Introduction

From digital marketplaces, social networks, to smart cities, modern multi-agent systems increasingly shape our world. A key challenge behind the design of multi-agent systems is the misaligned incentives of agents. The interaction among self-interested agents can easily lead a system to an undesirable state. This raises the fundamental problem of *incentive design*: how can we design a multi-agent system so that self-interested agents are incentivized to reach globally desirable outcomes?

Incentive design for multi-agent systems lies in the intersection of economics, computer science, operations research, and other fields. Standard approaches to incentive design include *mechanism design* (designing rewards/payments for the agents) and *information design* (controlling the information available to an agent). The traditional theory of mechanism and information design, however, often relies on idealized assumptions. A key assumption is *knowledgeable principal*: the principal (designer of the system) knows about the agents and the environment well enough to design the system optimally. While the mechanism design literature offers tools to elicit agents' private information, it still presumes that the principal knows, e.g., the agents' utility structures and the distribution of agents' private information. Another key assumption is *rational agent*: the agents can make decisions that best respond to other agents and the principal. In contrast, real-world decision-makers are known to be boundedly rational in various ways. Such assumptions limit the practicality of the traditional mechanism and information design theory.

This dissertation addresses incentive design problems under more realistic assumptions — specifically, without assuming knowledgeable principals and rational agents. Instead, the focus is on *learning-based* decision-makers. In numerous real-world systems (such as online advertising auctions), principals (auctioneers) and agents (bidders) exhibit learning behavior or directly use machine learning algorithms to make data-driven decisions – a practice that is gaining unprecedented popularity in the present day. This dissertation thus explores the implications of learning for incentive design problems in multi-agent systems. Three directions will be studied: (1) incentive design by learning principals; (2) incentive design for learning agents; and (3) incentive issues in machine learning systems.

1.1 Incentive Design by Learning Principals

The first part of this dissertation is about incentive design by learning principals, covering information design (Chapters 2 and 3) and mechanism design (Chapter 4).

As a microeconomic subject starting from the 1970s [Spe73, CS82, KG11], information design studies how information providers (principals) can strategically disclose information to influence the behavior of uninformed decision-makers (agents). A key aspect of information design is to model an agent's change of belief upon receiving a piece of information from the principal. The standard modeling approach is *Bayesian*: suppose the agent has a prior belief μ_0 about an unknown random variable ω called "the state of the world"; after receiving a signal s correlated with ω from the principal, the agent forms the posterior belief about the state of the world using Bayes rule. Despite being theoretically appealing, this Bayesian approach falls short in capturing the complicated, possibly non-Bayesian, belief updating processes of real-world decision-makers. Although previous works in information design has investigated various non-Bayesian updating processes (e.g., [AC16a, HL17, DP22, dCZ22, FHT24], they all assume that such processes are known to the principal. Chapters 2 and 3 of this dissertation study information design problems where the principal does not know the agent's non-Bayesian belief updating process. Instead, the principal *learns* the agent's belief updating process via interaction with the agent. The contributions of these two chapters are summarized below:

- **Chapter 2: Information Design with Unknown Prior.** Suppose the agent in an information design problem has a subjective prior belief μ^* that is "incorrect", i.e., not equal to the true distribution of the state of the world ω, and is unknown to the principal. Can the principal learn the agent's belief μ^* through repeated interactions with the agent, by designing information revelation policies adaptively? If yes, how fast can the principal do that?

  We provide positive answers to both questions by designing a learning algorithm for the principal to achieve $O(\log T)$ regret, namely, the cumulative loss for the principal due to learning is only logarithmic in the total number T of rounds of interactions. The algorithm we design is inspired by the "learning from revealed preference" idea in economics and circumvents the inevitable $\Omega(\sqrt{T})$ error of using samples to estimate a distribution. The algorithm features a multi-dimensional binary search that is more efficient than classical optimization algorithms such as the ellipsoid method.

- **Chapter 3: Information Design with Unknown Bias.** This chapter considers another form of non-Bayesian belief update: the agent has a correct prior belief about the state of the world, but is biased towards the prior when performing Bayes update upon receiving a signal. While this type of biased belief update is supported by human-subject experiments [TH21], theoretical attempts to *quantifying* the bias are under-explored. We provide the first theoretical framework to measure the level of bias of an agent using information design methodologies. The algorithm we design provably achieves the optimal sample complexity for this bias quantification problem.

Moving from information design to mechanism design, Chapter 4 of this dissertation studies incentive design in an archetypical type of mechanisms: *auctions*, in particular, *non-truthful* auctions. In non-truthful auctions such as first-price and all-pay auctions, the independent strategic behaviors of bidders, with the corresponding equilibrium notion – Bayesian Nash equilibria (BNE) – are known to be problematic. For example, independent learning dynamics of bidders may cause undesirable oscillations to the system and lead to outcomes with low welfare or revenue (as shown by previous works [EO07, BS22, BLO$^+$25] and my work [DHLZ22] in Chapter 6). The BNEs of non-truthful auctions are notoriously difficult to characterize or compute when bidders' private valuations are not identically and independently distributed [CP23, FRGHK24].

An alternative approach to designing better auction systems, then, is to *coordinate the bidders*. Coordination can potentially stabilize the system and lead to better outcomes than the independent outcomes. In modern auction systems such as online advertising auctions where bidders delegate the bidding task to platforms that run auto-bidding algorithms, those platforms can, at least in principle, coordinate different bidders' bids [DGPS23, CWD$^+$23]. A desideratum here is *incentive compatibility*: bidders should be willing to report their private values to the coordinator truthfully and submit the bids that are recommended by the coordinator. In other words, the coordinator has to find a *Bayesian version of correlated equilibrium* for the bidders. A Bayes correlated equilibrium (BCE), however, is sensitive to the distribution of bidders' private values. If the full distribution is unavailable, can the coordinator find a BCE using samples of bidders' private values? How many samples are needed?

- **Chapter 4: Learning to Coordinate Bidders in Non-Truthful Auctions.** This chapter initiates the study of the sample complexity of Bayes correlated equilibrium in non-truthful auctions. As there are multiple definitions of BCE in the literature [For06], we focus on the strategic-form BCE. We show that a polynomial number of samples, $\tilde{O}(\frac{n}{\varepsilon^2})$, is sufficient to find all strategic-form BCE in a large class of non-truthful auctions, including first-price and all-pay auctions. This moderate amount of samples demonstrates the practicality of learning to coordinate bidders in non-truthful auctions. Our technique is a non-trivial analysis of the pseudo-dimension of the class of all monotone bidding strategies of bidders. Our result and technique can be a starting point for the study of the sample complexity of coordinating players in more general mechanism design problems, which I believe is a fruitful agenda for future research.

Together, these chapters highlight the potential of learning-based approaches to enhance information and mechanism design when the principal is knowledge-constrained.

1.2 Incentive Design for Learning Agents

The second part of this dissertation is about incentive design for learning agents. As mentioned earlier, learning is arguably a more realistic behavioral model for agents than full rationality. Learning as a behavioral model dates back to the early economic literature on learning in games (e.g., [Bro51, FL98]) and has been actively studied by computer scientists and operations researchers in recent years (e.g., [NST15, BMSW18, DSS19, MMSS22, D'A23, GKS+24]). Chapter 5 of this dissertation studies a general class of principal-agent problems with a single learning agent. Chapter 6 then focuses on a specific problem (first-price auction) with multiple learning agents.

- **Chapter 5: Generalized Principal-Agent Problems with a Learning Agent.** Previous work on *playing against a learning agent* [DSS19, GKS+24] has made an interesting observation: if an agent's learning behavior satisfies a condition called "no swap regret", then the dynamic game between the principal and the learning agent converges to a repeated game where the agent best responds to the principal at every period, as the number of periods T approaches infinity. However, this observation was previously known for only a handful of complete-information games, such as bimatrix Stackelberg games and contract design; the generality of this observation remained open. Moreover, the rate of convergence of the game with a learning agent to the game with a best-responding agent was unknown.

Our work fills the above two gaps. First, we show that any principal-agent game where the learning agent has no private information (while the principal can be privately informed) converges to the game with a best-responding agent, as $T \to \infty$. This contribution generalizes the previous observation for complete-information games to a large class of incomplete-information games, such as Bayesian persuasion. Second, we characterize the convergence rate, proving that the principal's average utility in the T periods lies in the range of $\big[U^* - \Theta(\sqrt{\frac{\mathrm{SReg}(T)}{T}}), U^* + \Theta(\frac{\mathrm{SReg}(T)}{T})\big]$, where U^* is the principal's optimal utility in the game with a best-responding agent (known as the Stackelberg value), and $\mathrm{SReg}(T)$ is the swap regret of the learning agent. Interestingly, while this range converges to U^* as $T \to \infty$, the upper range $U^* + \Theta(\frac{\mathrm{SReg}(T)}{T})$ and the lower range $U^* - \Theta(\sqrt{\frac{\mathrm{SReg}(T)}{T}})$ are not symmetric — a new observation found by our work.

En route to obtaining the above two results, we develop a unified analytical framework to reduce any principal-agent problem with a learning agent to the problem with an approximately best-responding agent. Unlike the rough asymptotic analysis in previous work, our reduction is precise and enables an exact characterization of the principal's obtainable utility against a learning agent. We believe that this reduction is of independent interest.

- **Chapter 6: Multi-Agent Learning in Auctions.** This chapter studies multi-agent learning dynamics in a specific game: first-price auction. Whether the learning dynamics of multiple bidders converge to equilibrium in repeated first-price auctions is a long-standing open problem. This problem was not fully understood even in the case where bidders have fixed private values. Our work provides a complete characterization of the convergence properties of multi-agent learning in first-price auctions with fixed values, for a large class of natural learning algorithms (including, e.g., Multiplicative Weight Update and other no-regret learning algorithms). We identify the conditions under which such learning dynamics converge and do not converge. In the case of convergence, our proof features a combination of the iterative-elimination-of-dominated-strategy idea in game theory and a novel concentration analysis for infinite-horizon stochastic processes. Our technique has been adopted by later work to prove the convergence of multi-agent learning dynamics in more general games [BDO24].

1.3 Incentive Issues in Machine Learning Systems

The third part of this dissertation, more applied in nature, focuses on the incentive issues that arise in real-world machine learning systems. As the strategic behaviors of humans or algorithms are ubiquitous, understanding the impact of such behaviors is essential to the design of socially responsible AI systems. This part of my dissertation is based on my research at ByteDance in 2023 on one of the most successful commercial applications of machine learning algorithms: recommender systems.

- **Chapter 7: Incentives and Polarization in Recommender Systems.** Modern recommender systems use machine learning algorithms to predict users' preferences about items and recommend relevant items to users. Despite the enormous commercial success, recommender systems are known to cause adverse effects such as filter bubbles and polarization. The main solution proposed by previous works and used in practice to prevent such effects is to diversify the recommendation: recommending random items to broaden users' viewpoints. However, an important aspect is neglected by previous works: creators of items have incentives to make their items more attractive. Our work shows that, due to such strategic behavior of creators, diversification techniques cannot prevent (sometimes even worsen) the filter bubble and polarization effects in recommendation systems, both theoretically and empirically.

The reason is that different creators tend to create more similar items when some recommendations become random. We also show that, surprisingly, algorithms such as top-k truncation, which target algorithmic efficiency rather than recommendation diversity, can actually mitigate the polarization effect in recommendation systems with strategic creators. This work underscores the importance of incorporating strategic considerations into the design of socially responsible machine learning systems.

Part I

Incentive Design by Learning Principals

Chapter 2

Information Design with Unknown Prior

based on joint work with Ce Li [LL25]

This chapter focuses on information design by a learning principal.

2.1 Introduction

As a microeconomic subject with a long history (e.g., [Spe73, CS82, KG11]), *information design* studies how information providers can strategically disclose information to influence the behavior of decision makers. Information design has received tremendous attention over the years in many other fields, including computer science [DX16, BTCXZ24], and been applied to various domains: e.g., voting [AC16b, CCG20], online advertising [EFG+14, BBX18, BCM+22, AFT23], security games [RJJX15, XFC+16], and recommendation systems [MSSW16, IMSW20, ZIX21, FTX22, HMCG24].

Classic models of information design, such as *Bayesian persuasion* [KG11] and *cheap talk* [CS82], include a sender (principal), a receiver (agent), and a hidden state of the world $\omega \in \Omega$. The two players share a *common prior*: they both believe that the state ω follows a distribution $\mu_0 \in \Delta(\Omega)$. The sender observes the realization of the state ω and sends a randomized signal $s \in S$ to the receiver. Based on the signal s, the receiver performs Bayes update to obtain the posterior belief about the state and then makes a decision that determines the payoffs to both players. The goal of the sender is to design a signaling scheme to maximize their own payoff.

Common prior is arguably a strong assumption. In many cases, the receiver may have a different prior belief about the state of the world than the sender's. Previous works on information design with heterogeneous priors either take a distributional approach, assuming that the sender has a correct belief about the receiver's belief [KMZL17, GS19, Kos22], or take a worst-case approach, assuming that the receiver's belief is completely unknown and adversarial [CHJ20, DP22].

In this work, we study information design without the common prior assumption from a perspective that lies between the distributional and the worst-case approaches: the sender does not have distributional knowledge about the receiver's belief, but can learn the receiver's belief over time from repeated interactions with the receiver. This perspective is conceptually related to the "learning from revealed preference" literature [BV06, ZR12], which studies how to infer an agent's preference from the decisions made by the agent.

2.1.1 Overview of Our Contributions

We study a repeated Bayesian persuasion problem where the receiver has a subjective prior belief μ^* about the state of the world that is unknown to the sender. This captures settings where, for example, the receiver receives an external signal from the environment that is not observed by the sender. The sender and the receiver interact for T periods, where at each period, the sender designs a signaling scheme $\pi^{(t)}$ to map a state $\omega^{(t)}$ (i.i.d. sampled from the true distribution μ_0) to a signal $s^{(t)}$. Based on the subjective prior μ^* and the signal $s^{(t)}$, the receiver performs a Bayesian update to obtain the posterior belief about the state and then takes an optimal action. The action affects the payoffs to the two players and is observed by the sender. We aim to design an algorithm for the sender to learn to design good signaling schemes. The performance of a learning algorithm is measured by the *regret*: the sender's accumulated payoff under the optimal signaling scheme (with knowledge of μ^*) minus the actual accumulated payoff obtained by the algorithm. Due to the subjective nature of the receiver's prior, classical statistical methods to estimate a distribution using samples do not work. Even if we could sample from the receiver's prior belief μ^*, estimating μ^* by samples would cause an $\Omega(\sqrt{T})$ loss to the sender's payoff due to sampling error [ZIX21].

In this work, we design a learning algorithm for the sender to achieve $O(\log T)$ regret. This result circumvents the fundamental limitation of empirical estimation. Our algorithm is based on a *multi-dimensional binary search* to learn the unknown prior μ^* and a *robustification procedure* to obtain near-optimal signaling schemes. We illustrate the high-level idea below. Our key idea is to use the actions taken by the receiver to infer the receiver's prior μ^*. Consider the case with only two states ω_1, ω_2. Let π be a signaling scheme. Upon receiving a signal s from π, the receiver's posterior belief about the state is, by Bayes rule,

$$\mu^*(\omega_1|s) = \tfrac{\mu^*(\omega_1)\pi(s|\omega_1)}{\Pr(s)}, \ \mu^*(\omega_2|s) = \tfrac{\mu^*(\omega_2)\pi(s|\omega_2)}{\Pr(s)}, \text{ where } \Pr(s) = \mu^*(\omega_1)\pi(s|\omega_1) + \mu^*(\omega_2)\pi(s|\omega_2).$$

Let $v(a, \omega)$ be the receiver's utility when taking action a under state ω. If we observe that the receiver takes action a instead of another action a', then we know that the posterior expected utilities of the receiver by taking the two actions satisfy

$$\begin{aligned} & \mu^*(\omega_1|s)v(a,\omega_1) + \mu^*(\omega_2|s)v(a,\omega_2) \ \geq \ \mu^*(\omega_1|s)v(a',\omega_1) + \mu^*(\omega_2|s)v(a',\omega_2) \\ \Longrightarrow \ & \mu^*(\omega_1)\pi(s|\omega_1)\Big(v(a,\omega_1) - v(a',\omega_1)\Big) + \mu^*(\omega_2)\pi(s|\omega_2)\Big(v(a,\omega_2) - v(a',\omega_2)\Big) \ \geq \ 0 \\ \Longrightarrow \ & \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} \ \geq \ \frac{\pi(s|\omega_2)\big(v(a',\omega_2) - v(a,\omega_2)\big)}{\pi(s|\omega_1)\big(v(a,\omega_1) - v(a',\omega_1)\big)}. \end{aligned}$$

The above is an inequality regarding the unknown quantity $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$. By employing multiple different signaling schemes, we can obtain multiple inequalities for $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$, which allows us to use a binary search to estimate the prior μ^* with accuracy $\varepsilon = \frac{1}{T}$ in $O(\log T)$ periods. Our formal analysis will show how to generalize this intuition to the case with multiple states and multiple actions. Then, we compute an optimal signaling scheme from the estimated prior $\hat{\mu}^*$ and *robustify* the signaling scheme to ensure that it works well for the true prior μ^*. We thus obtain an algorithm that achieves $O(\log T)$ regret for the sender:

Main Result (Theorem 2.1): *We design a learning algorithm for the sender to achieve $O(\log T)$ regret for the repeated Bayesian persuasion problem where the receiver has an unknown subjective prior.*

2.1.2 Additional Related Works

Our work is related to the literature of *online Bayesian persuasion*, which designs algorithms for the sender to learn various unknown parameters in repeated Bayesian persuasion games, such as unknown utility functions [CCMG20, CMCG21, FTX22], unknown state distribution [CHJ20, ZIX21, WZF+22, HILS25], or both [BBC+24]. The main previous approach to learning the unknown distribution is empirical estimation, which suffers $\Omega(\sqrt{T})$ regret due to sampling error. We bypass this negative result by using the receiver's best-responding action to infer the prior, which is much more efficient and can achieve $O(\log T)$ regret. The idea of inferring the prior from the receiver's actions also appears in [HILS25], who aim to find the optimal learning algorithm that uses a constant number of samples of the receiver's actions. [HILS25]'s goal is challenging and their result only applies to a specific Bayesian persuasion setting with single-dimensional state and binary actions. In contrast, we have a less challenging goal: we do not aim to find the optimal learning algorithm and do not restrict to a constant number of samples. This allows us to deal with general instances with multiple states and more than two actions.

2.2 Background: Bayesian Persuasion

This section provides background on Bayesian persuasion. Bayesian persuasion is a game between a *sender* (she) and a *receiver* (he). There is a finite set of states of the world

$\Omega = \{\omega_1, \ldots, \omega_{|\Omega|}\}$. Unlike classical work that assumes a common prior [KG11], we assume that the sender and the receiver have different prior beliefs about the state of the world. The sender believes that ω follows a distribution $\mu_0 \in \Delta(\Omega)$, while the receiver believes that ω follows $\mu^* \in \Delta(\Omega)$.[1] The receiver has a finite set of actions $A = \{a_1, \ldots, a_{|A|}\}$. At the start of the game, the sender designs and announces a signaling scheme $\pi : \Omega \to \Delta(S)$, which is a mapping from every state to a probability distribution over signals in some finite signal set S. We use $\pi(s|\omega)$ to denote the probability that signal s is realized conditioning on state ω. Then, a state ω is sampled according to the sender's prior μ_0.[2] The receiver does not observe ω. Instead, he observes a signal $s \sim \pi(\cdot|\omega)$ drawn conditionally on ω according to π. Based on signal s, prior μ^*, and signaling scheme π, the receiver forms posterior belief μ_s^* about the state by Bayes' rule:

$$\mu_s^*(\omega) \;=\; \frac{\mu^*(\omega)\pi(s|\omega)}{\sum_{\omega_i \in \Omega} \mu^*(\omega_i)\pi(s|\omega_i)}, \quad \forall \omega \in \Omega, \tag{2.1}$$

and then takes an optimal action with respect to μ_s^*:

$$a_s^* \;\in\; \operatorname*{arg\,max}_{a \in A} \mathbb{E}_{\omega \sim \mu_s^*}[v(a, \omega)] \;=\; \operatorname*{arg\,max}_{a \in A} \sum_{\omega \in \Omega} \mu^*(\omega)\pi(s|\omega)v(a, \omega), \tag{2.2}$$

where $v : A \times \Omega \to [0, 1]$ is a bounded utility function of the receiver. The sender then obtains utility $u(a_s^*, \omega)$ where $u : A \times \Omega \to [0, 1]$ is the sender's utility function. The expected utility of the sender, as a function of priors μ_0, μ^*, and signaling scheme π, is

$$U(\mu_0, \mu^*, \pi) \;=\; \mathbb{E}\big[u(a_s^*, \omega)\big] \;=\; \sum_{\omega \in \Omega} \mu_0(\omega) \sum_{s \in S} \pi(s|\omega) u(a_s^*, \omega). \tag{2.3}$$

[1] The notation $\Delta(X)$ denotes the set of probability distributions over X.

[2] One can also assume that the state is sampled according to the receiver's prior μ^*. Our results will stay the same, except for a change of constants in the $O(\cdot)$ notations. We think sampling from the sender's prior is a more natural modeling choice.

The sender aims to find a signaling scheme π^* to maximize her expected utility:

$$\pi^* \in \operatorname*{arg\,max}_{\pi:\Omega\to\Delta(S)} U(\mu_0, \mu^*, \pi). \tag{2.4}$$

Let $U^* = U(\mu_0, \mu^*, \pi^*)$ denote the utility of an optimal signaling scheme. A signaling scheme π is *ε-approximately optimal* (or *ε-optimal*) if

$$U(\mu_0, \mu^*, \pi) \geq U^* - \varepsilon = \max_{\pi} U(\mu_0, \mu^*, \pi) - \varepsilon. \tag{2.5}$$

The *regret* of a signaling scheme π is the difference $U^* - U(\mu_0, \mu^*, \pi)$.

Direct and persuasive signaling schemes. For the above Bayesian persuasion problem, it is well known that there exists an optimal signaling scheme π^* that is *direct* and *persuasive*. This fact is known as the "revelation principle" in information design [KG11]. A *direct* signaling scheme $\pi : \Omega \to \Delta(A)$ maps every state to a probability distribution over actions, so every signal $a \in A$ can be regarded as an action recommendation for the receiver. A direct signaling scheme π is *persuasive for action/signal $a \in A$* if the recommended action a is optimal for the receiver: $\sum_\omega \mu^*(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0, \forall a' \in A$; and π is *persuasive* if it is persuasive for all actions $a \in A$. Note that persuasiveness is defined with respect to the receiver's prior belief μ^*. Let $\mathrm{Pers}(\mu^*)$ be the set of all persuasive signaling schemes under prior μ^*:

$$\mathrm{Pers}(\mu^*) := \Big\{\pi : \Omega \to \Delta(A) \mid \sum_{\omega\in\Omega} \mu^*(\omega)\pi(a|\omega)\big[v(a,\omega) - v(a',\omega)\big] \geq 0,\ \forall a, a' \in A\Big\}. \tag{2.6}$$

With attention restricted to persuasive signaling schemes, an optimal signaling scheme π^* can be computed efficiently by the following linear program:

$$\max_{\pi\in\mathrm{Pers}(\mu^*)} \sum_{\omega\in\Omega} \mu_0(\omega) \sum_{a\in A} \pi(a|\omega) u(a,\omega). \tag{2.7}$$

2.3 Main Result: Learning the Unknown Prior

This section designs an algorithm for the sender to learn to design approximately optimal signaling schemes in repeated Bayesian persuasion problems where the receiver's subjective prior belief is unknown. The formal model is as follows.

Model: The sender knows the utility functions of both players, has her own prior μ_0 about the state of the world, but does not know the receiver's prior μ^*. The two players interact for T periods. In each period $t \in \{1, \ldots, T\}$, the following happens in order:

- Based on history, the sender designs a signaling scheme $\pi^{(t)} : \Omega \to \Delta(S)$.
- A new state $\omega^{(t)} \sim \mu_0$ is independently sampled. A signal $s^{(t)} \sim \pi^{(t)}(\cdot|\omega^{(t)})$ is realized and sent to the receiver.
- The receiver performs a Bayesian update from μ^*, $\pi^{(t)}$, and $s^{(t)}$, and takes an optimal action $a^{(t)}$ based on the posterior belief $\mu^{(t)}_{s^{(t)}}$, as in Equation (2.2).
- The sender obtains utility $u(a^{(t)}, \omega^{(t)})$. The receiver obtains utility $v(a^{(t)}, \omega^{(t)})$.
- The sender observes the realized signal $s^{(t)}$ and the action $a^{(t)}$ taken by the receiver.

We do not require the sender to observe the state $\omega^{(t)}$; the sender does not observe the state when, e.g., the signal is sent by a third party other than the sender. The regret of the sender is the difference between the optimal utility U^* and the actual (expected) utility obtained during the T periods:

$$\mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T} \mathbb{E}\Big[U^* - u(a^{(t)}, \omega^{(t)})\Big] = T \cdot U^* - \mathbb{E}_{\pi^{(1)},\ldots,\pi^{(T)}}\Big[\sum_{t=1}^{T} U(\mu_0, \mu^*, \pi^{(t)})\Big]. \quad (2.8)$$

Clearly, if the sender can use an ε-approximately optimal signaling scheme π in all T periods, then her regret will be at most εT. Thus, the goal of the sender is to find approximately optimal signaling schemes while learning the receiver's unknown prior μ^* during the T periods.

Overview of our result: We will design a learning algorithm for the sender to achieve $O(\log T)$ regret. At a high level, the algorithm has an exploration-exploitation structure: first estimate the receiver's subjective prior belief μ^* with accuracy $\varepsilon = \frac{1}{T}$, then construct an $O(\varepsilon)$-approximately optimal signaling scheme from the estimated prior. The first phase takes $O(\log \frac{1}{\varepsilon}) = O(\log T)$ periods, suffering $O(\log T)$ regret. The second phase has $O(\varepsilon)$ regret per round. So, the total regret is at most $O(\log T) + O(\varepsilon)T = O(\log T)$.

The second phase of the algorithm uses the idea of *robustification of signaling schemes*, which we present in Section 2.3.1. Then in Section 2.3.2 we discuss how to estimate the receiver's prior efficiently. Finally, we present the full algorithm in Section 2.3.3.

2.3.1 Robustification of Signaling Schemes

A technique that we will use to design signaling schemes in the absence of the prior is *robustification.* Suppose that we are given an estimation $\hat{\mu}$ of the unknown prior μ^* with the ℓ_1-distance satisfying $\|\hat{\mu} - \mu^*\|_1 \le \varepsilon$. We can compute a signaling scheme $\hat{\pi}$ that is optimal for the estimated prior $\hat{\mu}$. As mentioned in Section 2.2, $\hat{\pi}$ can be assumed to be persuasive under $\hat{\mu}$, namely $\hat{\pi} \in \mathrm{Pers}(\hat{\mu})$. However, $\hat{\pi}$ might be not persuasive under the true prior μ^*. Moreover, $\hat{\pi}$ may even perform very poorly on μ^* because, although a signal s induces two similar posterior distributions $\hat{\mu}_s$ and μ^*_s under priors $\hat{\mu}$ and μ^*, the argmax actions of the receiver under the two posteriors might be very different. The idea of robustification is to slightly modify the signaling scheme $\hat{\pi}$ to be another scheme $\tilde{\pi}$ that is persuasive and approximately optimal for all the priors close to $\hat{\mu}$, including μ^*. Robustification requires some mild assumptions on the priors and the receiver's utility function:

Assumption 2.1 (Regularity on priors)**.** *There exists $p_0 > 0$ such that both players' priors satisfy: $\forall \omega \in \Omega$, $\mu^*(\omega) \ge p_0$ and $\mu_0(\omega) \ge p_0$.*

Assumption 2.2 (Regularity on receiver's utility)**.** *There exists $D > 0$ such that for any action $a \in A$ of the receiver, there exists a belief $\eta_a \in \Delta(\Omega)$ on which action a is better than any other action by a margin D: $\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega)] \ge \mathbb{E}_{\omega\sim\eta_a}[v(a',\omega)] + D$, $\forall a' \in A \setminus \{a\}$.*

Assumptions 2.1 and 2.2 are standard in the literature (e.g., [ZIX21]). And we assume that the sender knows p_0 and D. Why is Assumption 2.2 mild? If it does not hold, then there must exist an action $a_0 \in A$ such that, for any belief $\eta \in \Delta(\Omega)$, we have $\mathbb{E}_{\omega\in\eta}[v(a_0,\omega)] \le \mathbb{E}_{\omega\in\eta}[v(a',\omega)]$ for some $a' \in A \setminus \{a_0\}$. That means that the receiver can always take an action that is not a_0 without decreasing his utility, regardless of his belief. Thus, a_0 is a dominated action that can be deleted from the receiver's action set. After all dominated actions are deleted, Assumption 2.2 will be satisfied.

The following lemma formalizes the idea of robustification:

Lemma 2.1 (Robustification)**.** *Suppose $\varepsilon \le \frac{p_0^2 D}{4}$. Let $\hat{\pi}$ be a signaling scheme that is persuasive for the receiver under prior $\hat{\mu}$. We can convert $\hat{\pi}$ into another signaling scheme $\tilde{\pi}$ that satisfies the following:*

- *$\tilde{\pi}$ is persuasive for all receiver priors in $B_1(\hat{\mu}, \varepsilon) = \{\mu \in \Delta(\Omega) : \|\mu - \hat{\mu}\|_1 \le \varepsilon\}$.*
- *The sender's utilities under the two signaling schemes satisfy $U(\mu_0, \hat{\mu}, \tilde{\pi}) \ge U(\mu_0, \hat{\mu}, \hat{\pi}) - \frac{6\varepsilon}{p_0^2 D}$.*

The construction of the $\tilde{\pi}$ above uses a standard technique in the literature [ZIX21]: taking a mixture of signaling scheme $\hat{\pi}$, the signaling scheme that induces beliefs η_a for $a \in A$, and the signaling scheme that fully reveals the state, with mixture weights

$1-2\delta, \delta$, and δ. Since $\|\hat{\mu}-\mu^*\|_1 \le \varepsilon$, $\hat{\pi}$ is not persuasive for the receiver under prior μ^* by a margin of at most $O(\frac{\varepsilon}{p_0})$. The signaling scheme that induces η_a is strictly persuasive for the receiver by a margin of D according to Assumption 2.2. The fully-revealing signaling scheme is weakly persuasive. So, the persuasiveness of the mixture signaling scheme is $D \cdot \delta + 0 \cdot \delta - (1-2\delta) \cdot O(\frac{\varepsilon}{p_0}) \ge 0$ by choosing $\delta = O(\frac{\varepsilon}{p_0 D})$. Then, the loss of utility for the sender due to the mixture is at most $2\delta \cdot O(\frac{1}{p_0}) = O(\frac{\varepsilon}{p_0^2 D})$. See details in Section 2.5.2.

By applying Lemma 2.1 to the optimal signaling scheme for the estimated prior $\hat{\mu}$ that satisfies $\|\hat{\mu}-\mu^*\|_1 \le \varepsilon$, we immediately obtain the following corollary:

Corollary 2.1 (Robustification for optimal signaling scheme)**.** *Suppose* $\|\hat{\mu}-\mu^*\|_1 \le \varepsilon \le \frac{p_0^2 D}{4}$*. Let* $\hat{\pi}$ *be an optimal signaling scheme for receiver prior* $\hat{\mu}$*. We can convert* $\hat{\pi}$ *into another signaling scheme* $\tilde{\pi}$ *that is* $\frac{6\varepsilon}{p_0^2 D}$*-optimal for receiver prior* μ^**.*

2.3.2 Estimating Prior in $O(\log \frac{1}{\varepsilon})$ Periods

We then consider how to estimate the receiver's prior μ^*. Due to the subjective nature of μ^*, we cannot estimate it using samples. Instead, we design an algorithm that uses the actions taken by the receiver to infer μ^*. Recall that the receiver takes an action that is optimal on his posterior belief updated from the prior μ^* after receiving a signal. Such an action contains information about μ^*. By employing multiple different signaling schemes, the sender can gradually acquire more information about μ^*. Such a process requires a natural and mild assumption on the receiver's utility function $v(\cdot,\cdot)$:

Assumption 2.3 (Unique optimal action)**.** *For each state* $\omega \in \Omega$*, the optimal action* $a_\omega = \arg\max_{a \in A} v(a,\omega)$ *for the receiver is unique and strictly better than any other action by a positive margin of* G*:* $v(a_\omega,\omega) - v(a',\omega) > G > 0, \forall a' \in A \setminus \{a_\omega\}$*.*

We estimate the prior $\mu^* \in \Delta(\Omega)$ by estimating the ratio of probability $\frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}$ between every state ω_i ($i = 2, \ldots, |\Omega|$) and a fixed state ω_1. How do we estimate the probability ratio between two states? We first show how to do this for a pair of states, say ω_1, ω_2, whose receiver-optimal actions $a^*_{\omega_1}, a^*_{\omega_2}$ are different. We call such a pair of states *distinguishable*, and assume that such a pair exists:

Assumption 2.4 (Distinguishable states). *There exist two states $\omega_1, \omega_2 \in \Omega$ whose corresponding receiver-optimal actions are different: $a^*_{\omega_1} \neq a^*_{\omega_2}$.*

Assumption 2.4 makes the sender's learning problem non-trivial. If Assumption 2.4 does not hold, then all states in Ω will share the same receiver-optimal action a^*. Thus, regardless of the signaling scheme and the prior, the receiver will always find the action a^* to be optimal at his posterior, therefore take action a^*. The sender's expected utility will be the constant $\sum_{\omega \in \Omega} \mu^*(\omega) u(a^*, \omega)$, so she achieves 0 regret.

Algorithm 2.1 shows how to estimate the prior probability ratio $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$ between a pair of distinguishable states ω_1 and ω_2 using different signaling schemes in multiple periods. The main idea is a binary search. Specifically, if the sender uses a signaling scheme π that sends some signal s_0 only under states ω_1 and ω_2 (namely, $\pi(s_0|\omega) = 0$ for $\omega \neq \omega_1, \omega_2$), then the receiver will believe that the state must be ω_1 or ω_2 whenever he receives signal s_0. If the signaling ratio $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)}$ is zero, then the receiver will know for sure that the state is ω_1, thereby taking the optimal action a_1 for state ω_1. On the other hand, if the signaling ratio $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)}$ is large (say, $\geq \frac{1}{Gp_0}$), then the receiver will believe that the state is ω_2 with a high probability and take the optimal action for ω_2, which is different from a_1 by Assumption 2.4. Such reasoning suggests that there must exist some threshold $\tau \in [0, \frac{1}{Gp_0}]$ such that the receiver will start taking some action $\tilde{a} \neq a_1$ when the signaling ratio $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)}$ is above τ ($\tilde{a}$ is not necessarily equal to a_2). The sender can find the threshold τ by a binary search. The receiver must be indifferent between taking action a_1 and the different action $\tilde{a}$ at the signaling threshold: in other words, when $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = \tau$ we have

$$\mu^*(\omega_1)\pi(s_0|\omega_1)\big[v(a_1, \omega_1) - v(\tilde{a}, \omega_1)\big] + \mu^*(\omega_2)\pi(s_0|\omega_2)\big[v(a_1, \omega_2) - v(\tilde{a}, \omega_2)\big] = 0.$$

That implies

$$\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} = \frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} \cdot \frac{v(\tilde{a},\omega_2) - v(a_1,\omega_2)}{v(a_1,\omega_1) - v(\tilde{a},\omega_1)} = \tau \cdot \frac{v(\tilde{a},\omega_2) - v(a_1,\omega_2)}{v(a_1,\omega_1) - v(\tilde{a},\omega_1)},$$

which gives the value of the prior ratio $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$.

Algorithm 2.1: Binary Search for Prior Ratio Between Distinguishable States

```
Input     : two states $\omega_1, \omega_2$ whose receiver-optimal actions are different
Parameter: a desired accuracy $\varepsilon > 0$
Output    : an estimation $\hat{\rho}$ of the ratio $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$ that satisfies $|\hat{\rho} - \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}| \le \varepsilon$
Let $k = 0$, $\ell^{(0)} = 0$, $r^{(0)} = \frac{1}{Gp_0}$.
Let $a_1 = \arg\max_{a \in A} v(a, \omega_1)$, $\tilde{a} = \arg\max_{a \in A} v(a, \omega_2)$. ($a_1 \neq \tilde{a}$ by assumption)
while $r^{(k)} - \ell^{(k)} > \varepsilon G$ do
    Let $q = \frac{\ell^{(k)} + r^{(k)}}{2}$.
    Let $s_0$ be an arbitrary signal in $S$; let $\pi^{(k)}$ be a signaling scheme that satisfies
      $\frac{\pi^{(k)}(s_0|\omega_2)}{\pi^{(k)}(s_0|\omega_1)} = q$ and $\pi^{(k)}(s_0|\omega) = 0$ for $\omega \neq \omega_1, \omega_2$ (see the proof of Lemma 2.2
      for how to construct such a $\pi^{(k)}$).
    Use $\pi^{(k)}$ for multiple periods until signal $s_0$ is sent. Let $a^{(k)}$ be the action
      taken by the receiver when $s_0$ is sent.
    if $a^{(k)} = a_1$ then
        Let $\ell^{(k+1)} = q$, $r^{(k+1)} = r^{(k)}$.
    else
        Let $r^{(k+1)} = q$, $\tilde{a} = a^{(k)}$, $\ell^{(k+1)} = \ell^{(k)}$.
    $k = k + 1$.
Output $\hat{\rho} = \ell^{(k)} \cdot \frac{v(\tilde{a},\omega_2) - v(a_1,\omega_2)}{v(a_1,\omega_1) - v(\tilde{a},\omega_1)}$.
```

The performance of Algorithm 2.1 is given by Lemma 2.2, which shows that a good estimate of $\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)}$ can be obtained within a short amount of time with high probability.

Lemma 2.2. *The output $\hat{\rho}$ of Algorithm 2.1 satisfies $\hat{\rho} \le \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} \le \hat{\rho} + \varepsilon$, and Algorithm 2.1 terminates in at most $\frac{1}{p_0} \log_2 \frac{1}{G^2 p_0 \varepsilon}$ periods in expectation.*

Proof. See Section 2.5.3. □

What if the receiver-optimal actions of a pair of states are the same? Algorithm 2.2 deals with such a case, allowing the sender to estimate the prior ratio between any pair of states. The idea of Algorithm 2.2 is that, if states ω_i and ω_j have the same receiver-optimal actions, then both of them must be distinguishable from one state in a distinguishable pair of states, say $\omega_k \in \{\omega_1, \omega_2\}$. Thus, we estimate the ratios $\frac{\mu^*(\omega_i)}{\mu^*(\omega_k)}$ and $\frac{\mu^*(\omega_j)}{\mu^*(\omega_k)}$ separately, which will give $\frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}$ by a division.

Algorithm 2.2: Estimating Prior Ratio Between Any Pair of States

```
Input    : any two states $\omega_i, \omega_j \in \Omega$
Parameter: accuracy $\varepsilon > 0$
Output   : an estimation $\hat{\rho}_{ij}$ of the ratio $\frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}$
If $\omega_i$ and $\omega_j$ are distinguishable, i.e., $a^*_{\omega_i} \neq a^*_{\omega_j}$, then run Algorithm 2.1 on $\omega_i$ and
  $\omega_j$ with parameter $\varepsilon$.
Otherwise, i.e., $a^*_{\omega_i} = a^*_{\omega_j}$, find $\omega_k \in \{\omega_1, \omega_2\}$ such that $a^*_{\omega_k} \neq a^*_{\omega_i} = a^*_{\omega_j}$. Run
  Algorithm 2.1 with parameter $\varepsilon$ to obtain an estimate $\hat{\rho}_{ik}$ for $\frac{\mu^*(\omega_i)}{\mu^*(\omega_k)}$ and an
  estimate $\hat{\rho}_{jk}$ for $\frac{\mu^*(\omega_j)}{\mu^*(\omega_k)}$. Return $\hat{\rho}_{ij} = \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}}$.
```

Lemma 2.3. *Suppose $\varepsilon \le \frac{p_0}{2}$. For any two states $\omega_i, \omega_j \in \Omega$, the output $\hat{\rho}_{ij}$ of Algorithm 2.2 satisfies $|\hat{\rho}_{ij} - \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}| \le \frac{2\varepsilon}{p_0^2}$. Algorithm 2.2 terminates in at most $\frac{2}{p_0} \log_2 \frac{1}{G^2 p_0 \varepsilon}$ periods in expectation.*

Proof. The proof uses Lemma 2.2 twice. See details in Section 2.5.4. □

By estimating the prior ratio between any pair of states using Algorithm 2.2, we can estimate the entire unknown prior μ^*. This is described in Algorithm 2.3.

```
Algorithm 2.3: Estimating Receiver's Unknown Prior
  Parameter: $\varepsilon > 0$.
  Output    : An estimation $\hat{\mu}$ of the receiver's prior $\mu^*$
  For every state $\omega_i \in \Omega$, $i = 2, \dots, |\Omega|$, use Algorithm 2.2 on states $\omega_i$ and $\omega_1$ with
    parameter $\varepsilon$ to obtain an estimation $\hat{\rho}_i$ of the ratio $\frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}$.
  Compute $\hat{\mu}$ by letting $\begin{cases} \hat{\mu}(\omega_1) = 1/\big(1 + \sum_{i=2}^{|\Omega|} \hat{\rho}_i\big); \\ \hat{\mu}(\omega_i) = \hat{\rho}_i \hat{\mu}(\omega_1), \text{ for } i = 2, \dots, |\Omega|. \end{cases}$
```

Lemma 2.4. *Suppose* $\varepsilon \le \frac{p_0}{2}$. *The estimated prior* $\hat{\mu}$ *satisfies* $\|\hat{\mu} - \mu^*\|_1 \le \frac{6|\Omega|\varepsilon}{p_0^3}$. *Algorithm 2.3 terminates in at most* $\frac{2|\Omega|}{p_0} \log_2 \frac{1}{G^2 p_0 \varepsilon}$ *periods in expectation.*

Proof. The estimation $\hat{\rho}_i$ satisfies $|\hat{\rho}_i - \frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}| \le \varepsilon' = \frac{2\varepsilon}{p_0^2}$ by Lemma 2.3. Since μ^* is a probability distribution, we have $1 = \sum_{\omega \in \Omega} \mu^*(\omega) = \mu^*(\omega_1) + \mu^*(\omega_1) \sum_{i=2}^{|\Omega|} \frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}$, so

$$\mu^*(\omega_1) = \frac{1}{1 + \sum_{i=2}^{|\Omega|} \frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}}.$$

Because the function $f(x) = \frac{1}{1+x}$ is 1-Lipschitz ($|f'(x)| = \frac{1}{(1+x)^2} \le 1$), we have

$$|\hat{\mu}(\omega_1) - \mu^*(\omega_1)| = \Big| \frac{1}{1 + \sum_{i=2}^{|\Omega|} \hat{\rho}_i} - \frac{1}{1 + \sum_{i=2}^{|\Omega|} \frac{\mu^*(\omega_i)}{\mu^*(\omega_1)}} \Big| \le \Big| \sum_{i=2}^{|\Omega|} \hat{\rho}_i - \sum_{i=2}^{|\Omega|} \frac{\mu^*(\omega_i)}{\mu^*(\omega_1)} \Big| \le |\Omega| \varepsilon'.$$

Then, consider any $i = 2, \dots, |\Omega|$.

$$\begin{aligned} |\hat{\mu}(\omega_i) - \mu^*(\omega_i)| &= \big| \hat{\rho}_i \hat{\mu}(\omega_1) - \tfrac{\mu^*(\omega_i)}{\mu^*(\omega_1)} \mu^*(\omega_1) \big| \\ &\le \big| \hat{\rho}_i - \tfrac{\mu^*(\omega_i)}{\mu^*(\omega_1)} \big| \hat{\mu}(\omega_1) \ + \ \tfrac{\mu^*(\omega_i)}{\mu^*(\omega_1)} \big| \hat{\mu}(\omega_1) - \mu^*(\omega_1) \big| \\ &\le \varepsilon' \hat{\mu}(\omega_1) \ + \ \tfrac{\mu^*(\omega_i)}{\mu^*(\omega_1)} |\Omega| \varepsilon'. \end{aligned}$$

Therefore,

$$\begin{aligned}\|\hat{\mu}-\mu^*\|_1 &= |\hat{\mu}(\omega_1)-\mu^*(\omega_1)| + \sum_{i=2}^{|\Omega|} |\hat{\mu}(\omega_i)-\mu^*(\omega_i)| \\ &\le |\Omega|\varepsilon' + |\Omega|\varepsilon' + |\Omega|\varepsilon' \sum_{i=2}^{|\Omega|} \tfrac{\mu^*(\omega_i)}{\mu^*(\omega_1)} \le 2|\Omega|\varepsilon' + |\Omega|\varepsilon' \tfrac{1}{p_0} \le \tfrac{3|\Omega|\varepsilon'}{p_0} = \tfrac{6|\Omega|\varepsilon}{p_0^3}.\end{aligned}$$

□

2.3.3 Full Algorithm: Exploration-Exploitation

Finally, we present the full learning algorithm for the sender, Algorithm 2.4. First, we use Algorithm 2.3 to estimate the receiver's prior μ^* with a high precision, i.e., obtaining an $\hat{\mu}$ satisfying $\|\hat{\mu}-\mu^*\|_1 \le O(\varepsilon)$. Then, we use the robustification technique (Corollary 2.1) to obtain a signaling scheme $\tilde{\pi}$ that is persuasive and $O(\varepsilon)$-approximately optimal for μ^*. Using $\tilde{\pi}$ for the remaining periods thus incurs a small regret. The total regret of Algorithm 2.4 is formally characterized in Theorem 2.1.

Algorithm 2.4: Learning to Persuade a Receiver with Unknown Prior

```
Input     : Utility functions $u, v$. Sender's prior $\mu_0$. Total number of periods $T$.
Parameter: $\varepsilon > 0$.
Use Algorithm 2.3 with parameter $\varepsilon$ to obtain an estimation $\hat{\mu}$ of the receiver's
  prior $\mu^*$. Let $T_0$ be the number of periods in this process.
Then, use Corollary 2.1 on $\hat{\mu}$ with parameter $\frac{6|\Omega|\varepsilon}{p_0^3}$ to construct a signaling scheme
  $\tilde{\pi}$. Use $\tilde{\pi}$ for the remaining $T - T_0$ periods.
```

Theorem 2.1. *Assume Assumptions 2.1, 2.2, 2.3, 2.4. Choose* $\varepsilon = \frac{p_0^5 D}{24|\Omega|T}$*. The regret of Algorithm 2.4 is at most*

$$\frac{2|\Omega|}{p_0} \log_2 \frac{24|\Omega|T}{G^2 p_0^6 D} \; + \; 2 \; = \; O\big(\log T\big).$$

Proof. According to Lemma 2.4, the expected number of periods taken by Algorithm 2.3 is at most $\mathbb{E}[T_0] \le \frac{2|\Omega|}{p_0}\log_2\frac{1}{G^2 p_0 \varepsilon}$ and the estimated prior $\hat{\mu}$ satisfies $\|\hat{\mu}-\mu^*\|_1 \le \frac{6|\Omega|\varepsilon}{p_0^3}$. Using Corollary 2.1 on $\hat{\mu}$ with parameter $\frac{6|\Omega|\varepsilon}{p_0^3}$ (the condition $\frac{6|\Omega|\varepsilon}{p_0^3} \le \frac{p_0^2 D}{4}$ is satisfied by our choice of ε), the constructed scheme $\tilde{\pi}$ is persuasive for μ^* and is

$$\frac{6\cdot(\frac{6|\Omega|\varepsilon}{p_0^3})}{p_0^2 D} = \frac{36|\Omega|\varepsilon}{p_0^5 D}$$

-approximately optimal for μ^*. Thus, the regret of Algorithm 2.4 is upper bounded by

$$\begin{aligned}\mathrm{Reg}(T) \;&\le\; \underbrace{\mathbb{E}[T_0\cdot 1]}_{\text{regret in the first } T_0 \text{ periods}} \;+\; \underbrace{(T-\mathbb{E}[T_0])\cdot\frac{36|\Omega|\varepsilon}{p_0^5 D}}_{\text{regret in the remaining periods}}\\ &\le\; \frac{2|\Omega|}{p_0}\log_2\frac{1}{G^2 p_0\varepsilon} \;+\; T\cdot\frac{36|\Omega|\varepsilon}{p_0^5 D}\\ &\le\; \frac{2|\Omega|}{p_0}\log_2\frac{24|\Omega|T}{G^2 p_0^6 D} \;+\; 2\end{aligned}$$

with $\varepsilon = \frac{p_0^5 D}{24|\Omega|T}$. □

We provide two remarks on Theorem 2.1. First, the designed signaling schemes in the first T_0 periods are allowed to be non-persuasive and far from optimal. The goal of the sender in the first T_0 periods is to elicit useful information from the actions taken by receivers to estimate the prior, rather than to achieve persuasiveness and optimality.

Second, instead of estimating prior ratio by binary search (as in Algorithm 2.3), one might consider other multi-dimensional binary search methods, such as the ellipsoid method, to estimate the receiver's prior μ^*. However, the classical ellipsoid method for estimating a vector in $\mathbb{R}^{|\Omega|}$ with precision ε has query complexity $O(|\Omega|^2\log\frac{1}{\varepsilon})$, which is quadratic in the dimension $|\Omega|$ and worse than the $O(|\Omega|\log\frac{1}{\varepsilon})$ query complexity of our Algorithm 2.3.

2.4 Discussion

In summary, this chapter provides a learning perspective to the problem of information design with unknown subjective prior, complementing the distributional and worst-case perspectives in previous work. Using the idea of learning from revealed preference, we design an algorithm for the principal to achieve a logarithmic regret, circumventing the limitation of empirical estimation. Some directions for future work are given below:

- The receiver in our model best responds in each period myopically. Knowing that the sender is learning, the receiver may want to strategize for the long term. Such long-term strategic behavior might lead to interesting yet complicated equilibria.

- Our regret bound $O(\frac{1}{p_0}\log T)$ depends on the minimal prior probability p_0 of a state. Such an instance-dependent result is similar to the $O(\frac{1}{\Delta}\log T)$ regret bound in stochastic multi-armed bandit problems where Δ is the gap between the expected rewards of the optimal arm and the second-optimal arm. So, we believe that our instance-dependent result is well motivated. One way to remove the dependency on p_0 is to intentionally set $p_0 = O(\sqrt{\frac{\log T}{T}})$ and ignore the states whose prior probability is less than p_0. This would give a regret bound of $O(\frac{1}{p_0}\log T) + O(Tp_0) = O(\sqrt{T\log T})$. Whether a better result can be obtained is open.

- Can a lower bound on the sender's regret in the form of $\Omega(\frac{1}{p_0}\log T)$ or $\Omega(\sqrt{T})$ be proven? The classical information-theoretical approach to proving $\Omega(\sqrt{T})$ lower bound does not apply directly to our problem because the best-responding action of the receiver contains significant information about the prior distribution μ^*.

- Besides learning from revealed preference, another approach to learning the receiver's subjective prior is direct elicitation. To prevent the receiver from misreporting, some mechanism design ideas are needed [KMZL17, BBS18], and we expect to see here a fascinating mixture of information design, mechanism design, and machine learning.

2.5 Omitted Proofs in this Chapter

2.5.1 Useful Lemmas

Lemma 2.5 (Continuity of posterior)**.** *Let* $\pi : \Omega \to \Delta(S)$ *be any signaling scheme. Let* $\mu, \mu' \in \Delta(\Omega)$ *be two priors. Let* μ_s, μ'_s *be the posterior belief induced by signal* s *under* π *and prior* μ, μ' *respectively. Suppose* $\min_{\omega\in\Omega} \mu(\omega) \geq p_0 > 0$*. Then* $\|\mu_s - \mu'_s\|_1 \leq \frac{2}{p_0}\|\mu - \mu'\|_1$*.*

Proof. Let $\pi(s) = \sum_{\omega\in\Omega} \mu(\omega)\pi(s|\omega)$ and $\pi'(s) = \sum_{\omega\in\Omega} \mu'(\omega)\pi(s|\omega)$ be the probability of signal s under prior μ and μ' respectively. By the definition of μ_s, μ'_s and by triangle inequality,

$$\begin{aligned}\|\mu_s - \mu'_s\|_1 &= \sum_{\omega\in\Omega} \big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\big| \\ &\leq \sum_{\omega\in\Omega} \big|\tfrac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| + \sum_{\omega\in\Omega} \big|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)} - \tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\big|.\end{aligned}$$

For the first term above, $\sum_{\omega\in\Omega} \big|\frac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \frac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| = \sum_{\omega\in\Omega} \frac{\pi(s|\omega)}{\pi(s)}|\mu(\omega) - \mu'(\omega)|$. We note that, $\forall \omega \in \Omega$,

$$\tfrac{\pi(s|\omega)}{\pi(s)} = \tfrac{\pi(s|\omega)}{\sum_{\omega'\in\Omega} \mu(\omega')\pi(s|\omega')} \leq \tfrac{\pi(s|\omega)}{p_0 \sum_{\omega'\in\Omega} \pi(s|\omega')} \leq \tfrac{1}{p_0}. \tag{2.9}$$

Thus, $\sum_{\omega\in\Omega} \big|\frac{\mu(\omega)\pi(s|\omega)}{\pi(s)} - \frac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}\big| \leq \sum_{\omega\in\Omega} \frac{1}{p_0}|\mu(\omega) - \mu'(\omega)| = \frac{1}{p_0}\|\mu - \mu'\|_1$.

For the second term,

$$\begin{aligned}
\sum_{\omega\in\Omega}\Big|\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi(s)}-\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\Big| &= \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\Big|\tfrac{\pi'(s)-\pi(s)}{\pi(s)\pi'(s)}\Big| \\
&= \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\Big|\tfrac{\sum_{\omega'\in\Omega}(\mu'(\omega')-\mu(\omega'))\pi(s|\omega')}{\pi(s)\pi'(s)}\Big| \\
&\le \sum_{\omega\in\Omega}\mu'(\omega)\pi(s|\omega)\tfrac{\sum_{\omega'\in\Omega}|\mu'(\omega')-\mu(\omega')|\cdot\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)\pi'(s)} \\
&= \|\mu'-\mu\|_1\sum_{\omega\in\Omega}\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\tfrac{\max_{\omega'\in\Omega}\pi(s|\omega')}{\pi(s)} \\
\text{by (2.9)}\quad &\le \|\mu'-\mu\|_1\sum_{\omega\in\Omega}\tfrac{\mu'(\omega)\pi(s|\omega)}{\pi'(s)}\tfrac{1}{p_0} \\
&= \tfrac{1}{p_0}\|\mu'-\mu\|_1.
\end{aligned}$$

Therefore, we obtain $\|\mu_s-\mu'_s\|_1\le\frac{2}{p_0}\|\mu'-\mu\|_1$. □

2.5.2 Proof of Lemma 2.1

Let $\hat{\mu}\in\Delta(\Omega)$ be a receiver prior satisfying $\min_{\omega\in\Omega}\hat{\mu}(\omega)\ge p_0>0$. Let $B_1(\hat{\mu},\varepsilon)=\{\mu:\|\mu-\hat{\mu}\|_1\le\varepsilon\}$ be the set of priors within distance ε to $\hat{\mu}$. Suppose $\varepsilon\le\frac{p_0^2D}{4}$. Let $\hat{\pi}$ be a persuasive signaling scheme for receiver prior $\hat{\mu}$. We want to construct another signaling scheme $\tilde{\pi}$ that is persuasive for all receiver priors in $B_1(\hat{\mu},\varepsilon)$ and satisfies $U(\mu_0,\hat{\mu},\tilde{\pi})\ge U(\mu_0,\hat{\mu},\hat{\pi})-\frac{6\varepsilon}{p_0^2D}$. We do this in two steps: (1) First, we construct a signaling scheme $\tilde{\pi}$ that satisfies the requirements but is not a direct a signaling scheme; (2) Then, convert $\tilde{\pi}$ into a direct signaling scheme that still satisfies the requirements.

Step (1): Construct a non-direct signaling scheme $\tilde{\pi}$. Let

$$\delta=\tfrac{2\varepsilon}{p_0D}\le\tfrac{p_0}{2}.$$

Let $\hat{\pi}(a)$ be the unconditional probability that $\hat{\pi}$ sends signal a under prior $\hat{\mu}$: $\hat{\pi}(a) = \sum_{\omega\in\Omega}\hat{\mu}(a)\hat{\pi}(a|\omega)$. Let $\hat{\mu}_a \in \Delta(\Omega)$ be the posterior belief induced by signal a under prior $\hat{\mu}$:

$$\hat{\mu}_a(\omega) = \frac{\hat{\mu}(\omega)\hat{\pi}(a|\omega)}{\hat{\pi}(a)}, \quad \forall \omega \in \Omega.$$

Since $\hat{\pi}$ is persuasive, a must be an optimal action for the receiver on posterior $\hat{\mu}_a$:

$$\mathbb{E}_{\omega\sim\hat{\mu}_a}[v(a,\omega) - v(a',\omega)] \geq 0, \ \forall a' \neq a.$$

According to Assumption 2.2, there exists a belief $\eta_a \in \Delta(\Omega)$ for which $\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \geq D$. Consider the convex combination of $\hat{\mu}_a$ and η_a with coefficients $1-\delta, \delta$:

$$\xi_a = (1-\delta)\hat{\mu}_a + \delta\eta_a.$$

By the linearity of expectation, a must be better than any other action a' by δD on belief ξ_a:

$$\begin{aligned}&\mathbb{E}_{\omega\sim\xi_a}[v(a,\omega) - v(a',\omega)] \\ &= (1-\delta)\mathbb{E}_{\omega\sim\hat{\mu}_a}[v(a,\omega) - v(a',\omega)] + \delta\mathbb{E}_{\omega\sim\eta_a}[v(a,\omega) - v(a',\omega)] \ \geq \ \delta D.\end{aligned} \tag{2.10}$$

Let $\xi = \sum_{a\in A}\hat{\pi}(a)\xi_a \in \Delta(\Omega)$, and write $\hat{\mu}$ as the convex combination of ξ and another belief $\chi \in \Delta(\Omega)$:

$$\hat{\mu} \ = \ (1-y)\xi + y\chi \ = \ \sum_{a\in A}(1-y)\hat{\pi}(a)\xi_a \ + \ y\chi. \tag{2.11}$$

Lemma 2.6 (Proposition 1 of [ZIX21])**.** *If $\delta \le \frac{p_0}{2}$, then there exist χ on the boundary of $\Delta(\Omega)$ and $y \le \frac{\delta}{p_0} \le \frac{1}{2}$ that satisfy* (2.11)*.*

Since (2.11) is a convex decomposition of the prior $\hat{\mu}$, according to [KG11], there exists a signaling scheme $\tilde{\pi}$ that induces posterior ξ_a with probability $(1-y)\hat{\pi}(a)$, for $a \in A$, and the posterior that puts all probability on ω with probability $y\chi(\omega)$, for $\omega \in \Omega$. Namely, $\tilde{\pi}$ has signal space $S = A \cup \Omega$ and signal probability

$$\tilde{\pi}(s|\omega) = \begin{cases} \frac{(1-y)\hat{\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} & \text{for } s = a \in A; \\ \frac{y\chi(\omega)}{\hat{\mu}(\omega)} & \text{for } s = \omega \in \Omega; \\ 0 & \text{otherwise.} \end{cases}$$

It is not hard to verify that, under prior $\hat{\mu}$ and signaling scheme $\tilde{\pi}$, the posterior induced by signal $a \in A$ is equal to ξ_a, and the posterior induced by signal ω is the deterministic distribution on ω.

We show that, whenever $\tilde{\pi}$ sends an action recommendation $a \in A$, the recommendation is persuasive, under all receiver priors in $B_1(\hat{\mu}, \varepsilon)$.

Claim 2.1. *For all receiver priors $\mu \in B_1(\hat{\mu}, \varepsilon)$, any action recommendation $a \in A$ from $\tilde{\pi}$ is persuasive.*

Proof. Under signaling scheme $\tilde{\pi}$, the posteriors induced by signal a from priors μ, $\hat{\mu}$ are equal to μ_a, ξ_a, respectively. By the continuity of posterior (Lemma 2.5), we have $\|\mu_a - \xi_a\|_1 \le \frac{2}{p_0}\|\mu - \hat{\mu}\|_1 \le \frac{2\varepsilon}{p_0}$. Then, since the receiver's utility is in $[0,1]$, for any action $a' \ne a$, we have

$$\left|\mathbb{E}_{\omega\sim\mu_a}[v(a,\omega) - v(a',\omega)] - \mathbb{E}_{\omega\sim\xi_a}[v(a,\omega) - v(a',\omega)]\right| \le \|\mu_a - \xi_a\|_1 \le \tfrac{2\varepsilon}{p_0}.$$

Together with (2.10), we get $\mathbb{E}_{\omega\sim\mu_a}[v(a,\omega) - v(a',\omega)] \geq \delta D - \frac{2\varepsilon}{p_0} \geq 0$. Thus, the action recommendation a is persuasive by definition. □

We show that the signaling scheme $\tilde{\pi}$ is close to $\hat{\pi}$, in the following sense:

Claim 2.2. *For any $a \in A, \omega \in \Omega$, $|\tilde{\pi}(a|\omega) - \hat{\pi}(a|\omega)| \leq \frac{3\delta}{p_0}$.*

Proof. By definition,

$$\begin{aligned}
\left|\tilde{\pi}(a|\omega) - \hat{\pi}(a|\omega)\right| &= \left|\frac{(1-y)\hat{\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} - \frac{\hat{\pi}(a)\hat{\mu}_a(\omega)}{\hat{\mu}(\omega)}\right| \\
&\leq \frac{\hat{\pi}(a)}{\hat{\mu}(\omega)} \cdot |\xi_a(\omega) - \hat{\mu}_a(\omega)| \;+\; y \cdot \frac{\hat{\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} \\
&\leq \frac{\hat{\pi}(a)}{\hat{\mu}(\omega)} \cdot \delta \;+\; y \cdot \frac{\hat{\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} \\
&\leq \frac{\hat{\pi}(a)}{\hat{\mu}(\omega)} \cdot \delta + y \cdot 2 \\
&\leq \frac{1}{p_0} \cdot \delta + \frac{\delta}{p_0} \cdot 2 \;=\; \frac{3\delta}{p_0}
\end{aligned}$$

where the third "$\leq$" is because $\frac{(1-y)\hat{\pi}(a)\xi_a(\omega)}{\hat{\mu}(\omega)} \leq 1$ and $y \leq \frac{1}{2}$. □

Then, we show that the sender's utility under scheme $\tilde{\pi}$ is close to her utility under $\hat{\pi}$, which means that $\tilde{\pi}$ is an approximately optimal scheme for $\hat{\mu}$.

Claim 2.3. *The sender's utility $U(\mu_0, \hat{\mu}, \tilde{\pi}) \geq U(\mu_0, \hat{\mu}, \hat{\pi}) - \frac{3\delta}{p_0} = U(\mu_0, \hat{\mu}, \hat{\pi}) - \frac{6\varepsilon}{p_0^2 D}$.*

Proof. According to Claim 2.1, $\tilde{\pi}$ is persuasive for the receiver prior $\hat{\mu}$, so the sender's utility satisfies

$$\begin{aligned} U(\mu_0, \hat{\mu}, \tilde{\pi}) &= \sum_{\omega \in \Omega} \mu_0(\omega) \Big(\sum_{a \in A} \tilde{\pi}(a|\omega) u(a, \omega) \ + \ \tilde{\pi}(\omega|\omega) u(a^*(\omega), \omega) \Big) \\ &\geq \sum_{\omega \in \Omega} \mu_0(\omega) \sum_{a \in A} \tilde{\pi}(a|\omega) u(a, \omega) \\ &\geq \sum_{\omega \in \Omega} \mu_0(\omega) \sum_{a \in A} \hat{\pi}(a|\omega) u(a, \omega) \ - \ \tfrac{3\delta}{p_0} \cdot 1 \qquad \text{by Claim 2.2} \\ &= U(\mu_0, \hat{\mu}, \hat{\pi}) - \tfrac{3\delta}{p_0} \\ &= U(\mu_0, \hat{\mu}, \hat{\pi}) - \tfrac{6\varepsilon}{p_0^2 D}. \end{aligned}$$

□

Thus, $\tilde{\pi}$ satisfies the requirements we want.

Step (2): Convert $\tilde{\pi}$ into a direct signaling scheme. Then, we convert $\tilde{\pi}$ (whose signal space is $A \cup \Omega$) into a direct signaling scheme π^c (with signal space A) by coalescing the signals. Specifically, for each state ω, let π^c send signal a whenever $\tilde{\pi}$ sends signal a and signal ω if $a^*(\omega) = a$:

$$\pi^c(a|\omega) = \tilde{\pi}(a|\omega) + \mathbb{1}[a = a^*(\omega)] \cdot \tilde{\pi}(\omega|\omega). \tag{2.12}$$

Note that π^c is a valid signaling scheme because $\sum_{a \in A} \pi^c(a|\omega) = \sum_{a \in A} \tilde{\pi}(a|\omega) + \tilde{\pi}(\omega|\omega) = 1$. The coalesced signaling scheme π^c has the following property:

Lemma 2.7 (Coalescing)**.** *If all the action recommendations from $\tilde{\pi}$ are persuasive, then π^c is persuasive. And the sender's utility satisfies $U(\mu_0, \mu, \tilde{\pi}) = U(\mu_0, \mu, \pi^c)$.*

Proof. Under receiver prior μ and signaling scheme π^c, upon receiving signal a, for any action $a' \neq a$ we have

$$\begin{aligned}
&\sum_{\omega\in\Omega}\mu(\omega)\pi^c(a|\omega)[v(a,\omega)-v(a',\omega)]\\
&= \underbrace{\sum_{\omega\in\Omega}\mu(\omega)\tilde{\pi}(a|\omega)[v(a,\omega)-v(a',\omega)]}_{\geq 0 \text{ because } \tilde{\pi} \text{ is persuasive for } \mu} + \sum_{\omega\in\Omega}\mu(\omega)\underbrace{\mathbb{1}[a=a^*(\omega)]\tilde{\pi}(\omega|\omega)[v(a,\omega)-v(a',\omega)]}_{\geq 0 \text{ given } a=a^*(\omega)}\\
&\geq 0.
\end{aligned}$$

Therefore, π^c is persuasive for μ.

The sender's utility satisfies

$$\begin{aligned}
U(\mu_0,\mu,\pi^c) &= \sum_{\omega\in\Omega}\mu_0(\omega)\sum_{a\in A}\pi^c(a|\omega)u(a,\omega)\\
&= \sum_{\omega\in\Omega}\mu_0(\omega)\Big(\sum_{a\in A}\tilde{\pi}(a|\omega)u(a,\omega)+\tilde{\pi}(\omega|\omega)u(a^*(\omega),\omega)\Big) = U(\mu_0,\mu,\tilde{\pi})
\end{aligned}$$

by definition. □

Because $\tilde{\pi}$ is persuasive and satisfies $U(\mu_0,\hat{\mu},\tilde{\pi}) \geq U(\mu_0,\hat{\mu},\hat{\pi}) - \frac{6\varepsilon}{p_0^2 D}$, by the coalescing lemma, π^c satisfies the same properties, which proves Lemma 2.1.

2.5.3 Proof of Lemma 2.2

First, we note that the following two claims always hold during Algorithm 2.1:

- If the sender uses a signaling scheme π that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = \ell^{(k)}$, then the receiver will take action a_1 when signal s_0 is sent.
- If the sender uses a signaling scheme π that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = r^{(k)}$, then the receiver will take action $\tilde{a} \neq a_1$ when signal s_0 is sent.

These two claims hold automatically for $k \geq 1$ according to the definition of $\pi^{(k)}$, $\tilde{a}$, $\ell^{(k)}$, and $r^{(k)}$. So we only need to prove the two claims for $k = 0$. When $k = 0$, if the sender uses a signaling scheme π that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = \ell^{(0)} = 0$, then whenever signal s_0 is sent, the sender will believe that the state is ω_2 with probability 0 and is ω_1 with probability 1. Since $a_1 = \arg\max_{a\in A} v(a, \omega_1)$, the receiver will take a_1. If the sender uses a signaling scheme π that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = r^{(0)} = \frac{1}{Gp_0}$, then whenever signal s_0 is sent, the difference between the receiver's utility of taking action $\tilde{a} = \arg\max_{a\in A} v(a, \omega_2)$ and any action $a \neq \tilde{a}$ is

$$\begin{aligned}
&\sum_{\omega\in\Omega} \mu^*(\omega)\pi(s_0|\omega)\big[v(\tilde{a}, \omega) - v(a, \omega)\big] \\
&= \underbrace{\mu^*(\omega_1)}_{\le 1}\pi(s_0|\omega_1)\big[\underbrace{v(\tilde{a}, \omega_1) - v(a, \omega_1)}_{\ge -1}\big] + \underbrace{\mu^*(\omega_2)}_{\ge p_0}\pi(s_0|\omega_2)\big[\underbrace{v(\tilde{a}, \omega_2) - v(a, \omega_2)}_{>G}\big] \\
&> -\pi(s_0|\omega_1) + p_0 G \cdot \pi(s_0|\omega_2) \;=\; 0.
\end{aligned}$$

Therefore, the receiver will take action $\tilde{a}$.

Suppose the while loop of Algorithm 2.1 has finished. Due to the above two claims, if the sender uses a signaling scheme that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = \ell^{(k)}$, then when s_0 is sent the receiver will take action a_1, so the receiver's utility difference between a_1 and $\tilde{a}$ is ≥ 0:

$$\begin{aligned}
&\mu^*(\omega_1)\pi(s_0|\omega_1)\big[\underbrace{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)}_{>G>0}\big] + \mu^*(\omega_2)\pi(s_0|\omega_2)\big[v(a_1, \omega_2) - v(\tilde{a}, \omega_2)\big] \ge 0 \\
&\implies \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} \ge \frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} \cdot \frac{v(\tilde{a}, \omega_2) - v(a_1, \omega_2)}{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)} = \ell^{(k)} \cdot \frac{v(\tilde{a}, \omega_2) - v(a_1, \omega_2)}{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)} = \hat{\rho}.
\end{aligned}$$

If the sender uses a signaling scheme that satisfies $\frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} = r^{(k)}$, then when s_0 is sent the receiver will take action $\tilde{a}$, so the utility difference between a_1 and $\tilde{a}$ is ≤ 0:

$$\begin{aligned}
&\mu^*(\omega_1)\pi(s_0|\omega_1)\big[\underbrace{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)}_{>G>0}\big] + \mu^*(\omega_2)\pi(s_0|\omega_2)\big[v(a_1, \omega_2) - v(\tilde{a}, \omega_2)\big] \le 0 \\
&\implies \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} \le \frac{\pi(s_0|\omega_2)}{\pi(s_0|\omega_1)} \cdot \frac{v(\tilde{a}, \omega_2) - v(a_1, \omega_2)}{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)} = r^{(k)} \cdot \frac{v(\tilde{a}, \omega_2) - v(a_1, \omega_2)}{v(a_1, \omega_1) - v(\tilde{a}, \omega_1)}.
\end{aligned}$$

So, given $r^{(k)} - \ell^{(k)} \le \varepsilon G$, we have

$$\frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} - \hat{\rho} \;\le\; \big(r^{(k)} - \ell^{(k)}\big)\frac{v(\tilde{a},\omega_2) - v(a_1,\omega_2)}{v(a_1,\omega_1) - v(\tilde{a},\omega_1)} \;\le\; \varepsilon G \cdot \frac{1}{G} \;=\; \varepsilon.$$

This means that the output $\hat{\rho}$ of Algorithm 2.1 satisfies $\hat{\rho} \le \frac{\mu^*(\omega_1)}{\mu^*(\omega_2)} \le \hat{\rho} + \varepsilon$.

Then, we consider the expected number of periods needed by Algorithm 2.1. First, we note that, after each while loop, the difference $r^{(k)} - \ell^{(k)}$ shrinks by a half. The algorithm terminates when $r^{(k)} - \ell^{(k)} \le \varepsilon G$, so the total number k of while loops is at most

$$k \;\le\; \log_2 \frac{r^{(0)} - \ell^{(0)}}{\varepsilon G} \;=\; \log_2 \frac{1}{G^2 p_0 \varepsilon}. \tag{2.13}$$

Then, we consider the number of periods in each while loop. By definition, this is equal to the number of periods until signal s_0 is sent, whose expectation depends on the signaling scheme $\pi^{(k)}$. We construct $\pi^{(k)}$ as follows:

- If $q \le 1$, then let $\pi^{(k)}(s_0|\omega_2) = q$, $\pi^{(k)}(s_0|\omega_1) = 1$;
- If $q > 1$, then let $\pi^{(k)}(s_0|\omega_2) = 1$, $\pi^{(k)}(s_0|\omega_1) = 1/q$.

Note that this construction satisfies $\frac{\pi^{(k)}(s_0|\omega_2)}{\pi^{(k)}(s_0|\omega_1)} = q$, as needed in Algorithm 2.1. Since one of $\pi^{(k)}(s_0|\omega_2)$ and $\pi^{(k)}(s_0|\omega_1)$ is 1, the probability that s_0 is sent in each period is at least:

$$\pi^{(k)}(s_0) \;=\; \mu_0(\omega_1)\pi^{(k)}(s_0|\omega_1) + \mu_0(\omega_2)\pi^{(k)}(s_0|\omega_2) \;\ge\; \min\{\mu_0(\omega_1), \mu_0(\omega_2)\} \;\ge\; p_0.$$

So, by the property of geometric random variable, the expected number of periods until a signal s_0 is sent (namely, the expected number of periods in each while loop) is at most

$$\frac{1}{\pi^{(k)}(s_0)} \;\le\; \frac{1}{p_0}.$$

So, the total number of periods does not exceed $k \cdot \frac{1}{p_0} \;\le\; \frac{1}{p_0}\log_2 \frac{1}{G^2 p_0 \varepsilon}$ in expectation.

2.5.4 Proof of Lemma 2.3

If (ω_i, ω_j) is a pair of distinguishable states, then Lemma 2.2 shows that the estimate $\hat{\rho}_{ij}$ returned by Algorithm 2.1 satisfies $|\hat{\rho}_{ij} - \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}| \le \varepsilon \le \frac{2\varepsilon}{p_0^2}$.

If (ω_i, ω_j) is not a pair of distinguishable states, then by Lemma 2.2, we have the ratio estimates $\hat{\rho}_{ik}$ satisfying $\hat{\rho}_{ik} \le \frac{\mu^*(\omega_i)}{\mu^*(\omega_k)} \le \hat{\rho}_{ik} + \varepsilon$ and $\hat{\rho}_{jk}$ satisfying $\hat{\rho}_{jk} \le \frac{\mu^*(\omega_j)}{\mu^*(\omega_k)} \le \hat{\rho}_{jk} + \varepsilon$. So,

$$\hat{\rho}_{ij} \;=\; \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}} \;\le\; \frac{\frac{\mu^*(\omega_i)}{\mu^*(\omega_k)}}{\frac{\mu^*(\omega_j)}{\mu^*(\omega_k)} - \varepsilon} \;=\; \frac{\frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}}{1 - \varepsilon\frac{\mu^*(\omega_k)}{\mu^*(\omega_j)}}.$$

For real numbers $a \ge 0$ and $0 \le b \le \frac{1}{2}$, we have inequality $\frac{a}{1-b} = \frac{a(1-b)+ab}{1-b} = a + \frac{ab}{1-b} \le a + 2ab$. Under the assumption of $\varepsilon \le \frac{p_0}{2}$, we have $\varepsilon\frac{\mu^*(\omega_k)}{\mu^*(\omega_j)} \le \varepsilon\frac{1}{p_0} \le \frac{1}{2}$. So, we obtain

$$\hat{\rho}_{ij} \;\le\; \frac{\frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}}{1 - \varepsilon\frac{\mu^*(\omega_k)}{\mu^*(\omega_j)}} \;\le\; \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)} + 2\varepsilon\frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}\frac{\mu^*(\omega_k)}{\mu^*(\omega_j)} \;\le\; \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)} + \frac{2\varepsilon}{p_0^2}.$$

On the other hand,

$$\hat{\rho}_{ij} \;=\; \frac{\hat{\rho}_{ik}}{\hat{\rho}_{jk}} \;\ge\; \frac{\frac{\mu^*(\omega_i)}{\mu^*(\omega_k)} - \varepsilon}{\frac{\mu^*(\omega_j)}{\mu^*(\omega_k)}} \;=\; \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)} - \varepsilon\frac{\mu^*(\omega_k)}{\mu^*(\omega_j)} \;\ge\; \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)} - \frac{\varepsilon}{p_0} \;\ge\; \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)} - \frac{2\varepsilon}{p_0^2}.$$

So, we have $|\hat{\rho}_{ij} - \frac{\mu^*(\omega_i)}{\mu^*(\omega_j)}| \le \frac{2\varepsilon}{p_0^2}$.

Chapter 3

Information Design with Unknown Bias

joint work with

Yiling Chen, Ariel Procaccia

Aaditya Ramdas, Itai Shapira [CLP+24]

This chapter continues to study information design by a learning principal. While the previous chapter considers an agent with unknown subjective prior, this chapter considers an agent with the correct prior but an unknown bias when performing Bayes update.

3.1 Introduction

We begin with an example.

> *Guessing fair coin.* A bag contains two coins that look and feel identical, but one is a fair coin that, on a flip, comes up heads with probability 0.5, and the other is an unfair coin with probability 0.9 of heads. You reach into the bag, grab one of the coins and flip it once; it lands on heads. Since you are (hopefully) familiar with Bayes' rule, you conclude that the probability you are holding the fair coin is ≈ 0.36. Now suppose you are offered the following deal: if you pay \$1, you get to flip the same coin again, and if it comes up heads, you will receive \$1.4. Since you now believe that the probability of heads is 0.76, you take the deal (assuming you are risk-neutral) and earn 6 cents in expectation.

> If, by contrast, another risk-neutral person in the same situation decides to decline the same deal, they must believe that the probability they are holding the fair coin is greater than 0.47. That is, their belief is still very close to the prior of 0.5. We think of such a person as being *biased*, in the sense that they are unwilling to significantly update their beliefs, despite evidence to the contrary. □

Of course, failing to update one's beliefs about coin flips is not the end of the world. But this example serves to illustrate a broader phenomenon that, in our view, is both important and ubiquitous. In particular, the "stickiness" of — think of the controversy over Russian collusion in the 2016 US presidential election or the existence of weapons of mass destruction in Iraq in 2003. It is also prevalent in science, as exemplified by the polarized debate over the origins of the Covid pandemic [BCB+21].

The goal of this work is to develop algorithms that are able to *detect* bias in the form of non-Bayesian updating of beliefs. To our knowledge, we are the first to formalize and analytically address this problem, and we aim to build an initial framework that future work would build on. In the long term, we believe such algorithms could have many applications, including understanding to what degree the foregoing type of bias contributes to disagreement and polarization, and discounting the opinions of biased agents to improve collective decision making.

Our Approach. The first question we need to answer is how to *quantify* bias. In this first investigation, we adopt a linear model of bias that was proposed and used as a general belief updating model in economics [ENS10, HL17, dCZ22, TH21]. If the prior is μ_0 and the correct Bayesian posterior upon receiving a *signal* (or evidence) s is denoted μ_s, we posit that an agent with bias $w \in [0, 1]$ adopts the belief $w\mu_0 + (1 - w)\mu_s$. At the extremes, an agent with bias $w = 0$ performs perfect Bayesian updating and an agent with bias $w = 1$ cannot be convinced to budge from the prior.

The bigger conceptual question is how we can infer an agent's bias. To address it, we take an approach that is inspired by *information design* [KG11]. In our context, suppose that we (the *principal*) and the agent have asymmetric information: while both share a common (say public) prior about the state of the world, the principal knows the true (realized) state of the world, but the agent does not. The *principal* publicly commits to a (randomized) *signaling scheme* that specifies the probability of sending each possible signal given each possible realized state of the world. Given their knowledge of the latter, the principal draws a signal from the specified distribution and sends it to the agent. Upon receiving such a signal, the agent updates their beliefs about the state of the world (from the common prior) and then takes an *action* that maximizes their expected utility according to a given utility function. Similarly to the example we started with, it is the action taken by the agent that can (indirectly) reveal their degree of bias.

Note that the problem of estimating the exact level of bias reduces to the problem of detecting whether the agent's bias is above or below some threshold. Indeed, to estimate the level of bias to an accuracy of ϵ, $\log(1/\epsilon)$ such threshold queries suffice by using binary search. The challenge, then, is to design signaling schemes that test whether bias is above or below a given threshold in the most efficient way, that is, using a minimum number of signals in expectation.

Our Results. We design a polynomial-time algorithm that computes optimal signaling schemes, in Section 3.4. We first show that *constant* algorithms, which repeatedly use the same signaling scheme, are as powerful as *adaptive* algorithms, which can vary the scheme over time based on historical data (Lemma 3.2); we can therefore restrict our attention to constant algorithms. In Lemma 3.6, we establish a version of the *revelation principle* for the bias detection problem, which asserts that optimal signaling schemes need only use signals that can be interpreted as action recommendations. Finally, building on these insights, we show that the optimal solution to our problem is obtained by solving a "small" linear program (Algorithm 3.1 and Theorem 3.2).

In Section 3.5, we present a geometric characterization of optimal signaling schemes (Theorem 3.3), which sheds additional light on the performance of the algorithm. In particular, the characterization provides sufficient and necessary conditions for the testability of bias, and also identifies cases where only a single sample is needed for this task.

3.1.1 Related Work

There is a significant body of *experimental* work in the social sciences aiming to explain the failure of partisans to reach similar beliefs on factual questions where there is a large amount of publicly available evidence. The fact that biased belief updating occurs is undisputed, and the focus is on understanding the factors that play a role. In particular, a prominent line of work supports the (perhaps counterintuitive) hypothesis that the more cognitively sophisticated a partisan is, the more politically biased is their belief update process [TL06, TCK09, KPW$^+$12, Kah13]. These results are challenged by more recent work by [TPR20], who found that greater analytical thinking is associated with belief updates that are less biased, using an experimental design that explicitly measures the proximity of belief updates to a correct Bayesian posterior. While these studies provide empirical underpinnings for our theoretical model, their research questions are orthogonal to ours: we aim to measure the magnitude of bias regardless of its source.

Classical work in *information design* [CS82, KG11] studies how a principal can strategically provide information to induce an agent to take actions that are beneficial for the principal, assuming a perfectly Bayesian agent. Various relaxations of the perfectly Bayesian assumption have been investigated [AC16a, HL17, DP22, dCZ22, FHT24, YZ24, LC25]. The work [dCZ22] is closet to us, which studies biased belief update models including the linear model. However, their goal is to maximize the principal's utility with the agent's bias fully known. In our problem the agent's bias level is unknown, and the principal's goal is to infer the agent's bias level instead of maximizing their own utility. [TH21] present real-world experiments showing that human belief updates are close to a linear bias model, which supports our theoretical assumption.

3.2 Model: Bias Detection

Biased agent. Consider a standard Bayesian setting: the relevant *state of the world* is $\theta \in \Theta$, distributed according to some known prior distribution μ_0. If an agent were perfectly Bayesian, when receiving some new information ("signal") s and with the knowledge of the conditional distributions $P(s|\theta)$ for all θ, they would update their belief about the state of the world according to Bayes' Rule: $\mu_s(\theta) = P(\theta|s) = \frac{\mu_0(\theta)P(s|\theta)}{P(s)}$. We refer to μ_s as the true posterior belief induced by s. Being biased, the agent's belief after seeing s, denoted ν_s, is a convex combination of μ_s and μ_0:

$$\nu_s = w\mu_0 + (1-w)\mu_s,$$

where $w \in [0,1]$ is the unknown *bias level*, capturing the agent's inclination to retain their prior belief in the presence of new information. This linear model was proposed and adopted in economics for non-Bayesian belief updating [ENS10, HL17], to capture people's conservatism in processing new information and their tendency to protect their beliefs [War07].

The agent can choose an action from a finite set A and has a state-dependent utility function $U : A \times \Theta \to \mathbb{R}$. They receive utility $U(a,\theta)$ when taking action a in state θ. The agent will act according to their (biased) belief ν_s and choose an action a that maximizes their expected utility:

$$a \in \arg\max_{a \in A} \mathbb{E}_{\theta \sim \nu_s}[U(a,\theta)] = \arg\max_{a \in A} \sum_{\theta \in \Theta} \nu_s(\theta)U(a,\theta).$$

In the absence of any additional information, the agent operates based on the prior belief μ_0 and will select an action deemed optimal with respect to μ_0. We introduce the following mild assumption to ensure the uniqueness of this action:

Assumption 3.1. *There is a unique action that maximizes the expected utility based on the prior belief* μ_0*:* $|\arg\max_{a \in A}\{\sum_{\theta \in \Theta} \mu_0(\theta)U(a,\theta)\}| = 1$.

This assumption will be made throughout. We denote the unique optimal action on the prior belief as $a_0 = \arg\max_{a \in A}\{\sum_{\theta \in \Theta} \mu_0(\theta)U(a,\theta)\}$, and call it the *default action.*

Bias detection. The principal, who knows the prior μ_0 and the agent's utility function U, seeks to infer the agent's bias level from their action as efficiently as possible. The principal has an informational advantage — they observe the independent realizations of the state of the world at each time step. In other words, the principal knows θ_t, an independent sample drawn according to μ_0 at time t. The principal wants to design signaling schemes to strategically reveal information about θ_t to the agent, hoping to influence the agent's biased belief in a way that the agent's chosen actions reveal information about their bias level. Specifically, with a finite signal space S, the principal can commit to a *signaling scheme* $\pi_t : \Theta \to \Delta(S)$ at time t, where $\pi_t(s|\theta)$ specifies the probability of sending signal s in state θ at time t. After seeing a signal s_t, drawn according to $\pi_t(s|\theta_t)$ at time t, the agent takes action a_t that is optimal for their biased belief ν_{s_t}. The principal infers information about bias w from the history of signaling schemes, realized states, realized signals, and agent actions $\mathcal{H}_t = \{(\pi_1, \theta_1, s_1, a_1), \dots, (\pi_t, \theta_t, s_t, a_t)\}$. We denote by Π an adaptive algorithm that the principal uses to decide on the signaling scheme at time $t+1$ based on history $\mathcal{H}_t$.

Given a threshold $\tau \in (0,1)$, the principal wants to design Π to answer the following question:

Is the agent's bias level w greater than or equal to τ or less than or equal to τ?[1]

As noted earlier, by answering the above threshold question, one can also estimate the bias level w within accuracy ϵ by performing $\log(1/\epsilon)$ iterations of binary search. This effectively reduces the broader task of estimating w to a sequence of targeted threshold checks. By employing an adaptive signaling scheme, this approach lets us approximate w to any desired precision, providing an efficient solution to the bias estimation problem.

An algorithm Π for the above question terminates as soon as it can output a deterministic answer. The number of time steps for Π to terminate, denoted by $T_\tau(\Pi, w)$, is a random variable. The sample complexity of Π is defined to be the expected termination time in the worst case over $w \in [0, 1]$:

Definition 3.1 (sample complexity)**.** *The (worst-case) sample complexity of* Π *is defined as* $T_\tau(\Pi) = \max_{w\in[0,1]} \mathbb{E}[T_\tau(\Pi, w)]$.

Taking the worst case over $w \in [0, 1]$ in the above definition is not overly pessimistic. As we will show in the proofs, the worst case in fact happens at $w \in [\tau - \varepsilon, \tau + \varepsilon]$ for some $\varepsilon > 0$, which makes intuitive sense. Therefore, the sample complexity can be equivalently defined as $T_\tau(\Pi) = \max_{w\in[\tau-\varepsilon,\tau+\varepsilon]} \mathbb{E}[T_\tau(\Pi, w)]$.

Our goal is to design an algorithm Π that can determine whether $w \geq \tau$ or $w \leq \tau$ with minimal sample complexity. Specifically, we want to solve the minimax problem

$$\min_{\Pi} \max_{w\in[0,1]} \mathbb{E}[T_\tau(\Pi, w)].$$

We say an algorithm Π is *constant* if it keeps using the same signaling scheme repeatedly until termination. Constant algorithms are a special case of *non-adaptive* algorithms, which may vary the signaling schemes over time but are independent of historical data.

[1] One may want to test $w \geq \tau$ or $w < \tau$ instead. But this requires assumptions on tie-breaking when the agent has multiple optimal actions. Indifference at $w = \tau$ allows us to avoid such assumptions.

Preliminaries. We now introduce the well-known splitting lemma from the information design literature [AMS95, KG11]. It relates a signaling scheme with a set of induced true posteriors for a Bayesian agent and a distribution over the set of true posteriors.

Lemma 3.1 (Splitting Lemma, e.g., [KG11])**.** *Let π be a signaling scheme where each signal $s \in S$ is sent with unconditional probability $\pi(s) = \sum_{\theta\in\Theta} \mu_0(\theta)\pi(s|\theta)$ and induces true posterior μ_s. Then, the prior μ_0 equals the convex combination of $\{\mu_s\}_{s\in S}$ with weights $\{\pi(s)\}_{s\in S}$: $\mu_0 = \sum_{s\in S} \pi(s)\mu_s$. Conversely, if the prior can be expressed as a convex combination of distributions $\mu'_s \in \Delta(\Theta)$: $\mu_0 = \sum_{s\in S} p_s\mu'_s$, where $p_s \geq 0, \sum_{s\in S} p_s = 1$, then there exists a signaling scheme π where each signal s is sent with unconditional probability $\pi(s) = p_s$ and induces posterior μ'_s.*

The splitting lemma is also referred as the Bayesian consistency condition. It allows one to think about choosing a signaling scheme as choosing a set of true posteriors, $\{\mu_s\}_{s\in S}$, and a distribution over the set, $\{\pi(s)\}_{s\in S}$, in a Bayesian consistent way.

3.3 Warm-Up: A Two-State, Two-Action Example

How can the principal design a signaling scheme to learn the agent's bias level? We use a simple two-state, two-action example to demonstrate how inducing a specific true posterior belief will allow the principal to determine whether $w \geq \tau$ or $w \leq \tau$.

The two states of the world are represented as {Good, Bad}. The agent has two possible actions: Active and Passive. Taking the Passive action always yields a utility of 0, independently of the state. For the Active action, the utility is a if the state is Good and $-b$ otherwise; $a, b > 0$. We use the probability of the Good state to represent a belief, so the prior is a number $\mu_0 \in [0, 1]$, which is only a slight abuse of notation. With belief $\mu \in [0, 1]$ for the Good state (and $1 - \mu$ for the Bad state), the agent's expected utility for choosing the Active action is $a\mu - b(1 - \mu) = (a + b)\mu - b$. Thus, the Active action is better than the Passive action (so the agent will take Active) if

$$(a+b)\mu - b > 0 \qquad \iff \qquad \mu > \frac{b}{a+b} =: \mu^*. \tag{3.1}$$

Conversely, the Passive action is better if $\mu < \mu^*$. Here, $\mu^* = \frac{b}{a+b}$ is an *indifference belief* where the agent is indifferent between the two actions. We assume that the prior μ_0 satisfies $0 < \mu_0 < \mu^*$, so the agent chooses the Passive action by default.

Consider the following constant signaling scheme π_τ with two signals $\{G, B\}$:

- If the state is Good, send signal G with probability one.
- If the state is Bad, send signal B with probability $\frac{\mu^* - \mu_0}{(\mu^* - \tau\mu_0)(1-\mu_0)}$ and signal G with the complement probability.

We will show that, by repeatedly using π_τ, we can test whether the agent's bias w is $\le \tau$ or $\ge \tau$. By Bayes' Rule, the true posterior beliefs (for the Good state) associated with the two signals are $\mu_B = 0$ (i.e., on receiving B, the agent knows the state is Bad for sure) and

$$\mu_G = P(\text{Good}|G) = \frac{\mu_0 \cdot \pi_\tau(G|\text{Good})}{\mu_0 \cdot \pi_\tau(G|\text{Good}) + (1-\mu_0) \cdot \pi_\tau(G|\text{Bad})} = \frac{\mu^* - \tau\mu_0}{1-\tau}.$$

Notably, the posterior μ_G satisfies the following property: if the agent's bias level w is exactly equal to τ, then the agent's biased belief is equal to the indifference belief:

$$\text{when } w = \tau, \qquad \nu_G = \tau\mu_0 + (1-\tau)\mu_G = \mu^*.$$

We also note the inequality $\mu_0 < \mu^* < \mu_G$. As a result, if the agent's bias level w is $> \tau$, then the biased belief will be smaller than μ^*, and otherwise the opposite is true:

$$\text{for } w > \tau, \quad w\mu_0 + (1-w)\mu_G < \mu^*; \qquad \text{for } w < \tau, \quad w\mu_0 + (1-w)\mu_G > \mu^*.$$

By Equation (3.1), this means that the agent will take the Passive action if $w > \tau$, and the Active action if $w < \tau$ (on receiving G). Therefore, by observing which action is taken by the agent when signal G is sent, we can immediately tell whether $w \le \tau$ or $w \ge \tau$. This leads to the following:

Theorem 3.1. *In the two-state, two-action example, for any threshold* $\tau \in [0, \frac{1-\mu^*}{1-\mu_0}]$, *the above constant signaling scheme* π_τ *can test whether the agent's bias* w *satisfies* $w \le \tau$ *or* $w \ge \tau$*: specifically, whenever the signal* G *is sent,*

- *if the agent takes action* Active*, then* $w \le \tau$*,*
- *if the agent takes action* Passive*, then* $w \ge \tau$*.*

The sample complexity of this scheme is $\frac{\mu^*-\mu_0}{\mu_0(1-\tau)} + 1$*, which increases with* τ*.*

Proof. The range $\tau \in [0, \frac{1-\mu^*}{1-\mu_0}]$ ensures that the probability $\pi_\tau(B|\text{Bad}) = \frac{\mu^*-\mu_0}{(\mu^*-\tau\mu_0)(1-\mu_0)}$ is in $[0, 1]$. The two items in the theorem follow from the argument before the theorem statement. The sample complexity is equal to the expected number of time steps until a G signal is sent, which is a geometric random variable with success probability $P(G) = \mu_0\pi_\tau(G|\text{Good}) + (1-\mu_0)\pi_\tau(G|\text{Bad}) = \frac{\mu_0(1-\tau)}{\mu^*-\tau\mu_0}$. So the sample complexity is equal to the mean $\frac{1}{P(G)} = \frac{\mu^*-\mu_0}{\mu_0(1-\tau)} + 1$. □

The *main intuition* behind this result is that in order to test whether $w \ge \tau$ or $w \le \tau$, we design a signaling scheme where certain signals induce posteriors that make the agent *indifferent between two actions if the agent's bias level is exactly* τ. Then, the action actually taken by the agent will directly reveal whether $w \ge \tau$ or $w \le \tau$. Such signals are *useful* signals, but not all signals are necessarily useful. The sample complexity is then determined by the total probability of useful signals. This intuition will carry over to computing the optimal signaling scheme for the general case in Section 3.4.

Finally, we remark that using the constant signaling scheme π_τ constructed above to test $w \ge \tau$ or $w \le \tau$ is in fact the optimal adaptive algorithm, according to the results we will present in Section 3.4. So, the minimal sample complexity to test whether $w \geq \tau$ or $w \leq \tau$ in this two-state, two-action example is exactly $\frac{\mu^* - \mu_0}{\mu_0(1-\tau)} + 1$ as shown in Theorem 3.1.

3.4 General Case: Computing the Optimal Signaling Scheme

In this section, we generalize the initial observations from the previous section to the case with any number of actions and states and general utility function U. We will show how to compute the optimal algorithm (signaling scheme) to test the agent's bias level. There are three key ingredients. First, we prove that we can use a constant signaling scheme. Second, we develop a "revelation principle" to further simplify the space of signaling schemes. Building on these two steps, we show that the optimal signaling scheme can be computed by a linear program.

3.4.1 Optimality of Constant Signaling Schemes

This subsection shows that adaptive algorithms are no better than constant algorithms for the problem of testing $w \geq \tau$ or $w \leq \tau$. Therefore, to find the algorithm with minimal sample complexity, we only need to consider constant algorithms/signaling schemes.

Lemma 3.2. *Fix $\tau \in (0, 1)$. For the problem of testing whether $w \geq \tau$ or $w \leq \tau$, the sample complexity of any adaptive algorithm is at least that of the optimal constant algorithm (i.e., using a fixed signaling scheme repeatedly).*

To prove this lemma, we introduce some notations. For any action $a \in A \setminus \{a_0\}$, define vector

$$c_a = (c_{a,\theta})_{\theta \in \Theta} = \big(U(a_0, \theta) - U(a, \theta)\big)_{\theta \in \Theta} \in \mathbb{R}^{|\Theta|}, \tag{3.2}$$

whose components are the utility differences between the default action a_0 and any other action a at different states $\theta \in \Theta$. Let $R_{a_0} \subseteq \Delta(\Theta)$ be the region of beliefs under which the agent strictly prefers a_0 over any other action:

$$R_{a_0} = \big\{\mu \in \Delta(\Theta) \mid \forall a \in A \setminus \{a_0\},\, c_a^\top \mu > 0\big\}.$$

It is the intersection of $|A|-1$ open halfspaces with the probability simplex $\Delta(\Theta)$. As the agent strictly prefers a_0 at the prior μ_0, we have $\mu_0 \in R_{a_0}$. The boundary of this region, ∂R_{a_0}, is the set of beliefs where the agent is indifferent between a_0 and at least one other action $a \in A \setminus \{a_0\}$ and a_0 and a are both (weakly) better than any other action:

$$\partial R_{a_0} = \big\{\mu \in \Delta(\Theta) \mid \exists a \in A \setminus \{a_0\}, c_a^\top \mu = 0 \text{ and } \forall a' \in A \setminus \{a_0\},\, c_{a'}^\top \mu \geq 0\big\}. \tag{3.3}$$

Lastly, the exterior of R_{a_0}, denoted as $\mathrm{ext}R_{a_0}$, comprises the set of beliefs where the agent strictly prefers not to choose a_0:

$$\mathrm{ext}R_{a_0} = \Delta(\Theta) \setminus (R_{a_0} \cup \partial R_{a_0}) = \big\{\mu \in \Delta(\Theta) \mid \exists a \in A \setminus \{a_0\},\, c_a^\top \mu < 0\big\}.$$

Given a signaling scheme π, we classify its signals into three types based on the location of the biased belief associated with the signal with respect to the region R_{a_0}.

Definition 3.2. *Let $\tau \in (0,1)$ be a parameter. Let $s \in S$ be a signal from a signaling scheme π, with associated true posterior μ_s and τ-biased posterior $\mu_s^\tau = \tau\mu_0 + (1-\tau)\mu_s$. We say s is*

- *an **internal signal** if $\mu_s^\tau \in R_{a_0}$;*
- *a **boundary signal** if $\mu_s^\tau \in \partial R_{a_0}$;*
- *an **external signal** if $\mu_s^\tau \in \mathrm{ext}R_{a_0}$.*

The above classification helps to formalize the idea of whether a signal is "useful" for bias detection. A boundary signal is useful because the action taken by the agent after receiving a boundary signal immediately tells whether $w \geq \tau$ or $w \leq \tau$:

Lemma 3.3. *When a boundary signal is realized, the agent's action immediately reveals whether $w \geq \tau$ or $w \leq \tau$. Specifically, if the agent chooses action a_0, then $w \geq \tau$; otherwise, $w \leq \tau$.*

Proof. If the agent's bias level satisfies $w < \tau$, then the biased belief $\nu_s = w\mu_0 + (1-w)\mu_s$ must be inside R_{a_0} (because $\mu_s^\tau = \tau\mu_0 + (1-\tau)\mu_s$ is on the boundary of R_{a_0} and $\mu_0 \in R_{a_0}$), so the agent strictly prefers the default action a_0. If $w > \tau$, then the biased belief ν_s is outside of R_{a_0}, so the agent will not take action a_0. □

An external signal might also be useful in revealing whether $w \geq \tau$ or $w \leq \tau$ if the agent is indifferent between some actions a_1, a_2 other than a_0 at the τ-biased belief μ_s^τ. However, the following lemma shows that, in such cases, we can always modify the signaling scheme to turn the external signal into a boundary signal. This modification will increase the total probability of useful signals and hence reduce the sample complexity. The proof of this lemma is in Section 3.7.1.

Lemma 3.4. *Suppose Π is an adaptive algorithm that uses signaling schemes with internal, boundary, and external signals. Then, there exists another adaptive algorithm Π' with equal or lower sample complexity that employs only signaling schemes with internal and boundary signals.*

An internal signal, on the other hand, is not useful for testing $w \geq \tau$ or $w \leq \tau$, for the following reason. For an internal signal, the biased belief with bias level τ, μ_s^τ, lies inside R_{a_0}. Since R_{a_0} is an open region, there must exist a small number $\varepsilon > 0$ such that when the agent has bias level $w = \tau + \varepsilon$ or $\tau - \varepsilon$, the biased belief with bias level w, $w\mu_0 + (1-w)\mu_s$, is also inside the region R_{a_0}, so the agent will take action a_0. As the agent takes a_0 under both $w = \tau + \varepsilon$ and $\tau - \varepsilon$, we cannot distinguish these two cases, so this signal is not helpful in determining $w \geq \tau$ or $w \leq \tau$. The following lemma formalizes the idea that internal signals are not useful (with proof in Section 3.7.2):

Lemma 3.5. *To test whether $w \geq \tau$ or $w \leq \tau$, any adaptive algorithm that uses signaling schemes with boundary and internal signals cannot terminate until a boundary signal is sent.*

Proof of Lemma 3.2. By Lemma 3.4, the optimal adaptive algorithm only uses signaling schemes with boundary and internal signals. By Lemma 3.5, the algorithm cannot terminate until a boundary signal is sent. By Lemma 3.3, the algorithm terminates when a boundary signal is sent. We conclude that the termination time of any adaptive algorithm cannot be better than the constant algorithm that keeps using the signaling scheme that maximizes the total probability of boundary signals. □

3.4.2 Revelation Principle

To compute the optimal constant signaling scheme, we need another technique that is similar to the *revelation principle* in the information design literature [KG11, DX16]. The revelation principle says that, in some information design problems, it is without loss of generality to consider only "direct" signaling schemes where signals are recommendations of actions for the agent: namely, the signal space $S = A$, and when the principal sends signal a, it should be optimal for the agent to take action a given the posterior belief induced by signal a. Unlike classical information design problems where the agent is unbiased, our problem involves a biased agent, so we need a different revelation principle: the signals are still action recommendations, but when the principal sends signal a, action a is optimal for an agent with bias level exactly τ; moreover, if $a \neq a_0$, then an agent with bias level τ will be indifferent between a and a_0. This insight is formalized in the following lemma:

Lemma 3.6 (revelation principle for bias detection)**.** *Let π be an arbitrary signaling scheme that can test $w \geq \tau$ or $w \leq \tau$. Then, there exists another signaling scheme π' that can do so with signal space $S = A$ such that:*

(1) Given signal $a \in A$, action a is an optimal action for any agent with bias level $w = \tau$.

(2) Given signal $a \in A \setminus \{a_0\}$, actions a and a_0 are both optimal for any agent with bias level $w = \tau$. As a corollary, if the agent's bias level $w < \tau$, then the agent strictly prefers a over a_0; and if $w > \tau$, then the agent strictly prefers a_0 over any other actions.

(3) The sample complexity satisfies $T_\tau(\pi') \leq T_\tau(\pi)$.

In the above signaling scheme π', every $a \in A \backslash \{a_0\}$ is a boundary signal (Definition 3.2), which is useful for testing bias: given signal $a \in A \setminus \{a_0\}$, if the agent takes action a_0, then it must be $w \geq \tau$; otherwise $w \leq \tau$. The signal a_0 is internal and not useful for determining $w \geq \tau$ or $w \leq \tau$. So, the sample complexity of π' is equal to the expected time steps until a signal in $A \setminus \{a_0\}$ is sent.

The idea behind Lemma 3.6 is *combination of signals*. Suppose there is a signaling scheme that can determine whether $w \geq \tau$ or $w \leq \tau$ with a signal space larger than A. There must exist two signals s and s' under which the agent is indifferent between a_0 and some action $a \neq a_0$ if the agent's bias level is exactly τ. We can then combine the two signals into a single signal s'' under which the agent remains indifferent between a_0 and a, yielding a new signaling scheme with a smaller signal space. Repeating this can reduce the signal space to size $|A|$. See Section 3.7.3 for the full proof.

3.4.3 Algorithm for Computing the Optimal Signaling Scheme

Finally, we present an algorithm to compute the optimal (minimal sample complexity) signaling scheme to test whether $w \geq \tau$ or $w \leq \tau$. The revelation principle in the previous subsection ensures that we only need a direct signaling scheme where signals are action recommendations. The optimal direct signaling scheme turns out to be solvable by a linear program, detailed in Algorithm 3.1. In the linear program, the constraint in Equation (3.5) ensures that whenever the principal recommends action $a \in A$, it is optimal for an agent with bias level τ to take action a; this satisfies condition (1) in the revelation principle (Lemma 3.6). The indifference constraint (Equation (3.5)) ensures that when the recommended action a is not a_0, an agent with bias level τ is indifferent between a and a_0; this satisfies condition (2) in the revelation principle. The objective (Equation (3.4)) is to maximize the probability of useful signals (those in $A \setminus \{a_0\}$), hence minimize the sample complexity.

Algorithm 3.1: Linear program to compute the optimal signaling scheme

```
Input    : prior $\mu_0$, utility function $U$, and the parameter $\tau \in (0, 1)$
Variable: signaling scheme $\pi$, consisting of $\pi(a|\theta)$ for $a \in A, \theta \in \Theta$
Denote $\Delta U(a, a', \theta) = U(a, \theta) - U(a', \theta)$. Solve the following linear program:
    $$\text{Maximize}_\pi \quad \sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \pi(a|\theta)\mu_0(\theta) \qquad (3.4)$$
subject to:
    Optimality of $a$ over other actions: $\forall a \in A, \forall a' \in A \setminus \{a\}$
        $$\sum_{\theta \in \Theta} \pi(a|\theta) \cdot \mu_0(\theta) \Big[(1-\tau)\Delta U(a, a', \theta) + \tau \sum_{\theta' \in \Theta} \mu_0(\theta') \Delta U(a, a', \theta')\Big] \geq 0; \qquad (3.5)$$
    Indifference between $a$ and $a_0$: $\forall a \in A \setminus \{a_0\}$,
        $$\sum_{\theta \in \Theta} \pi(a|\theta) \cdot \mu_0(\theta) \Big[(1-\tau)\Delta U(a, a_0, \theta) + \tau \sum_{\theta' \in \Theta} \mu_0(\theta') \Delta U(a, a_0, \theta')\Big] = 0; \qquad (3.6)$$
    Probability distribution constraints: $\forall \theta \in \Theta$,
        $$\sum_{a \in A} \pi(a|\theta) = 1 \quad \text{and} \quad \forall a \in A, \ \pi(a|\theta) \geq 0.$$
```

Theorem 3.2. *Algorithm 3.1 finds a constant signaling scheme for testing $w \geq \tau$ or $\leq \tau$ that is optimal among all adaptive signaling schemes. The sample complexity of* *the optimal signaling scheme is* $1/p^*$, *where* p^* *is the optimal objective value of the linear program* (3.4).

Using the above optimal signaling scheme, whenever the principal recommends an action a other than a_0, the agent's action immediately reveals whether $w \geq \tau$ or $w \leq \tau$: if the agent indeed follows the recommendation or takes any other action than a_0, then the bias must be small ($w \leq \tau$); if the agent takes a_0 instead, the bias must be large ($w \geq \tau$). Thus, the expected sample complexity is equal to the expected number of iterations until a signal in $A \setminus \{a_0\}$ is sent, which is $1/p^*$.[2]

The linear program in Algorithm 3.1 has a polynomial size in $|A|$ (the number of actions) and $|\Theta|$ (the number of states), so it is a polynomial-time algorithm. The solution p^* depends on the geometry of the problem instance and does not seem to have a closed-form expression.

The rest of this section proves Theorem 3.2. The proof requires another lemma:

Lemma 3.7. *Given a signaling scheme* $\pi = (\pi(a|\theta))_{a\in A, \theta\in\Theta}$ *and an agent's bias level* w, *after signal* a *is sent, the agent strictly prefers action* a_1 *over* a_2 *under the biased belief if and only if:*

$$\sum_{\theta\in\Theta} \pi(a|\theta) \cdot \mu_0(\theta) \Big[(1-w)\Delta U(a_1, a_2, \theta) + w \sum_{\theta'\in\Theta} \mu_0(\theta')\Delta U(a_1, a_2, \theta') \Big] > 0.$$

Proof. The agent's biased belief under signal a and bias level w is given by $(1 - w)\frac{\mu_0(\theta)\pi(a|\theta)}{\sum_{\theta'\in\Theta}\mu_0(\theta')\pi(a|\theta')} + w\mu_0(\theta), \ \forall \theta \in \Theta$. The condition for the agent to strictly prefer a_1 over a_2 is that the expected utility under the biased belief when choosing a_1 is greater

[2] We can also derive a high-probability guarantee: with $t \geq \frac{1}{p^*}\log\frac{1}{\delta}$ iterations, we can determine whether $w \geq \tau$ or $w \leq \tau$ with probability at least $1 - \delta$. This is because the probability that no useful signal is sent after t iterations is at most $(1-p^*)^t \leq \delta$ when $t \geq \frac{1}{p^*}\log\frac{1}{\delta}$.

than that of a_2:

$$\sum_{\theta\in\Theta}\left((1-w)\frac{\mu_0(\theta)\pi(a|\theta)}{\sum_{\theta'\in\Theta}\mu_0(\theta')\pi(a|\theta')}+w\mu_0(\theta)\right)\Delta U(a_1,a_2,\theta)>0,$$

where $\Delta U(a_1,a_2,\theta)=U(a_1,\theta)-U(a_2,\theta)$. Multiplying by $\sum_{\theta'\in\Theta}\mu_0(\theta')\pi(a|\theta')$, we obtain:

$$(1-w)\sum_{\theta\in\Theta}\mu_0(\theta)\pi(a|\theta)\Delta U(a_1,a_2,\theta)+w\sum_{\theta\in\Theta}\mu_0(\theta)\sum_{\theta'\in\Theta}\mu_0(\theta')\pi(a|\theta')\Delta U(a_1,a_2,\theta)>0.$$

Factoring out the terms, this can be rewritten as:

$$\sum_{\theta\in\Theta}\pi(a|\theta)\mu_0(\theta)\left((1-w)\Delta U(a_1,a_2,\theta)+w\sum_{\theta'\in\Theta}\mu_0(\theta')\Delta U(a_1,a_2,\theta')\right)>0.$$

This final expression is positive if and only if the agent to strictly prefer a_1 over a_2. □

Proof of Theorem 3.2. According to Lemma 3.2 (constant algorithms are optimal) and Lemma 3.6 (revelation principle), to find an optimal adaptive algorithm we only need to find the optimal constant signaling scheme that satisfies the conditions in Lemma 3.6. We verify that the signaling scheme computed from the linear program in Algorithm 3.1 satisfies the conditions in Lemma 3.6:

- The optimality constraint (Equation (3.5)) in the linear program, together with Lemma 3.7, ensures that: whenever signal $a\in A$ is sent, action a is weakly better than any other action for an agent with bias level $w=\tau$. This satisfies the first condition in Lemma 3.6.
- The indifference constraint (Equation (3.5)), together with Lemma 3.7, ensures that: whenever $a\in A\setminus\{a_0\}$ is sent, the agent is indifferent between action a and a_0 if the bias level $w=\tau$. Then, by the optimality constraint (Equation (3.5)), we have both

a and a_0 being optimal actions. This satisfies the second condition in Lemma 3.6.

We then argue that the solution of the linear program is the optimal signaling scheme that satisfies the conditions of Lemma 3.6. According to our argument after Lemma 3.6, only the signals in $A \setminus \{a_0\}$ are useful signals, so the sample complexity is equal to the expected number of time steps until a signal in $A \setminus \{a_0\}$ is sent. The probability that a signal in $A \setminus \{a_0\}$ is sent at each time step is

$$\sum_{a \in A \setminus \{a_0\}} \pi(a) = \sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \mu_0(\theta) \pi(a|\theta).$$

The expected number of time steps is the inverse $\frac{1}{\sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \mu_0(\theta)\pi(a|\theta)}$ (because the number of time steps is a geometric random variable). The linear program maximizes the probability $\sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \mu_0(\theta)\pi(a|\theta)$, so it minimizes the sample complexity. □

3.5 Geometric Characterization

To complement the algorithmic solution presented in the previous section, this section provides a geometric characterization of the bias detection problem. We identify the conditions under which testing whether $w \geq \tau$ or $w \leq \tau$ can be done in only *one* sample, in finite number of samples, or cannot be done at all (which is the scenario where the linear program in Algorithm 3.1 is infeasible).

By Assumption 3.1 (a_0 is strictly better than other actions at prior μ_0), we have:

$$c_a^\top \mu_0 = \sum_{\theta \in \Theta} \mu_0(\theta) \big(U(a_0, \theta) - U(a, \theta) \big) > 0, \quad \forall a \in A \setminus \{a_0\},$$

where c_a is as defined in Equation (3.2). Define I_a as the set of indifference beliefs between action a and a_0, which is the intersection of the hyperplane $\{x \mid c_a^\top x = 0\}$ and the probability simplex $\Delta(\Theta)$:

$$I_a := \{\mu \in \Delta(\Theta) \mid c_a^\top \mu = 0\}.$$

Given a parameter $\tau \in (0, 1)$, for which we want to test whether $w \geq \tau$ or $w \leq \tau$, let

$$I_{a,\tau} := \{\mu \in \Delta(\Theta) \mid (1-\tau)\mu + \tau\mu_0 \in I_a\}$$

be the set of posterior beliefs for which, if the agent's bias level is exactly τ, then the agent's biased belief will fall within the indifference set I_a.

Lemma 3.8. *$I_{a,\tau}$ is equal to the intersection of the probability simplex $\Delta(\Theta)$ and a translation of the hyperplane $\{x \mid c_a^\top x = 0\}$: $I_{a,\tau} = \big\{\mu \in \Delta(\Theta) \mid c_a^\top \mu = -\frac{\tau}{1-\tau} c_a^\top \mu_0\big\}$.*

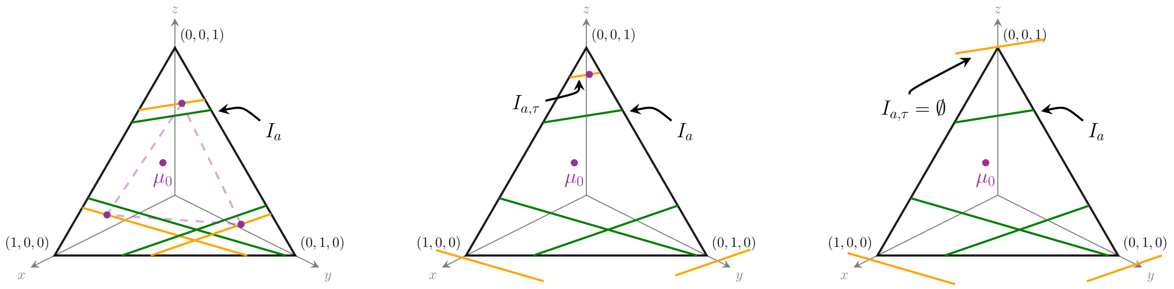
Proof. For $\mu \in \Delta(\Theta)$, by convexity of $\Delta(\Theta)$, we have $(1-\tau)\mu + \tau\mu_0 \in \Delta(\Theta)$. Then,

$$\begin{aligned} \mu \in I_{a,\tau} &\iff (1-\tau)\mu + \tau\mu_0 \in I_a \iff c_a^\top((1-\tau)\mu + \tau\mu_0) = 0 \\ &\iff (1-\tau)c_a^\top \mu + \tau c_a^\top \mu_0 = 0 \iff c_a^\top \mu = -\frac{\tau}{1-\tau} c_a^\top \mu_0. \end{aligned}$$ □

With this representation of $I_{a,\tau}$ in hand, we can now present a geometric characterization of the testability of bias.

Theorem 3.3 (geometric characterization). *Fix $\tau \in (0, 1)$. The problem of testing $w \geq \tau$ or $w \leq \tau$*

- *Can be solved with a* single sample *(the sample complexity is 1) if and only if the prior μ_0 is in the convex hull formed by the translated sets $I_{a,\tau}$ for all non-default actions $a \in A \setminus \{a_0\}$: i.e., $\mu_0 \in \mathrm{ConvexHull}\big(\bigcup_{a \in A \setminus \{a_0\}} I_{a,\tau}\big)$.*
- *Can be solved (with finite sample complexity) if and only if $I_{a,\tau} \neq \emptyset$ for at least one $a \in A \setminus \{a_0\}$.*



(a) A single sample

(b) Finite sample complexity

(c) Cannot be solved

Figure 3.1: Geometric Characterization of Bias Detection

- *Cannot be solved if* $I_{a,\tau} = \emptyset$ *for all* $a \in A \setminus \{a_0\}$.

Figure 3.1 illustrates the three cases of Theorem 3.3. The triangle is the probability simplex over three states, where μ_0 is the prior belief. Each point in the simplex is an unbiased belief, corresponding to an optimal action for the agent. Each green curve indicates the indifference beliefs between the default action a_0 and another action a, namely, the I_a. Orange curves ($I_{a,\tau}$) are translated versions of these indifference curves; a posterior on these curves means that the agent's biased belief (at bias level τ) aligns with the green curves. From (a) to (c), τ increases, translating the orange curves further.

- Figure (a) corresponds to case 1 in Theorem 3.3, where μ_0 can be represented as a convex combination of points on the translated curves, allowing bias level detection with a single sample. In this case, the solution of the linear program in Algorithm 3.1 satisfies $\sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \pi(a|\theta)\mu_0(\theta) = 1$, meaning that useful signals are sent with probability 1, which allows us to tell whether $w \geq \tau$ or $w \leq \tau$ immediately.

- Figure (b) corresponds to case 2 in Theorem 3.3. In this case, the total probability of useful signals satisfies $\sum_{a \in A \setminus \{a_0\}} \sum_{\theta \in \Theta} \pi(a|\theta)\mu_0(\theta) < 1$, so the sample complexity is more than 1.

- Figure (c) corresponds to case 3 in Theorem 3.3, where the bias level cannot be tested against τ with finitely many samples. In this case, the linear program in Algorithm 3.1 is not feasible, so $w \geq \tau$ or $w \leq \tau$ cannot be determined; importantly, this is not a limitation of our particular algorithm, but is a general impossibility in our model.

The rest of this section proves Theorem Theorem 3.3.

3.5.1 Proof of Theorem 3.3

We first prove the first part of Theorem 3.3, then prove the the second and third parts.

Proof of Part 1 of Theorem 3.3

We want to prove that $w \ge \tau$ or $w \le \tau$ can be tested with a single sample **if and only if** the prior μ_0 is in the convex hull formed by the translated sets $I_{a,\tau}$ for all non-default actions $a \in A \setminus \{a_0\}$: $\mu_0 \in \mathrm{ConvexHull}\big(\cup_{a \in A\setminus\{a_0\}} I_{a,\tau} \big)$.

The "if" part. Suppose $\mu_0 \in \mathrm{ConvexHull}\big(\cup_{a \in A\setminus\{a_0\}} I_{a,\tau} \big)$, namely, there exist a set of positive weights $\{p_s\}_{s\in S}$ and a set of posterior beliefs $\{\mu_s\}_{s\in S}$ such that

$$\mu_0 = \sum_{s\in S} p_s \mu_s,$$

where each $\mu_s \in I_{a,\tau}$ for some $a \in A \setminus \{a_0\}$. By definition, the τ-biased belief $\tau\mu_0 + (1-\tau)\mu_s$ is in the indifference set I_a. Recall the definition of the boundary set ∂R_{a_0} (Equation (3.3)), which is the set of beliefs under which the agent is indifferent between a_0 and some other action and these two actions are better any other actions. The τ-biased belief $\tau\mu_0 + (1-\tau)\mu_s \in I_a$ may or may not belong to ∂R_{a_0}, depending on whether a and a_0 are better than any other actions:

- If $\tau\mu_0 + (1-\tau)\mu_s \in \partial R_{a_0}$, then s is a boundary signal (by Definition 3.2) and hence useful for testing whether $w \ge \tau$ or $w \le \tau$ (Lemma 3.3). Denote $\mu'_s = \mu_s$ in this case.

- If $\tau\mu_0 + (1-\tau)\mu_s \notin \partial R_{a_0}$, then there must exist some action a' that is strictly better than a and a_0 for the agent at the τ-biased belief, hence $\tau\mu_0 + (1-\tau)\mu_s \in \text{ext}R_{a_0}$ (so s is an external signal). Then, according to the argument in Lemma 3.4, we can find another belief μ_s' on the line segment between μ_s and μ_0 such that the τ-biased version of μ_s' lies exactly on the boundary set ∂R_{a_0}:

$$\tau\mu_0 + (1-\tau)\mu_s' \in \partial R_{a_0}, \qquad \mu_s' = t\mu_s + (1-t)\mu_0 \text{ for some } t \in [0,1].$$

After the above discussion, we have found a μ_s' that is either equal to μ_s or on the line segment between μ_s and μ_0, for every $s \in S$. So, μ_0 can be written as a convex combination of $\{\mu_s'\}_{s\in S}$:

$$\mu_0 = \sum_{s\in S} p_s' \mu_s'.$$

Moreover, the μ_s' defined above satisfies $\tau\mu_0 + (1-\tau)\mu_s' \in \partial R_{a_0}$. So, a signal inducing true posterior μ_s' will be a boundary signal and useful for testing $w \geq \tau$ or $w \leq \tau$ (Lemma 3.3). Finally, by the splitting lemma (Lemma 3.1), we know that there must exist a signaling scheme π' with signal space S where each signal $s \in S$ indeed induces posterior μ_s'. Such a signaling scheme sends useful (boundary) signals with probability 1. Hence, the sample complexity of it is 1.

The "only if" part. Suppose whether $w \geq \tau$ or $w \leq \tau$ can be tested with a single sample. This means that the optimal signaling scheme obtained from the linear program in Algorithm 3.1 must satisfy $\sum_{a\in A\setminus\{a_0\}} \pi(a) = \sum_{a\in A\setminus\{a_0\}} \sum_{\theta\in\Theta} \pi(a|\theta)\mu_0(\theta) = 1$, namely, the total probability of useful signals (signals in $A \setminus \{a_0\}$) is 1. Then, by the splitting lemma, the prior μ_0 can be expressed as the convex combination

$$\mu_0 = \sum_{a\in A\setminus\{a_0\}} \pi(a)\mu_a$$

where $\pi(a) = \sum_{\theta \in \Theta} \mu_0(\theta)\pi(a|\theta)$ is the unconditional probability of signal a and μ_a is the true posterior induced by signal a. Moreover, the indifference constraint (3.6) in the linear program ensures that the agent is indifferent between a and a_0 upon receiving signal a if the agent has bias level τ: mathematically, $\tau\mu_0 + (1-\tau)\mu_a \in I_a$. This means $\mu_a \in I_{a,\tau}$ by definition. So, we obtain

$$\mu_0 \in \text{ConvexHull}\Bigg(\bigcup_{a \in A \setminus \{a_0\}} I_{a,\tau} \Bigg).$$

Proof of Parts 2 and 3 of Theorem 3.3

We first prove that, if whether $w \geq \tau$ or $w \leq \tau$ can be tested with finite sample complexity, then $I_{a,\tau} \neq \emptyset$ for at least one $a \in A \setminus \{a_0\}$.

According to Lemma 3.2, if we can test whether $w \geq \tau$ or $w \leq \tau$ with finite sample complexity using adaptive algorithms, then we can do this using a constant signaling scheme. Lemma 3.4 further ensures that we can do this using a constant signaling scheme π with only boundary and internal signals. But according to Lemma 3.5, internal signals are not useful for testing $w \geq \tau$ or $w \leq \tau$. So, the signaling scheme π must send some boundary signal s with positive probability. Let μ_s be the true posterior induced by s. By the definition of boundary signal, $\tau\mu_0 + (1-\tau)\mu_s \in \partial R_{a_0}$, implying that the agent is indifferent between a_0 and some action $a \in A \setminus \{a_0\}$ if their belief is $\tau\mu_0 + (1-\tau)\mu_s$ (and a_0 and a are better than any other actions). This means $\tau\mu_0 + (1-\tau)\mu_s \in I_a$, so $\mu_s \in I_{a,\tau}$ by definition. Hence, $I_{a,\tau} \neq \emptyset$.

We then prove the opposite direction: if $I_{a,\tau} \neq \emptyset$ for at least one $a \in A \setminus \{a_0\}$, then whether $w \geq \tau$ or $w \leq \tau$ can be tested with finite sample complexity.

Let $a_1 \in A \setminus \{a_0\}$ be an action for which $I_{a_1,\tau} \neq \emptyset$. We claim that:

Claim 3.1. *There exists a state $\theta_1 \in \Theta$ for which the agent weakly prefers action a_1 over action a_0 if the true posterior is state θ_1 with probability 1 and the agent has bias level τ.*

In notation, let $e_{\theta_1} \in \Delta(\Theta)$ be the vector whose θ_1th component is 1 *and other components are* 0*. The agent weakly prefers action a_1 over action a_0 under belief $\tau\mu_0 + (1-\tau)e_{\theta_1}$.*

Proof. Suppose on the contrary that no such state θ_1 exists. Then the agent strictly prefers a_0 over a_1 under belief $\tau\mu_0 + (1-\tau)e_\theta$ for every state $\theta \in \Theta$. This implies that, for any belief $\mu \in \Delta(\Theta)$, the agent should also strictly prefer a_0 over a_1 under the belief $\tau\mu_0 + (1-\tau)\mu$, due to linearity of the agent's utility with respect to the belief. The agent strictly preferring a_0 over a_1 implies $\tau\mu_0 + (1-\tau)\mu \notin I_a$, so μ cannot be in $I_{a,\tau}$ by definition. This holds for any $\mu \in \Delta(\Theta)$, so $I_{a,\tau} = \emptyset$, a contradiction. □

Let θ_1 be the state in the above claim. The prior μ_0 can be trivially written as the convex combination of e_{θ_1} and e_θ for other states θ:

$$\mu_0 = \mu_0(\theta_1)e_{\theta_1} + \sum_{\theta \in \Theta \setminus \{\theta_1\}} \mu_0(\theta)e_\theta.$$

Since the agent does not prefer a_0 under belief $\tau\mu_0 + (1-\tau)e_{\theta_1}$, the belief $\tau\mu_0 + (1-\tau)e_{\theta_1}$ cannot be in the region R_{a_0}. The prior μ_0 is in the region R_{a_0}. Consider the line segment connecting e_{θ_1} and the prior μ_0. There must exist a point $\mu' = te_{\theta_1} + (1-t)\mu_0$ on the line segment such that the τ-biased belief $\tau\mu_0 + (1-\tau)\mu'$ lies exactly on the boundary of R_{a_0}. Clearly, the prior can also be written as a convex combination of μ' and e_θ for $\theta \in \Theta \setminus \{\theta_1\}$:

$$\mu_0 = p'\mu' + \sum_{\theta \in \Theta \setminus \{\theta_1\}} p'_\theta e_\theta.$$

Then by the splitting lemma (Lemma 3.1), there exists a signaling scheme with $|\Theta|$ signals where one signal induces posterior μ' and the other signals induce posteriors $\{e_\theta\}_{\theta \in \Theta \setminus \{\theta_1\}}$. In particular, the signal inducing μ' is a boundary signal since $\tau\mu_0 + (1-\tau)\mu' \in \partial R_{a_0}$ by construction. By Lemma 3.3, that signal is useful for testing $w \geq \tau$ or $w \leq \tau$. When that signal is sent (which happens with positive probability $p' > 0$ at each time step), we can tell $w \geq \tau$ or $w \leq \tau$. This finishes the proof.

The two directions proved above together prove the parts 2 and 3 of Theorem 3.3.

3.6 Discussion

Our approach has some limitations; here we discuss the two that we view as most significant.

First, we have assumed a linear model of bias. While the linear model is common in the literature [ENS10, HL17, dCZ22, TH21], we also consider a more general model of bias in our published paper [CLP+24]: as the bias level w increases from 0 to 1, the agent's belief changes from the true posterior μ_s to the prior μ_0 according to some general continuous function $\phi(\mu_0, \mu_s, w)$. We show that, as long as the function ϕ satisfies a certain single-crossing property (as w increases, once the agent starts to prefer the default action a_0, they will not change the preferred action anymore), our results regarding the optimality of constant signaling schemes and the geometric characterization still hold, while the revelation principle and the linear program algorithm no longer work because ϕ is not linear. We consider it an interesting challenge to come up with more general models of bias that are still tractable, in the sense that one can efficiently design good signaling schemes with reasonable sample complexity bounds.

Second, we have assumed that the agent's prior is the same as the real prior from which states of the world are drawn. But what if the agent's prior is different? Our results directly extend to the case where the agent has a wrong, *known* prior. If the agent's prior is unknown, then our problem becomes significantly more challenging. More generally, the agent may have a private type that determines both their prior and utility and is unknown to the principal. We conjecture that testing the agent's bias in this case becomes impossible, because if different types consistently take "opposite" actions, then the actions provide no information about the agent's bias.

Despite these limitations, we view our paper as making significant progress on a novel problem that seems fundamental. Our results suggest that practical algorithms for detecting bias in belief update are within reach and, in the long term, may lead to new insights on issues of societal importance. In particular, we anticipate future research in more complex situations such as combining decisions of many experts (human or AI) after measuring and accounting for their individual biases.

3.7 Omitted Proofs in this Chapter

3.7.1 Proof of Lemma 3.4

Proof. Suppose that, during its operation, Π selects a signaling scheme π that includes an external signal $s \in S$. By definition, for an external signal, the τ-biased belief $\mu_s^\tau = \tau\mu_0 + (1-\tau)\mu_s$ is in $\text{ext}R_{a_0}$. This implies that the true posterior μ_s, derived from the signaling scheme π and the prior μ_0, also lies in $\text{ext}R_{a_0}$. Consequently, the line segment connecting μ_s and μ_0, represented as $\{(1-t)\mu_s + t\mu_0 \mid t \in [0,1]\}$, must intersect the boundary ∂R_{a_0} at some point. Denote this intersection by $\mu^* = (1-t^*)\mu_s + t^*\mu_0 \in \partial R_{a_0}$.

We will adjust the original signaling scheme π. To do so, define $\tilde{\mu}_s$ as the belief whose τ-biased version equals μ^*:

$$\tau\mu_0 + (1-\tau)\tilde{\mu}_s = \mu^* \quad \iff \quad \tilde{\mu}_s = \frac{(t^*-\tau)\mu_0 + (1-t^*)\mu_s}{1-\tau}.$$

Under the original signaling scheme π, according to the splitting lemma (Lemma 3.1), the prior μ_0 can be represented as a convex combination of μ_s and the posteriors associated with other signals $s' \in S \setminus \{s\}$:

$$\mu_0 = p_s\mu_s + \sum_{s' \in S \setminus \{s\}} p_{s'}\mu_{s'}.$$

If we change μ_s to $\tilde{\mu}_s$, then we obtain a new convex combination (this is valid because $\tilde{\mu}_s$ is on the line segment from μ_s to μ_0):

$$\mu_0 = \tilde{p}_s \tilde{\mu}_s + \sum_{s' \in S \setminus \{s\}} \tilde{p}_{s'} \mu_{s'},$$

where

$$\tilde{p}_s = \frac{p_s}{1 - t^* + t^* p_s} \quad \text{and} \quad \forall s' \in S \setminus \{s\},\ \tilde{p}_{s'} = \frac{1 - t^*}{1 - t^* + t^* p_s} p_{s'}.$$

Then, by the splitting lemma (Lemma 3.1), there exists a signaling scheme π' with $|S|$ signals where signal s induces posterior $\tilde{\mu}_s$ and other signals s' induces $\mu_{s'}$. Note that the τ-biased version of $\tilde{\mu}_s$ satisfies $\tau \mu_0 + (1 - \tau)\tilde{\mu}_s = \mu^* \in \partial R_{a_0}$, so s is a boundary signal under signaling scheme π'.

Since s is a boundary signal, we can immediately tell whether $w \geq \tau$ or $w \leq \tau$ according to Lemma 3.3 when s is sent and end the algorithm. If any signal s' other than s is sent, the induced posterior μ_s is the same as the posterior in the original signaling scheme π, so the agent will take the same action, and we can just follow the rest of the original algorithm Π. But we note that the probability of signal s being sent under the new signaling scheme π' is larger than or equal to the probability under the original signaling scheme π:

$$\tilde{p}_s = \frac{p_s}{1 - t^* + t^* p_s} \geq p_s.$$

So, in expectation, we can end the algorithm faster by using $\tilde{\pi}$ than using π. Hence, by repeating the above procedure to replace all the signaling schemes in the original algorithm Π that use external signals, we obtain a new algorithm Π' that only uses boundary and internal signals with smaller or equal sample complexity. □

3.7.2 Proof of Lemma 3.5

Proof. Let Π be any adaptive algorithm using signaling schemes with boundary and internal signals. Let $\mathcal{H}_t = \{(\pi_1, \theta_1, s_1, a_1), \ldots, (\pi_t, \theta_t, s_t, a_t)\}$ be any history that can happen during the execution of Π. If no boundary signal has been sent, then every realized signal s_k is an internal signal in the respective signaling scheme π_k, with the τ-biased posterior satisfying $\mu^\tau_{s_k} = \tau\mu_0 + (1-\tau)\mu_{s_k} \in R_{a_0}$. Because $R_{a_0} = \big\{\mu \in \Delta(\Theta) \mid \forall a \in A \setminus \{a_0\}, c_a^\top \mu > 0\big\}$ is an open region, there must exist some $\varepsilon_k > 0$ such that the ℓ_1-norm ball $B_{\varepsilon_k}(\mu^\tau_{s_k}) = \{\mu \in \Delta(\Theta) : \|\mu - \mu^\tau_{s_k}\|_1 \le \varepsilon_k\}$ is a subset of R_{a_0}. Let $\varepsilon = \min_{k=1}^t \varepsilon_k > 0$. Then $B_\varepsilon(\mu^\tau_{s_k}) \subseteq R_{a_0}$ for every $k = 1, \ldots, t$. Suppose the agent's bias level w is in the range $[\tau - \frac{\varepsilon}{2}, \tau + \frac{\varepsilon}{2}]$. Then, for every signal s_k, the agent's biased belief $\nu_{s_k} = w\mu_0 + (1-w)\mu_{s_k}$ satisfies:

$$\|\nu_{s_k} - \mu^\tau_{s_k}\|_1 = \|(w-\tau)(\mu_0 - \mu_{s_k})\|_1 \le |w - \tau| \cdot \|\mu_0 - \mu_{s_k}\|_1 \le \varepsilon.$$

This means

$$\nu_{s_k} \in B_\varepsilon(\mu^\tau_{s_k}) \subseteq R_{a_0}.$$

So, the agent should take action a_0 given signal s_k. Note that this holds for every $k = 1, \ldots, t$ and any $w \in [\tau - \frac{\varepsilon}{2}, \tau + \frac{\varepsilon}{2}]$. So we cannot determine whether $w \ge \tau$ or $w \le \tau$ so far. We have to run the algorithm until a boundary signal is sent. □

3.7.3 Proof of Lemma 3.6

Proof. Let π be a signaling scheme that can test whether $w \ge \tau$ or $w \le \tau$. According to Lemma 3.4, π can be assumed to only use boundary and internal signals. Recall that a signal s is boundary if the τ-biased belief $\mu^\tau_s = \tau\mu_0 + (1-\tau)\mu_s$ lies on the boundary set ∂R_{a_0}. For $a \in A \setminus \{a_0\}$, let B_a be the set of beliefs under which the agent is indifferent between a and a_0 and a and a_0 are both better than other actions:

$$B_a = \{\mu \in \Delta(\Theta) \mid c_a^\top \mu = 0 \ \text{ and } \ \forall a' \in A, c_{a'}^\top \mu \geq 0\}.$$

The boundary set ∂R_{a_0} can be written as the union of B_a for $a \in A \setminus \{a_0\}$:

$$\partial R_{a_0} = \bigcup_{a \in A \setminus \{a_0\}} B_a.$$

Then, we classify the boundary signals into $|A|-1$ sets $\{S_a\}_{a \in A \setminus \{a_0\}}$ according to which B_a sets their τ-biased beliefs belong to: namely, the set S_a contains boundary signals s under which

$$\tau\mu_0 + (1-\tau)\mu_s \in B_a.$$

We then *combine* the signals in S_a. Specifically, consider the normalized weighted average of the true posterior beliefs associated with the signals in S_a, denoted by μ_a:

$$\mu_a = \sum_{s \in S_a} \frac{\pi(s)}{\sum_{s' \in S_a} \pi(s')} \mu_s.$$

Note that the τ-biased version of μ_a is also in the set B_a because B_a is a convex set:

$$\tau\mu_0 + (1-\tau)\mu_a = \sum_{s \in S_a} \frac{\pi(s)}{\sum_{s' \in S_a} \pi(s')} \Big(\tau\mu_0 + (1-\tau)\mu_s\Big) \in B_a.$$

This means that if a signal a induces true posterior μ_a, then this signal is a boundary signal.

After defining μ_a as above for every $a \in A \setminus \{a_0\}$, let's consider the set of internal signals of the signaling scheme π, which we denote by S_I. For each internal signal $s \in S_I$, the τ-biased belief satisfies

$$\tau\mu_0 + (1-\tau)\mu_s \in R_{a_0}.$$

Similar to above, we combine all the signals in S_I: define μ_{a_0} to be the normalized weighted average of the posteriors associated with all internal signals:

$$\mu_{a_0} = \sum_{s \in S_I} \frac{\pi(s)}{\sum_{s' \in S_I} \pi(s')} \mu_s.$$

Then, the τ-biased version of μ_{a_0} must be in R_{a_0} because R_{a_0} is a convex set:

$$\tau\mu_0 + (1-\tau)\mu_{a_0} = \sum_{s \in S_I} \frac{\pi(s)}{\sum_{s' \in S_I} \pi(s')} \Big(\tau\mu_0 + (1-\tau)\mu_s\Big) \in R_{a_0}.$$

This means that, if a signal induces posterior μ_{a_0}, then this signal is internal.

From the splitting lemma (Lemma 3.1), we know that the convex combination of the original posteriors $\sum_{s \in S} \pi(s)\mu_s$ is equal to the prior μ_0. This means that the following convex combination of the new posteriors $\{\mu_a\}_{a \in A \setminus a_0}$ and μ_{a_0} is also equal to the prior:

$$\sum_{a \in A \setminus a_0} \sum_{s' \in S_a} \pi(s')\mu_a + \sum_{s' \in S_I} \pi(s')\mu_{a_0} = \sum_{a \in A \setminus a_0} \sum_{s \in S_a} \pi(s)\mu_s + \sum_{s \in S_I} \pi(s)\mu_s = \sum_{s \in S} \pi(s)\mu_s = \mu_0$$

where the convex combination weight of μ_a is $\sum_{s' \in S_a} \pi(s')$ for every $a \in A \setminus \{a\}$ and the convex combination weight of μ_{a_0} is $\sum_{s' \in S_I} \pi(s')$. One can easily verify that the weights sum to 1. Then, by the splitting lemma (Lemma 3.1), there exists a signaling scheme π' with signal space of size $|A|$ (so we simply denote the signal space by A) where each signal $a \in A$ induces posterior μ_a. We show that this new signaling scheme π' satisfies the properties in Lemma 3.6:

- Signal a_0 induces posterior μ_{a_0} whose τ-biased version satisfies $\tau\mu_0 + (1-\tau)\mu_{a_0} \in R_{a_0}$. So, given signal a_0, action a_0 is the optimal action for an agent with bias level τ.
- For each signal $a \in A \setminus \{a_0\}$, the induced posterior μ_a satisfies $\tau\mu_0 + (1-\tau)\mu_a \in$

$B_a \subseteq \partial R_{a_0}$. So, by the definition of B_a, an agent with bias level τ is indifferent between actions a and a_0 and these two actions are better than other actions. Also, this signal is a boundary signal by Definition 3.2, which satisfies the following according to Lemma 3.3: if the agent's bias level $w < \tau$, then the agent strictly prefers a over a_0; if $w > \tau$, then the agent strictly prefers a_0 over a.

- The sample complexity of π' is the same as π because: (1) the sample complexity is equal to the inverse of the total probability of boundary signals (as a corollary of Lemma 3.5), and (2) the total probability of boundary signals of the two signaling schemes are the same:

  $$\sum_{a \in A \setminus \{a_0\}} \pi'(a) = \sum_{a \in A \setminus \{a_0\}} \sum_{s' \in S_a} \pi(s') = \sum_{s \in \cup_{a \in A \setminus \{a_0\}} S_a} \pi(s).$$

  So, $T_\tau(\pi') = T_\tau(\pi)$.

□

Chapter 4

Learning to Coordinate Bidders in Non-Truthful Auctions

based on joint work with Hu Fu [FL20]

Switching from information design to mechanism design, this chapter studies incentive design problems in an archetypical type of mechanisms – *auctions*.

4.1 Introduction

Non-truthful auctions, among which the most ubiquitous and fundamental example is perhaps the first-price auction (FPA), are gaining more popularity over truthful auctions (such as the second-price auction) in the online advertising markets for various reasons in recent years [AL18, Slu19, Raj19, PLST20, GWMS22].

Non-truthful auctions require strategic bidding. In auctions like FPA, the independent choices of bidding strategies by bidders, with the corresponding equilibrium notion – Bayes Nash equilibrium (BNE) – are known to be problematic. For example, the independent learning dynamics of bidders may oscillate, causing undesirable instability to the system, or lead to outcomes with low welfare or low revenue (as shown by previous works [EO07, BS22, BLO$^+$25] and my work [DHLZ22] in Chapter 6). The BNE of FPA is notoriously difficult to characterize or compute when bidders' private valuations are not identically and independently distributed [CP23, FRGHK24].

A potential approach to designing better auction systems, then, is to *coordinate the* *bidders*. Correlated equilibria are easier to compute than BNE [FRGHK24], and coordination can potentially stabilize the system and lead to better outcomes for both the bidders and the auctioneer than the independent outcomes. In modern auction systems such as online advertising auctions where bidders delegate the bidding task to platforms that run auto-bidding algorithms [ABB$^+$24], those platforms can, at least in principle, coordinate different bidders' bids [DGPS23, CWD$^+$23]. A desideratum here is *incentive compatibility*: bidders should be willing to report their private values to the coordinator truthfully and submit the bids that are recommended by the coordinator. In other words, the coordinator has to find a *Bayesian correlated equilibrium* (BCE) for the bidders. A BCE, however, is sensitive to the distribution of bidders' private values, which is often unavailable in practice. Can the coordinator find a BCE using samples of bidders' private values? How many samples are needed? These two questions are the focus of our work.

Overview of Our Contributions We initiate the study of the sample complexity of Bayesian correlated equilibrium in non-truthful auctions. As there are multiple notions of BCE in the literature [For06], we focus on the strategic-form BCE. We show that the strategic-form ε-BCEs of a large class of non-truthful auctions (including first-price and all-pay auctions) can be found with a polynomial number of samples, $\tilde{O}(\frac{n}{\varepsilon^2})$, where n is the number of bidders.[1] This result holds for any distributions of bidders' private values that are bounded and independent. The moderate amount of samples demonstrates the practicality of learning to coordinate bidders in non-truthful auctions.

A more technical contribution of our work is a *utility estimation* result. We show that bidders' expected utilities under all monotone bidding strategies can be estimated with high precision using $\tilde{O}(\frac{n}{\varepsilon^2})$ samples. This result implies that ε-BCEs can be learned from the samples. Interestingly, the utility estimation result does not hold for all bidding strategies (including non-monotone ones), as we will show in Proposition 4.2. But because

[1] The $\tilde{O}(\cdot)$ notation omits logarithmic factors.

strategic-form BCEs are always monotone (Proposition 4.1), it suffices to prove the utility estimation result for monotone bidding strategies only. We connect this problem to the PAC (probably approximately correct) learning literature, and prove the utility estimation result by a non-trivial analysis of the pseudo-dimension of the class of utility functions under monotone strategies.

4.1.1 Related Works

Coordination in Auctions. Bidder's coordination, or collusion, is a long-standing topic in the traditional auction literature [GM87, MZ91, MM92, MM07, HPT08, LMS11] and has recently been studied in the online ad auction domain as well [DGP20, DGPS23, CWD+23]. Contrary to the previous view that collusion can undermine the auctioneer's revenue, we take a positive viewpoint here: coordination might be desirable for the system designer. This is because: (1) coordination can potentially prevent the unstable strategizing behaviors of independent bidders; (2) the set of BCEs is larger than the set of BNEs in theory, so the coordinator can potentially induce an equilibrium with a (weakly) higher revenue or welfare than any independent equilibrium.

Bayesian Correlated Equilibrium in General Games. Incentivizing bidders to coordinate is closely related to finding a Bayes correlated equilibrium in the auction game. The classical notion of correlated equilibrium [Aum74] is defined for complete information games. For incomplete information games like auctions with private values, the literature has defined multiple notions of Bayes correlated equilibria, such as strategic form BCE, agent normal form BCE, and communication equilibrium [Mye82, For06, BM16, Fuj23]. We consider the strategic form BCE [For06] where the coordinator recommends randomized joint bidding strategies to all bidders without knowing the bidders' private values. This type of BCE satisfies monotonicity (as we will show in Proposition 4.1) and does not alter any bidder's belief about other bidders' private values. These two crucial properties ensure the learnability of the BCE when bidders' value distribution is unknown.

Sampling from Value Distributions. An assumption of our work is that the learner has sample access to the underlying distribution of bidders' values. This is a standard assumption in the literature of learning in mechanism design [e.g. CR14, MR15, MR16, BSV16, BSV18, GN17, Syr17, GW18, GHZ19, BCD20, YB21, GHTZ21]. While most of those works study revenue maximization in truthful auctions, we consider the under-explored problems of utility estimation and equilibrium learning in non-truthful auctions.

Value samples have been assumed in the context of learning in non-truthful auctions [BSV19, Vit21]. Just as in classical microeconomics, prior knowledge (in the form of samples here) comes from market research, survey, simulation etc., and is not assumed to be from past bidding history. We distance our approach from the line of work on learning non-truthful auctions where samples are from past bidding history [CHN17, HT19]. This latter approach, with obvious merits, has its limitations. Crucially, it assumes that the observed bidding in a non-truthful auction is at equilibrium, which may not be the case in reality. Also, to avoid strategic issues between auctions, the bidders need to be short-lived or myopic. The two approaches (value samples vs. bid samples) complement each other even in learning problems for non-truthful auctions. This work takes the first approach, and leaves the direction with bid samples as an enticing open question.

Utility Estimation in Games. Given a non-truthful auction, [BSV19] studied the number of value samples needed to learn the maximal utility a bidder could gain by non-truthful bidding, when all other bidders are truthful. In comparison, we learn utilities when all bidders use arbitrary monotone bidding strategies; this suffices for the study of virtually all properties of an auction, including the task of [BSV19].[2]

[2] Our results imply that the maximal utility (w.r.t the opponents' value distribution) obtained by non-truthful bidding can be approximated by the maximal obtainable utility w.r.t. the empirical distribution, which can be computed by enumerating the samples in the empirical distribution because a

[AVGCU19, MTG20, DHZ$^+$23] studied utility estimation and equilibrium learning for general normal-form games with random utility matrices (with sample access). Similar to us, they frame the problem as a PAC learning problem and bound the number of samples using complexity measures (Rademacher complexity, covering number) of some function classes. But they did not characterize those complexity measures for specific games. Bounding those complexity measures is generally challenging, if not impossible. For example, without monotonicity in bidding strategies, the pseudo-dimension of utilities in first-price auctions is unbounded, as implied by our Proposition 4.2.

Equilibrium Computation in Auctions. There is a large literature on the computation of equilibrium in non-truthful auctions [e.g. MMRS94, FG03, GR08, EMR09, WSZ20, CP23, FRGH$^+$21, FRGHK24]. Although there has been major progress on the computation of BNE in first-price auctions with common prior distributions [WSZ20, CP23], this problem turns out to be PPAD-hard with subjective prior distributions [FRGH$^+$21]. On the other hand, a BCE is known to be easier to compute than a BNE [FRGHK24], which provides an additional motivation for us to study the BCE of non-truthful auctions.

4.2 Preliminary: Auctions, BNE, and BCE

Auctions. Consider a single-item auction with n bidders denoted by $[n] = \{1, \ldots, n\}$. Each bidder $i \in [n]$ has a private value v_i drawn from a distribution D_i supported on $T_i \subseteq [0, H] \subseteq \mathbb{R}_+$, where H is an upper bound on the bidder's value. The size of the support $|T_i|$ can be infinite. Different bidders' values are independent and can be non-identically distributed, so the joint value distribution $\boldsymbol{D} = \prod_{i=1}^{n} D_i$ is a product distribution. Each bidder i makes a sealed-envelope bid of $b_i \in [0, H]$. The auction maps the vector of

best-responding bid must be equal to (or slightly more than) some opponent's value from the empirical distribution.

bids $\boldsymbol{b} = (b_1, \dots, b_n)$ to *allocation* and *payments*, where allocation $x_i(\boldsymbol{b}) \in [0, 1]$ is the probability with which bidder i receives the item, with $\sum_{i=1}^n x_i(\boldsymbol{b}) \le 1$, and payment $p_i(\boldsymbol{b})$ is the payment made by bidder i to the auctioneer. Bidder i's *ex post utility* is denoted by

$$U_i(v_i, \boldsymbol{b}) \coloneqq v_i x_i(\boldsymbol{b}) - p_i(\boldsymbol{b}). \tag{4.1}$$

We focus on the allocation rule where the bidder with the highest bid wins, with ties broken randomly: $x_i(\boldsymbol{b}) = \frac{\mathbb{1}[b_i = \max_{j \in [n]} b_j]}{|\arg\max_{j \in [n]} b_j|}$. Auctions with reserve prices can be modeled by adding an additional bidder who always bids the reserve price.

We consider any payment function of the following form:

$$p_i(\boldsymbol{b}) = x_i(\boldsymbol{b}) f_i(b_i) + g_i(b_i). \tag{4.2}$$

where functions f_i and g_i satisfy $0 \le f_i(b_i), g_i(b_i) \le H$. For example,

- in the *first-price auction* (FPA), the highest bidder wins the item and pays her bid, and other bidders pays zero: $p_i^{\mathrm{FPA}}(\boldsymbol{b}) = x_i(\boldsymbol{b}) b_i$.
- In the *all-pay auction* (APA), the highest bidder wins the item but all bidders pay their bids: $p_i^{\mathrm{APA}}(\boldsymbol{b}) = b_i$. APA is a good model for, e.g., crowdsourcing [CHS12].

We assume that, fixing the bids $\boldsymbol{b}_{-i}$ of other bidders, if bidder i wins the item at bid b_i, then her payment must be strictly increasing in her bid:

$$x_i(b_i, \boldsymbol{b}_{-i}) > 0 \implies p_i(b_i', \boldsymbol{b}_{-i}) > p_i(b_i, \boldsymbol{b}_{-i}), \quad \forall b_i' > b_i. \tag{4.3}$$

This condition is satisfied by both first-price and all-pay auctions.

Strategies and Equilibria. A *(bidding) strategy* $\sigma_i : T_i \to [0, H]$ is a mapping from the bidder's value v_i to bid $b_i = \sigma_i(v_i)$. Let $\Sigma_i = T_i \to [0, H]$ be the strategy space of bidder i, and let $\boldsymbol{\Sigma} = \prod_{i=1}^{n} \Sigma_i$ be the joint strategy space of all bidders. Let $\boldsymbol{b} = \boldsymbol{\sigma}(\boldsymbol{v}) = (\sigma_1(v_1), \dots, \sigma_n(v_n))$ denote the bids of all bidders, and $\boldsymbol{b}_{-i} = \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})$ denote the bids of bidders except i. When other bidders use strategies $\boldsymbol{\sigma}_{-i}$, bidder i with value v_i and bid b_i obtains *interim utility*

$$\begin{aligned} u_i(v_i, b_i, \boldsymbol{\sigma}_{-i}) &\coloneqq \mathbb{E}_{\boldsymbol{v}_{-i}\sim \boldsymbol{D}_{-i}}\left[U_i(v_i, b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))\right] \\ &= \mathbb{E}_{\boldsymbol{v}_{-i}\sim \boldsymbol{D}_{-i}}\left[v_i x_i(b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) - p_i(b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))\right]. \end{aligned} \tag{4.4}$$

We define Bayes Nash equilibrium for the auction game:

Definition 4.1 (Bayes Nash equilibrium)**.** *For $\varepsilon \geq 0$, a joint bidding strategy $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ is a (pure-strategy) ε-*Bayes Nash equilibrium (ε-BNE) *for value distribution $\boldsymbol{D} = \prod_{i=1}^{n} D_i$ if for each bidder $i \in [n]$, any value $v_i \in T_i$, any bid $b_i' \in [0, H]$,*

$$u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}) \;\geq\; u_i(v_i, b_i', \boldsymbol{\sigma}_{-i}) - \varepsilon.$$

When $\varepsilon = 0$, $\boldsymbol{\sigma}$ is a Bayes Nash equilibrium (BNE).

We also define a Bayes correlated equilibrium for the auction game. Correlated equilibria are typically defined for complete information games. For incomplete information games like auctions, there are multiple definitions of Bayes correlated equilibria in the literature [For06]. We consider the "strategic form Bayes correlated equilibrium" in [For06], which regards the incomplete information game as a normal form game where a player's pure strategy is the mapping σ_i. A *correlation device*, or *mediator*, can sample a joint strategy $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ from a joint distribution $Q \in \Delta(\boldsymbol{\Sigma})$, and recommend each strategy σ_i to the respective bidder i, while ensuring that no bidder has incentive to deviate from the recommended strategy.

Definition 4.2 (Bayes correlated equilibrium)**.** *For $\varepsilon \geq 0$, a distribution $Q \in \Delta(\boldsymbol{\Sigma})$ over joint bidding strategies is an ε-*Bayes correlated equilibrium *(ε-*BCE*) for value distribution $\boldsymbol{D} = \prod_{i=1}^{n} D_i$ if for each bidder $i \in [n]$, any value $v_i \in T_i$, any deviation function $\phi_i : \Sigma_i \times T_i \to [0, H]$,*

$$\mathbb{E}_{\boldsymbol{\sigma} \sim Q}\big[u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i})\big] \;\geq\; \mathbb{E}_{\boldsymbol{\sigma} \sim Q}\big[u_i(v_i, \phi_i(\sigma_i, v_i), \boldsymbol{\sigma}_{-i})\big] - \varepsilon.$$

When $\varepsilon = 0$, σ is a Bayes correlated equilibrium (BCE)*.*

A pure-strategy BNE is a BCE where bidders' joint strategy $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ is deterministic. A mixed-strategy BNE is a BCE where bidders' strategies $\sigma_1, \dots, \sigma_n$ are randomized and independent.

We say a bidding strategy σ_i is *monotone* if it is weakly increasing: $v \geq v' \Rightarrow \sigma_i(v) \geq \sigma_i(v')$, A joint bidding strategy $\boldsymbol{\sigma}$ is monotone if all individual strategies $\sigma_1, \dots, \sigma_n$ are monotone. A BCE $Q \in \Delta(\boldsymbol{\Sigma})$ is monotone if every joint strategy $\boldsymbol{\sigma}$ sampled from Q is monotone. [MR00] show that the BNEs of auction games are "essentially monotone". We generalize their result to BCEs.

Proposition 4.1. *Under Assumption* (4.3)*, any BCE $Q \in \Delta(\boldsymbol{\Sigma})$ of the auction game is "essentially monotone" in the following sense: for any joint strategy $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$ sampled from Q, every bidder i's strategy $\sigma_i(v_i)$ is weakly increasing except when v_i is too low that bidder i wins the item with probability* 0*.*

Proof. Let $Q \in \Delta(\boldsymbol{\Sigma})$ be a BCE, with $\boldsymbol{\sigma} \sim Q$. Suppose bidder i's strategy σ_i is not weakly increasing on two values $v_i < v_i'$, namely, $b_i = \sigma_i(v_i) > b_i' = \sigma_i(v_i')$. By the definition of BCE, conditioning on bidder i being recommended σ_i, we have

$$\mathbb{E}_{\boldsymbol{\sigma}_{-i}|\sigma_i}\left[u_i(v_i, b_i, \boldsymbol{\sigma}_{-i})\right] \;\geq\; \mathbb{E}_{\boldsymbol{\sigma}_{-i}|\sigma_i}\left[u_i(v_i, b_i', \boldsymbol{\sigma}_{-i})\right].$$

Define interim allocation $x_i(b_i) = \mathbb{E}_{\boldsymbol{\sigma}_{-i}|\sigma_i} \mathbb{E}_{\boldsymbol{v}_{-i}\sim \boldsymbol{D}_{-i}}[x_i(b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))]$ and interim payment $p_i(b_i) = \mathbb{E}_{\boldsymbol{\sigma}_{-i}|\sigma_i} \mathbb{E}_{\boldsymbol{v}_{-i}\sim \boldsymbol{D}_{-i}}[p_i(b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))]$. Then we have

$$v_i x_i(b_i) - p_i(b_i) \;\geq\; v_i x_i(b_i') - p_i(b_i'). \tag{4.5}$$

Switching the roles of v_i and v_i',

$$v_i' x_i(b_i') - p_i(b_i') \;\geq\; v_i' x_i(b_i) - p_i(b_i). \tag{4.6}$$

Adding (4.5) and (4.6), we obtain

$$\big(v_i' - v_i\big) \cdot \big[x_i(b_i') - x_i(b_i)\big] \;\geq\; 0.$$

Since $v_i' > v_i$, we obtain $x_i(b_i') \geq x_i(b_i)$. But under the assumption of $b_i > b_i'$, we have $x_i(b_i') \leq x_i(b_i)$ because the function $x_i(\cdot)$ is weakly increasing. Therefore, it must be

$$x_i(b_i') = x_i(b_i). \tag{4.7}$$

Plugging (4.7) into (4.5) and (4.6), we obtain

$$p_i(b_i') = p_i(b_i).$$

If $x_i(b_i') = x_i(b_i) > 0$, then we have $p_i(b_i) > p_i(b_i')$ by Assumption (4.3), which leads to a contradiction. So, it must be $x_i(b_i') = x_i(b_i) = 0$, which means that bidder i never wins the item under values v_i and v_i'. □

Since BCE is essentially monotone and any essentially monotone BCE can be converted to a monotone BCE without affect any bidder's expected utility, we will restrict attentions to monotone BCE. The set of monotone ε-BCEs depends on the bidders' value distribution

$\boldsymbol{D}$. We denote this set by

$$\mathrm{BCE}(\boldsymbol{D}, \varepsilon) = \Big\{ Q \in \Delta(\boldsymbol{\Sigma}) \;\Big|\; Q \text{ is monotone and is an } \varepsilon\text{-BCE on value distribution } \boldsymbol{D} \Big\}.$$

Our goal Our goal is to learn the set $\mathrm{BCE}(\boldsymbol{D}, \varepsilon)$ when bidders' value distribution $\boldsymbol{D}$ is unknown and can only be accessed by sampling. We aim to characterize the number of samples that are needed to achieve this goal.

4.3 Sample Complexity of Estimating Utility

A crucial step to learn the set of BCEs in an auction with unknown distribution $\boldsymbol{D}$ is to estimate the bidders' expected utility for any given joint bidding strategy $\boldsymbol{\sigma}$. We call this problem *utility estimation.* The utility estimation problem is also interesting by itself, so we study the sample complexity of utility estimation in this section.

Formally, we are given a set of m samples $\mathcal{S} = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$ from the value distribution $\boldsymbol{D} = \prod_{i=1}^{n} D_i$, where each sample $\boldsymbol{v}^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$ contains the values of all bidders, we aim the estimate the expected utility of every bidder under every possible joint bidding strategy $\boldsymbol{\sigma}$. A *utility estimation algorithm*, denoted by $\mathcal{A}$, takes the samples $\mathcal{S}$, bidder index i, value v_i, and all bidders' strategies $\boldsymbol{\sigma}$ as input, outputs $\mathcal{A}(\mathcal{S}, i, v_i, \boldsymbol{\sigma})$ to estimate bidder i's interim utility $u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}) = \mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{D}_{-i}}[U_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))]$.

Definition 4.3 (utility estimation). *Let $\mathcal{B} \subseteq \boldsymbol{\Sigma}$ be a set of joint bidding strategies. For $\varepsilon > 0, \delta \in (0, 1)$, we say an algorithm $\mathcal{A}$ (ε, δ)-*estimates with m sample the utilities over $\mathcal{B}$ *if, for any value distribution $\boldsymbol{D}$, with probability at least $1 - \delta$ over the random draw of m samples from $\boldsymbol{D}$, for any joint bidding strategy $\boldsymbol{\sigma} \in \mathcal{B}$, for each* *bidder $i \in [n]$ and any value $v_i \in T_i$,*

$$\left| \mathcal{A}(\mathcal{S}, i, v_i, \boldsymbol{\sigma}) - u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}) \right| < \varepsilon.$$

We aim to estimate the interim utility in the above definition, instead of the ex ante utility $\mathbb{E}_{\boldsymbol{v}\sim\boldsymbol{D}}[U_i(v_i, \boldsymbol{\sigma}(\boldsymbol{v}))]$, because the ex ante utility is difficult to estimate due to the randomness of bidder i's own value. Even in a first-price auction with a single bidder, the bidder's ex ante utility $\mathbb{E}_{v_i\sim D_i}[v_i - \sigma_i(v_i)]$ cannot be estimated using finitely many samples for all distributions D_i and for all strategies σ_i simultaneously. The interim utility $u_i(v_i, b_i, \boldsymbol{\sigma}_{-i}) = \mathbb{E}_{\boldsymbol{v}_{-i}\sim\boldsymbol{D}_{-i}}[U_i(v_i, b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}))]$, on the other hand, does not involve randomization over bidder i's own value and is easier to estimate.

We note that the utility estimation problem cannot be solved if $\mathcal{B}$ contains all possible bidding strategies, including non-monotone and monotone ones.

Proposition 4.2. *The utility estimation problem cannot be solved with finitely many samples if $\mathcal{B}$ contains all possible bidding strategies.*

Proof. Consider an auction with two bidders, the first bidder having value 1 and bidding $\frac{1}{2}$, and the second bidder's value v_2 uniformly drawn from $[0, 1]$. Any finite set of samples of v_2 has probability measure 0 in the distribution of v_2. Therefore on any set of samples, there are bidding strategies of bidder 2 that look the same on the sampled values but give bidder 1 drastically different utilities in expectation on the value distribution. □

Given the above impossibility result, we restrict $\mathcal{B}$ to be the set of monotone joint bidding strategies. This is without loss of generality according to Proposition 4.1.

4.3.1 Upper Bound of Sample Complexity of Utility Estimation

In this subsection, we show that $\tilde{O}(n/\varepsilon^2)$ value samples suffice for estimating the interim utilities for all monotone bidding strategies. The estimation algorithm is the empirical distribution estimator, which outputs the expected utility on the uniform distribution over the samples.

Definition 4.4. *The* empirical distribution estimator*, denoted by* Emp*, estimates interim utilities on the uniform distribution over the samples. Formally, on samples* $\mathcal{S} = \{\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(m)}\}$*, for bidder* i *with value* v_i*, for joint bidding strategy* $\boldsymbol{\sigma}$*,*

$$\mathrm{Emp}(\mathcal{S}, i, v_i, \boldsymbol{\sigma}) \coloneqq \frac{1}{m}\sum_{j=1}^{m} U_i\big(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i}^{(j)})\big).$$

We now state an upper bound on the sample complexity of utility estimation, which is $\tilde{O}(\frac{H^2}{\varepsilon^2}n)$ when ignoring logarithmic factors.

Theorem 4.1 (Utility estimation by empirical distribution)**.** *Suppose* $T_i \subseteq [0, H]$*. For any* $\varepsilon > 0, \delta \in (0,1)$*, there is*

$$M = O\left(\frac{H^2}{\varepsilon^2}\left[n \log n \log\left(\frac{H}{\varepsilon}\right) + \log\left(\frac{n}{\delta}\right)\right]\right),$$

such that for any $m \geq M$*, the empirical distribution estimator* Emp (ε, δ)*-estimates with* m *samples the utilities over the set of all monotone bidding strategies.*

Pseudo-Dimension and the Proof of Theorem 4.1

To prove Theorem 4.1, we use a tool called *pseudo-dimension* from stasitical learning theory (see, e.g., [AB09]), which captures the complexity of a class of functions.

Definition 4.5. *Let $\mathcal{H}$ be a class of real-valued functions on input space $\mathcal{X}$. A set of inputs $x_1, \ldots, x_m$ is said to be* pseudo-shattered *if there exist* witnesses $t_1, \ldots, t_m \in \mathbb{R}$ *such that for any label vector $\boldsymbol{l} \in \{1, -1\}^m$, there exists $h_{\boldsymbol{l}} \in \mathcal{H}$ such that* $\operatorname{sgn}(h_{\boldsymbol{l}}(x_i) - t_i) = l_i$ *for each $i = 1, \ldots, m$, where* $\operatorname{sgn}(y) = 1$ *if $y > 0$ and -1 if $y < 0$. The* pseudo-dimension *of $\mathcal{H}$,* $\operatorname{Pdim}(\mathcal{H})$*, is the size of the largest set of inputs that can be pseudo-shattered by $\mathcal{H}$.*

Definition 4.6. *For $\varepsilon > 0, \delta \in (0,1)$, a class of functions $\mathcal{H} : \mathcal{X} \to \mathbb{R}$ is (ε, δ)-*uniformly convergent with sample complexity M *if for any $m \geq M$, for any distribution D on $\mathcal{X}$, if $x^{(1)}, \ldots, x^{(m)}$ are i.i.d. samples from D, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$, $\left| \mathbb{E}_{x \sim D}[h(x)] - \frac{1}{m} \sum_{j=1}^{m} h(x^{(j)}) \right| < \varepsilon$.*

Theorem 4.2 (See, e.g., [AB09])**.** *Let $\mathcal{H}$ be a class of functions with range $[0, H]$ and pseudo-dimension $d =$* $\operatorname{Pdim}(\mathcal{H})$*, for any $\varepsilon > 0$, $\delta \in (0,1)$, $\mathcal{H}$ is (ε, δ)-uniformly convergent with sample complexity $O\left(\frac{H^2}{\varepsilon^2} \left[d \log(\frac{H}{\varepsilon}) + \log(\frac{1}{\delta}) \right] \right)$.*

We prove Theorem 4.1 by treating the utilities on monotone bidding strategies as a class of functions, whose uniform convergence implies that Emp learns the interim utilities.

For each bidder i, let $h^{v_i, \boldsymbol{\sigma}}$ be the function that maps the opponents' values to bidder i's ex post utility:

$$h^{v_i, \boldsymbol{\sigma}}(\boldsymbol{v}_{-i}) = U_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})).$$

Let $\mathcal{H}_i$ be the set of all such functions corresponding to the set of monotone strategies,

$$\mathcal{H}_i = \left\{ h^{v_i, \boldsymbol{\sigma}}(\cdot) \;\middle|\; v_i \in T_i, \quad \boldsymbol{\sigma} \text{ is monotone} \right\}.$$

By Equation (4.4), the expectation of $h^{v_i,\boldsymbol{\sigma}}(\cdot)$ over $\boldsymbol{D}_{-i}$ is the interim utility of bidder i:

$$\mathbb{E}_{\boldsymbol{v}_{-i}\sim\boldsymbol{D}_{-i}}\left[h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i})\right] = u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}).$$

By Definition 4.4, on samples $\mathcal{S} = \{\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(m)}\}$, $\mathrm{Emp}(\mathcal{S}, i, v_i, \boldsymbol{\sigma}) = \frac{1}{m}\sum_{j=1}^{m} h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i}^{(j)})$. Thus,

$$\left|\mathrm{Emp}(\mathcal{S}, i, v_i, \boldsymbol{\sigma}) - u_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i})\right| = \left|\mathbb{E}_{\boldsymbol{v}_{-i}\sim\boldsymbol{D}_{-i}}\left[h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i})\right] - \frac{1}{m}\sum_{j=1}^{m} h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i}^{(j)})\right|. \tag{4.8}$$

The right hand side of (4.8) is the difference between the expectation of $h^{v_i,\boldsymbol{\sigma}}$ on the distribution $\boldsymbol{D}_{-i}$ and that on the empirical distribution with samples drawn from $\boldsymbol{D}_{-i}$. Now by Theorem 4.2, to bound the number of samples needed by Emp to (ε, δ)-estimate the utilities over monotone strategies, it suffices to bound the pseudo-dimension of $\mathcal{H}_i$. With the following key lemma, the proof of Theorem 4.1 is completed by observing that the range of each $h^{v_i,\boldsymbol{\sigma}}$ is within $[-H, H]$ and by taking a union bound over $i \in [n]$.

Lemma 4.1. $\mathrm{Pdim}(\mathcal{H}_i) = O(n \log n)$.

The proof of Lemma 4.1 follows a powerful framework introduced by [MR16] and [BSV18] for bounding the pseudo-dimension of a class $\mathcal{H}$ of functions: given inputs that are to be pseudo-shattered, fixing any witnesses, one classifies the functions in $\mathcal{H}$ into subclasses, such that the functions in the same subclass output the same label on all the inputs; by counting and bounding the number of subclasses, one can bound the number of shattered inputs. Our proof follows this strategy. To bound the number of subclasses, we make use of the monotonicity of bidding functions, which is specific to our problem.

Proof of Lemmea 4.1. By definition, given any v_i and $\boldsymbol{\sigma}$ (with $b_i = \sigma_i(v_i)$), the output of $h^{v_i,\boldsymbol{\sigma}}$ on input $\boldsymbol{v}_{-i}$ is

$$h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i}) = v_i x_i(\boldsymbol{b}) - p_i(\boldsymbol{b}) = v_i x_i(\boldsymbol{b}) - x_i(\boldsymbol{b}) f_i(b_i) - g_i(b_i)$$
$$= \big(v_i - f_i(b_i)\big) x_i(\boldsymbol{b}) - g_i(b_i).$$

Because the allocation $x_i(\boldsymbol{b})$ is that the highest bidder wins with random tie breaking, $h^{v_i,\boldsymbol{\sigma}}(\boldsymbol{v}_{-i})$ must take one of the following $n+1$ values:

$$v_i - f_i(b_i) - g_i(b_i),\ \tfrac{v_i - f_i(b_i)}{2} - g_i(b_i),\ \ldots,\ \tfrac{v_i - f_i(b_i)}{n} - g_i(b_i),\ 0 - g_i(b_i).$$

This value is fully determined by the $n-1$ comparisons $b_i \lesseqgtr \sigma_j(v_j^k)$, one for each $j \neq i$.

Let $\boldsymbol{v}_{-i}^{(1)}, \ldots, \boldsymbol{v}_{-i}^{(m)}$ be any m inputs. We argue that the function class $\mathcal{H}_i$ can be divided into $O(m^{2n})$ sub-classes $\{\mathcal{H}_i^{\mathbf{k}}\}_{\mathbf{k}\in[m+1]^{2(n-1)}}$ such that each sub-class $\mathcal{H}_i^{\mathbf{k}}$ generates at most $O(m^n)$ different label vectors on the m inputs. Thus $\mathcal{H}_i$ generates at most $O(m^{3n})$ label vectors in total. To pseudo-shatter m inputs, we need $O(m^{3n}) \geq 2^m$, which implies $m = O(n \log n)$.

We now define sub-classes $\{\mathcal{H}_i^{\mathbf{k}}\}_{\mathbf{k}}$, each indexed by $\mathbf{k} \in [m+1]^{2(n-1)}$. For each dimension $j \in [n] \setminus \{i\}$, we sort the m inputs by their j-th coordinates non-decreasingly, and use $\pi(j,\cdot)$ to denote the resulting permutation over $\{1, 2, \ldots, m\}$: formally, $v_j^{(\pi(j,1))} \leq v_j^{(\pi(j,2))} \leq \cdots \leq v_j^{(\pi(j,m))}$. For each function $h^{v_i,\boldsymbol{\sigma}}(\cdot)$, for each j, we define two special positions:

$$k_{j,1} = \max\Big\{0,\ \big\{k : \sigma_j(v_j^{(\pi(j,k))}) < b_i\big\}\Big\},$$
$$k_{j,2} = \min\Big\{m+1,\ \big\{k : \sigma_j(v_j^{(\pi(j,k))}) > b_i\big\}\Big\}.$$

These two positions are well defined because $\sigma_j(\cdot)$ is monotone. By definition, if $k_{j,1} <$

$k_{j,2} - 1$, then for any k such that $k_{j,1} < k < k_{j,2}$, we must have $\sigma_j(v_j^{(\pi(j,k))}) = b_i$. We let a function $h^{v_i,\boldsymbol{\sigma}}(\cdot)$ to belong to the sub-class $\mathcal{H}_i^{\mathbf{k}}$ where the index $\mathbf{k}$ is $(k_{j,1}, k_{j,2})_{j\in[n]\setminus\{i\}}$. The number of sub-classes is the number of indices, which is bounded by $(m+1)^{2(n-1)}$.

We now show that the functions within a sub-class $\mathcal{H}_i^{\mathbf{k}}$ give rise to at most $(m+1)^n$ label vectors on the m inputs. Let us focus on one such class with index $\mathbf{k}$. On the k-th input $\boldsymbol{v}_{-i}^{(k)}$, a function's membership in $\mathcal{H}_i^{\mathbf{k}}$ suffices to specify whether bidder i is a winner on this input, and, if so, the number of other bidders winning at a tie. Therefore, the class index $\mathbf{k}$ determines a mapping $c : [m] \to \{0, 1, \ldots, n\}$, with $c(k) > 0$ meaning bidder i is a winner on input $\boldsymbol{v}_{-i}^{(k)}$ at a tie with $c(k) - 1$ other bidders, and $c(k) = 0$ meaning bidder i is a loser on input $\boldsymbol{v}_{-i}^{(k)}$. Then, the output of a function $h^{v_i,\boldsymbol{\sigma}}(\cdot) \in \mathcal{H}_i^{\mathbf{k}}$ on input $\boldsymbol{v}_{-i}^{(k)}$ is $\frac{v_i - f_i(b_i)}{c(k)} - g_i(b_i)$ if $c(k) > 0$ and $-g_i(b_i)$ otherwise. The same utility is output on two inputs $\boldsymbol{v}_{-i}^{(k)}$ and $\boldsymbol{v}_{-i}^{(k')}$ whenever $c(k) = c(k')$. Consider the set $S \subseteq [m]$ of inputs that are mapped to one integer by c, and fix any $|S|$ witnesses. By varying the function in the subclass $\mathcal{H}_i^{\mathbf{k}}$, we can generate at most $|S| + 1 \leq m + 1$ patterns of labels on the input set S, because we are comparing the same utility with $|S|$ witnesses. The label vector for the entire input set $[m]$ is the concatenation of these patterns of labels. Since the image of c has $n + 1$ integers, and there are at most $(m+1)^n$ label vectors.

To conclude, the total number of label vectors generated by $\mathcal{H}_i = \bigcup_{\mathbf{k}} \mathcal{H}_i^{\mathbf{k}}$ is at most

$$(m+1)^{2(n-1)}(m+1)^{n+1} \leq (m+1)^{3n}.$$

To pseudo-shatter m inputs, we need $(m+1)^{3n} \geq 2^m$, which implies $m = O(n \log n)$. □

4.3.2 Lower Bound of Sample Complexity of Utility Estimation

We give an information-theoretic lower bound on the number of samples needed for any algorithm to estimate utilities over monotone strategies in a first-price auction. The lower bound matches our upper bound up to polylogarithmic factors.

Theorem 4.3. *For any $\varepsilon < \frac{1}{4000}, \delta < \frac{1}{20}$, there is a family of product value distributions for which no algorithm can (ε, δ)-estimate utilities over the set of all monotone bidding strategies with $m \leq \frac{1}{4\times 10^8} \cdot \frac{n}{\varepsilon^2}$ samples.*

The proof of Theorem 4.3 is in Section 4.6.1. As a sketch, the product value distributions we construct encode length $n-1$ binary strings by having a slightly unfair Bernoulli distribution for each bidder, the bias shrinking as n grows large. We then show that, if with a few samples a learning algorithm can estimate utilities for all monotone bidding strategies, then there must exist two product value distributions from the family that differ at only one coordinate, and yet they can be told apart by the learning algorithm. This must violate the well-known information-theoretic lower bound for distinguishing two distributions [Man11].

4.4 Sample Complexity of Learning BCE

This section studies how to learn BCEs using samples from the value distribution $\boldsymbol{D}$.

4.4.1 Estimating Utility by Empirical Product Distributions

Section 4.3.1 shows that the empirical distribution estimator approximates interim utilities with high probability. However, this does not immediately imply that the auction on the empirical distribution is a close approximation to the auction on the original distribution $\boldsymbol{D}$. This is because the empirical distribution over samples is *correlated* — the values $\boldsymbol{v}^{(j)} = (v_1^{(j)}, \ldots, v_n^{(j)})$ are drawn as a vector, instead of independently. The equilibria (either BCE or BNE) with respect to this correlated empirical distribution do not correspond to the equilibria on the original product distribution $\boldsymbol{D}$. Therefore, it is desirable that utilities can also be estimated on a *product* distribution arising from the samples, where each bidder's value is independently drawn, uniformly from the m samples of her value. We show that this can indeed be done, without a substantial increase in the sample complexity. The key technical step, Lemma 4.2, is a reduction from learning on empirical distribution to learning on empirical product distribution. We believe this lemma is of independent interest. In fact, in the published paper [FL20] we use Lemma 4.2 to derive the sample complexity for another problem (Pandora's Box problem).

Definition 4.7. *Given samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ from a product distribution $\boldsymbol{D} = \prod_{i=1}^n D_i$, let E_i be the uniform distribution over $\{x_i^{(1)}, \ldots, x_i^{(m)}\}$. The* empirical product distribution *is the product distribution $\boldsymbol{E} = \prod_{i=1}^n E_i$.*

Definition 4.8. *For $\varepsilon > 0, \delta \in (0,1)$, a class of functions $\mathcal{H} : \prod_{i=1}^n T_i \to \mathbb{R}$ is (ε, δ)-*uniformly convergent on product distribution with sample complexity *M if for any $m \geq M$, for any product distribution $\boldsymbol{D}$ on $\prod_{i=1}^n T_i$, if $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ are i.i.d. samples from $\boldsymbol{D}$, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$,*

$$\left| \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{D}}\left[h(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{E}}\left[h(\boldsymbol{x})\right] \right| \; < \; \varepsilon,$$

where $\boldsymbol{E} = \prod_{i=1}^n E_i$ is the empirical product distribution.

Lemma 4.2. *Let $\mathcal{H}$ be a class of functions from a product space $\boldsymbol{T} = \prod_{i=1}^n T_i$ to $[0, H]$. If $\mathcal{H}$ is (ε, δ)-uniformly convergent with sample complexity $m(\varepsilon, \delta)$, then $\mathcal{H}$ is $\left(2\varepsilon, \frac{H\delta}{\varepsilon}\right)$-uniformly convergent on product distribution with sample complexity $m(\varepsilon, \delta)$. In other words, $\mathcal{H}$ is (ε', δ')-uniformly convergent with sample complexity $m(\frac{\varepsilon'}{2}, \frac{\delta'\varepsilon'}{2H})$.*

Lemma 4.2 is closely related to a concentration inequality by [DHP16], who show that for any *single* function $h : \boldsymbol{T} \to [0, H]$, the expectation of h on the empirical product distribution $\boldsymbol{E}$ is close to its expectation on the original distribution $\boldsymbol{D}$ with high probability. Our lemma generalizes [DHP16] to the simultaneous concentration for a family of functions, and seems more handy for applications such as ours.

Proof of Lemma 4.2. Write the m samples $\mathcal{S} = \{\boldsymbol{x}^1, \dots, \boldsymbol{x}^m\}$ from $\boldsymbol{D}$ as an $m \times n$ matrix (x_i^j), where each row $j \in [m]$ represents a sample $\boldsymbol{x}^j$, and each column $i \in [n]$ consists of the m values sampled from D_i. Then, we draw n permutations $\pi_1, ..., \pi_n$ of $[m] = \{1, \dots, m\}$ independently and uniformly at random, and permute the m elements in column i by π_i. Regard each new row j as a new sample, denoted by $\tilde{\boldsymbol{x}}^j = (x_1^{\pi_1(j)}, x_2^{\pi_2(j)}, ..., x_n^{\pi_n(j)})$. Given $\pi_1, \dots, \pi_n$, the "permuted samples" $\{\tilde{\boldsymbol{x}}^1, \dots, \tilde{\boldsymbol{x}}^m\}$ have the same distributions as m i.i.d. random draws from $\boldsymbol{D}$.

For $h \in \mathcal{H}$, let $p_h = \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{D}}[h(\boldsymbol{x})]$. By the definition of (ε, δ)-uniform convergence (not on product distribution),

$$\Pr_{\mathcal{S},\pi}\left[\exists h \in \mathcal{H},\ \left|p_h - \frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right| \geq \varepsilon\right] \leq \delta. \tag{4.9}$$

For a set of fixed samples $\mathcal{S} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^m)$, recall that E_i is the uniform distribution over $\{x_i^1, \dots, x_i^m\}$, and $\boldsymbol{E} = \prod_{i=1}^n E_i$. We show that the expected value of h on $\boldsymbol{E}$ satisfies $\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{E}}[h(\boldsymbol{x})] = \mathbb{E}_{\pi}[\frac{1}{m}\sum_{j=1}^m h(\tilde{\boldsymbol{x}}^j)]$. This is because

$$\begin{aligned}
&\mathbb{E}_{\pi}\left[\frac{1}{m}\sum_{i=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right] = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\pi}\left[h(\tilde{\boldsymbol{x}}^j)\right] \\
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{(k_1,\dots,k_n)\in[m]^n} h(x_1^{k_1}, \dots, x_n^{k_n}) \cdot \Pr_{\pi}\left[\pi_1(j) = k_1, \dots, \pi_n(j) = k_n\right] \\
&= \frac{1}{m}\sum_{j=1}^{m}\sum_{(k_1,\dots,k_n)\in[m]^n} h(x_1^{k_1}, \dots, x_n^{k_n}) \cdot \frac{1}{m^n} \\
&= \frac{1}{m^n}\sum_{(k_1,\dots,k_n)\in[m]^n} h(x_1^{k_1}, \dots, x_n^{k_n}) \\
&= \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{E}}\left[h(\boldsymbol{x})\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\left|p_h - \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{E}}\left[h(\boldsymbol{x})\right]\right| &= \left|p_h - \mathbb{E}_{\pi}\left[\frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right]\right| \\
&\le \mathbb{E}_{\pi}\left[\left|p_h - \frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right|\right] \\
&\le \Pr_{\pi}\left[\left|p_h - \frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right| \ge \varepsilon\right] \cdot H \\
&\qquad + \left(1 - \Pr_{\pi}\left[\left|p_h - \frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right| \ge \varepsilon\right]\right) \cdot \varepsilon \\
&\le \Pr_{\pi}\left[\mathrm{Bad}(h, \pi, \mathcal{S})\right] \cdot H + \varepsilon,
\end{aligned}$$

where we define event

$$\mathrm{Bad}(h, \pi, \mathcal{S}) = \mathbb{I}\left[\left|p_h - \frac{1}{m}\sum_{j=1}^{m} h(\tilde{\boldsymbol{x}}^j)\right| \ge \varepsilon\right].$$

We note that, whenever $\left|p_h - \mathbb{E}_{\boldsymbol{v}\sim\boldsymbol{E}}[h(\boldsymbol{v})]\right| \ge 2\varepsilon$, we have $\Pr_{\pi}[\mathrm{Bad}(h, \pi, \mathcal{S})] \ge \frac{\varepsilon}{H}$.

Finally, consider the random draw of samples $\mathcal{S} \sim \boldsymbol{D}$,

$$\begin{aligned}
&\Pr_{\mathcal{S}}\left[\exists h \in \mathcal{H},\ \left|p_h - \mathbb{E}_{\boldsymbol{v}\sim\boldsymbol{E}}\left[h(\boldsymbol{v})\right]\right| \ge 2\varepsilon\right] \\
&\le \Pr_{\mathcal{S}}\left[\exists h \in \mathcal{H},\ \Pr_{\pi}\left[\mathrm{Bad}(h, \pi, \mathcal{S})\right] \ge \frac{\varepsilon}{H}\right] \\
&\le \Pr_{\mathcal{S}}\left[\Pr_{\pi}\left[\exists h \in \mathcal{H},\ \mathrm{Bad}(h, \pi, \mathcal{S}) \text{ holds}\right] \ge \frac{\varepsilon}{H}\right] \\
&\le \frac{H}{\varepsilon}\,\mathbb{E}_{\mathcal{S}}\left[\Pr_{\pi}\left[\exists h \in \mathcal{H},\ \mathrm{Bad}(h, \pi, \mathcal{S}) \text{ holds}\right]\right] && \text{by Markov inequality} \\
&= \frac{H}{\varepsilon}\Pr_{\mathcal{S},\pi}\left[\exists h \in \mathcal{H},\ \mathrm{Bad}(h, \pi, \mathcal{S}) \text{ holds}\right] \\
&\le \frac{H\delta}{\varepsilon} && \text{by (4.9).}
\end{aligned}$$

□

Combining Theorem 4.1 with Lemma 4.2, we derive a result of utility estimation by empirical product distribution.

Definition 4.9. *The* empirical product distribution estimator Empp *estimates interim utilities of a bidding strategy on the empirical product distribution* $\boldsymbol{E} = \prod_{i=1}^{n} E_i$*. Formally, for bidder* i *with value* v_i*, for bidding strategy profile* $\boldsymbol{\sigma}$*,*

$$\mathrm{Empp}(\mathcal{S}, i, v_i, \boldsymbol{\sigma}) \coloneqq \mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{E}_{-i}} \left[U_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) \right]. \tag{4.10}$$

Theorem 4.4 (Utility estimation by empirical product distribution)**.** *Suppose* $T_i \subseteq [0, H]$*. Let* $\boldsymbol{D}$ *be a product distribution on* $\prod_{i=1}^{n} T_i$*. For any* $\varepsilon > 0, \delta \in (0, 1)$*, there is*

$$M = O\left(\frac{H^2}{\varepsilon^2} \left[n \log n \log\left(\frac{H}{\varepsilon} \right) + \log\left(\frac{n}{\delta} \right) \right] \right), \tag{4.11}$$

such that for any $m \geq M$*, the empirical product distribution estimator* Empp (ε, δ)*-estimates with* m *samples the utilities over the set of all monotone bidding strategies.*

4.4.2 Learning Equilibrium from Samples

We are now ready to present our results for learning equilibria (BCE and BNE) using samples from the value distribution $\boldsymbol{D}$. By Theorem 4.4, utilities of the bidders on $\boldsymbol{D}$ can be approximated by the utilities on the empirical product distribution $\boldsymbol{E}$, therefore the auctions on the two distributions share the same set of approximate equilibria:

Theorem 4.5. *Suppose* $T_i \subseteq [0, H]$ *and* $\boldsymbol{D}$ *is a product distribution on* $\prod_{i=1}^{n} T_i$*. For any* $\varepsilon, \varepsilon' > 0, \delta \in (0, 1)$*, by drawing* $m \geq$ (4.11) *samples from* $\boldsymbol{D}$*, with probability at least* $1 - \delta$*, we have: Any monotone* ε'*-BCE* Q *on the empirical product distribution* $\boldsymbol{E} = \prod_{i=1}^{n} E_i$ *is a monotone* $(\varepsilon' + 2\varepsilon)$*-BCE on* $\boldsymbol{D}$*. Conversely, any monotone* ε'*-BCE* Q *on* $\boldsymbol{D}$ *is a monotone* $(\varepsilon' + 2\varepsilon)$*-BCE on* $\boldsymbol{E}$*. Formally:*

$$\mathrm{BCE}(\boldsymbol{E}, \varepsilon') \subseteq \mathrm{BCE}(\boldsymbol{D}, \varepsilon' + 2\varepsilon), \quad \mathrm{BCE}(\boldsymbol{D}, \varepsilon') \subseteq \mathrm{BCE}(\boldsymbol{E}, \varepsilon' + 2\varepsilon).$$

Proof. We will prove $\mathrm{BCE}(\boldsymbol{E}, \varepsilon') \subseteq \mathrm{BCE}(\boldsymbol{D}, \varepsilon' + 2\varepsilon)$. The other direction is analogous. According to Theorem 4.4, for any bidder i with value v_i and bid b_i, for any monotone strategies $\boldsymbol{\sigma}_{-i}$ of other bidders, bidder i's interim utilities on distributions $\boldsymbol{D}$ and $\boldsymbol{E}$ satisfy:

$$\left| \mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{D}_{-i}} \Big[U_i(v_i, b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) \Big] - \mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{E}_{-i}} \Big[U_i(v_i, b_i, \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) \Big] \right| \ \leq \ \varepsilon. \tag{4.12}$$

Let Q be any monotone ε'-BNE Q on $\boldsymbol{E}$. By Definition 4.2, bidder i's utility gain by deviating according to deviation function ϕ_i satisfies

$$\mathbb{E}_{\boldsymbol{\sigma} \sim Q} \left[\mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{E}_{-i}} \Big[U_i(v_i, \phi_i(\sigma_i, v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) - U_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) \Big] \right] \ \leq \ \varepsilon'.$$

Applying (4.12) to above, we obtain

$$\mathbb{E}_{\boldsymbol{\sigma} \sim Q} \left[\mathbb{E}_{\boldsymbol{v}_{-i} \sim \boldsymbol{D}_{-i}} \Big[U_i(v_i, \phi_i(\sigma_i, v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) - U_i(v_i, \sigma_i(v_i), \boldsymbol{\sigma}_{-i}(\boldsymbol{v}_{-i})) \Big] \right] \ \leq \ \varepsilon' + 2\varepsilon,$$

which implies that Q is an $(\varepsilon' + 2\varepsilon)$-BCE on $\boldsymbol{D}$. □

An implication of Theorem 4.5 is the following: if a mediator wants to coordinate the bidders in a non-truthful auction (such as first-price auction) but does not know the bidders' value distribution, the mediator can still achieve that by computing an approximate correlated equilibrium for the bidders using samples from the distribution. Theorem 4.5 characterizes the number of samples needed.

A similar conclusion also holds for BNEs:

Theorem 4.6. *Under the same condition as Theorem 4.5,*

$$\mathrm{BNE}(\boldsymbol{E}, \varepsilon') \subseteq \mathrm{BNE}(\boldsymbol{D}, \varepsilon' + 2\varepsilon), \quad \mathrm{BNE}(\boldsymbol{D}, \varepsilon') \subseteq \mathrm{BNE}(\boldsymbol{E}, \varepsilon' + 2\varepsilon).$$

Given some recent progress on the computation of BNE in first-price auctions on *discrete* distributions [WSZ20, CP23], we present an interesting corollary of Theorem 4.6: if there exists an algorithm that can compute BNE for a first-price auction on discrete value distributions, then there also exists an algorithm that can compute approximate BNE on *any* distributions, by simply sampling from the distribution and running the former algorithm on the empirical product distribution (which is discrete).

Corollary 4.1. *If there exists an algorithm that computes monotone ε'-BNE for the first-price auction on any discrete product value distributions $\boldsymbol{D} = \prod_{i=1}^{n} D_i$, then there exists a sample-access algorithm that computes $(\varepsilon' + \varepsilon)$-BNE for the first-price auction on any product distributions $\boldsymbol{F} = \prod_{i=1}^{n} F_i$ with high probability. If the running time of the former algorithm is polynomial in $\frac{1}{\varepsilon'}$ and the support size of each discrete D_i, then the running time of the latter algorithm is $\mathrm{poly}(\frac{1}{\varepsilon}, \frac{1}{\varepsilon'})$, which does not depend on the support size of F_i (and F_i can be continuous).*

4.5 Discussion

In this work, we obtained the first sample complexity result for learning strategic form Bayesian correlated equilibria in non-truthful auctions such as first-price and all-pay auctions. En route, we showed that bidders' expected utilities can be estimated using a moderate amount of value samples for all monotone bidding strategies. Our work can be a starting point for several future research directions:

- **Other types of BCE.** The learnability of strategic form BCE in non-truthful auctions relies on its simple form: recommending monotone joint bidding strategies to bidders without eliciting bidders' values. Other types of BCEs, such as a communication equilibrium which includes an elicitation phase and a recommendation phase [Mye82, For06], need not have a monotone structure, and we do not know whether they are efficiently learnable.

- **Correlated value distribution.** Our results also depend on bidders' values being drawn independently. With correlated values, the conditional distribution of opponents' values changes with a bidder's own value, and any naïve utility estimation algorithm needs a number of samples that grows linearly with the size of a bidder's value space. It is interesting whether there are meaningful tractable middle grounds between independent distributions and arbitrary correlated distributions.

- **Multi-item auctions and general games.** Monotonicity is a natural assumption on bidding strategies in a single-item auction, but it does not generalize to multi-parameter settings, where equilibria are difficult to characterize. It is an interesting question whether our results can be generalized to multi-item auctions, such as simultaneous first-price auctions, via more general, lossless structural assumptions on the bidding strategies. One can ask an even more general question: when can BCE be learned from type samples in general incomplete-information games?

4.6 Omitted Proofs in this Chapter

4.6.1 Lower Bound: Proof of Theorem 4.3

Fixing $\varepsilon > 0$, fixing $c_1 = 2000$, we first define two value distributions. Let D^+ be a distribution supported on $\{0, 1\}$, and for $v \sim D^+$, $\Pr[v = 0] = 1 - \frac{1+c_1\varepsilon}{n}$, and $\Pr[v = 1] = \frac{1+c_1\varepsilon}{n}$. Similarly define D^-: for $v \sim D^-$, $\Pr[v = 0] = 1 - \frac{1-c_1\varepsilon}{n}$, and $\Pr[v = 1] = \frac{1-c_1\varepsilon}{n}$.

Let $\mathrm{KL}(D^+; D^-)$ denote the KL-divergence between the two distributions.

Claim 4.1. $\mathrm{KL}(D^+; D^-) = O(\frac{\varepsilon^2}{n})$.

Proof. By definition,

$$\begin{aligned}
\mathrm{KL}(D^+; D^-) &= \frac{1+c_1\varepsilon}{n}\ln\left(\frac{1+c_1\varepsilon}{1-c_1\varepsilon}\right) + \frac{n-1-c_1\varepsilon}{n}\ln\left(\frac{n-1-c_1\varepsilon}{n-1+c_1\varepsilon}\right) \\
&= \frac{1}{n}\ln\left(\frac{1+c_1\varepsilon}{1-c_1\varepsilon}\cdot\frac{(1-\frac{c_1\varepsilon}{n-1})^{n-1}}{(1+\frac{c_1\varepsilon}{n-1})^{n-1}}\right) + \frac{c_1\varepsilon}{n}\ln\left(\frac{1+c_1\varepsilon}{1-c_1\varepsilon}\cdot\frac{1+\frac{c_1\varepsilon}{n-1}}{1-\frac{c_1\varepsilon}{n-1}}\right) \\
&\le \frac{1}{n}\ln\left(\frac{1+c_1\varepsilon}{1-c_1\varepsilon}\cdot\frac{\left(1-\frac{c_1\varepsilon}{n-1}\right)^{n-1}}{1+c_1\varepsilon}\right) + \frac{2c_1\varepsilon}{n}\ln\left(1+\frac{2c_1\varepsilon}{1-c_1\varepsilon}\right) \\
&\le \frac{1}{n}\ln\left(\frac{1-c_1\varepsilon+\frac{1}{2}(c_1\varepsilon)^2}{1-c_1\varepsilon}\right) + \frac{8c_1^2\varepsilon^2}{n} \\
&\le \frac{10c_1^2\varepsilon^2}{n}.
\end{aligned}$$

In the last two inequalities we used $c_1\varepsilon < \frac{1}{2}$ and $\ln(1+x) \le 1+x$ for all $x > 0$. □

It is well known that an upper bound on KL-divergence implies an information-theoretic lower bound on the number of samples to distinguish two distributions (e.g., [Man11]).

Corollary 4.2. *Given t i.i.d. samples from D^+ or D^-, if $t \le \frac{n}{80c_1^2\varepsilon^2}$, no algorithm $\mathcal{H}$ that maps samples to $\{D^+, D^-\}$ can do the following: when the samples are from D^+, $\mathcal{H}$ outputs D^+ with probability at least $\frac{2}{3}$, and if the samples are from D^-, $\mathcal{H}$ outputs D^- with probability at least $\frac{2}{3}$.*

We now construct product distributions using D^+ and D^-. For any $S \subseteq [n-1]$, define product distribution $\boldsymbol{D}_S$ to be $\prod_i D_i$ where $D_i = D^+$ if $i \in S$, and $D_i = D^-$ if $i \in [n-1] \setminus S$, and F_n is a point mass on value 1. For any $j \in [n-1]$ and $S \subseteq [n-1]$, distinguishing $\boldsymbol{D}_{S\cup\{j\}}$ and $\boldsymbol{D}_{S\setminus\{j\}}$ by samples from the product distribution is no easier than distinguishing D^+ and D^-, because the coordinates of the samples not from D_j contains no information about D_j.

Corollary 4.3. *For any $j \in [n-1]$ and $S \subseteq [n-1]$, given t i.i.d. samples from $\boldsymbol{D}_{S\cup\{j\}}$ or $\boldsymbol{D}_{S\setminus\{j\}}$, if $t \le \frac{n}{80c_1^2\varepsilon^2}$, no algorithm $\mathcal{H}$ can do the following: when the samples are from $\boldsymbol{D}_{S\cup\{j\}}$, $\mathcal{H}$ outputs $\boldsymbol{D}_{S\cup\{j\}}$ with probability at least $\frac{2}{3}$, and when the samples are from $\boldsymbol{D}_{S\setminus\{j\}}$, $\mathcal{H}$ outputs $\boldsymbol{D}_{S\setminus\{j\}}$ with probability at least $\frac{2}{3}$.*

We now use Corollary 4.3 to derive an information-theoretic lower bound on estimating utilities for monotone bidding strategies, for distributions in $\{\boldsymbol{D}_S\}_{S\subseteq[n]}$.

Proof of Theorem 4.3. Without loss of generality, assume n is odd. Let S be an arbitrary subset of $[n-1]$ of size either $\lfloor n/2 \rfloor$ or $\lceil n/2 \rceil$. We focus on the interim utility of bidder n with value 1 and bidding $\frac{1}{2}$. Denote this bidding strategy by σ_n. The other bidders may adopt one of two bidding strategies. One of them is σ^+: $\sigma^+(0) = 0$ and $\sigma^+(1) = \frac{1}{2} + \eta$ for sufficiently small $\eta > 0$. The other bidding strategy $\sigma^-(\cdot)$ maps all values to 0. For $T \subseteq [n-1]$, let $\boldsymbol{\sigma}_T$ be the profile of bidding strategies where $\sigma_i = \sigma^+$ for bidder $i \in T$, and $\sigma_i = \sigma^-$ for bidder $i \in [n-1] \setminus T$.

For the distribution $\boldsymbol{D}_S$, the interim utility of bidder n is

$$\begin{aligned} u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_T\right) &= \frac{1}{2} \Pr\left[\max_{i\in T} v_i = 0\right] \\ &= \frac{1}{2}\left(1 - \frac{1+c_1\varepsilon}{n}\right)^{|S\cap T|}\left(1 - \frac{1-c_1\varepsilon}{n}\right)^{|T\setminus S|} \\ &= \frac{1}{2}\left(1 - \frac{1+c_1\varepsilon}{n}\right)^{|T|}\left(\frac{n-1+c_1\varepsilon}{n-1-c_1\varepsilon}\right)^{|T\setminus S|}. \end{aligned}$$

Therefore, for $T, T' \subseteq [n-1]$ with $|T| = |T'|$,

$$\begin{aligned} \frac{u_n(1, \frac{1}{2}, \boldsymbol{\sigma}_T)}{u_n(1, \frac{1}{2}, \boldsymbol{\sigma}_{T'})} &= \left(1 + \frac{2c_1\varepsilon/(n-1)}{1 - \frac{c_1\varepsilon}{n-1}}\right)^{|T\setminus S| - |T'\setminus S|} \\ &\ge 1 + \frac{2c_1\varepsilon}{n-1} \cdot (|T \setminus S| - |T' \setminus S|); \end{aligned}$$

Suppose $|T \setminus S| \geq |T' \setminus S|$ and $|T| = |T'| \geq \lfloor \frac{n}{2} \rfloor$, then

$$u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_T\right) - u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_{T'}\right) \geq (|T \setminus S| - |T' \setminus S|) \cdot \frac{2c_1\varepsilon}{n-1} \cdot u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_{T'}\right)$$
$$\geq (|T \setminus S| - |T' \setminus S|) \cdot \frac{2c_1\varepsilon}{n-1} \cdot \frac{1}{8e^2}, \tag{4.13}$$

where the last inequality is because $u_n(1, \frac{1}{2}, \boldsymbol{\sigma}_{T'}) \geq \frac{1}{2}(1-\frac{2}{n})^n = \frac{1}{2}[(1-\frac{2}{n})^{\frac{n}{2}}]^2 \geq \frac{1}{2}(\frac{1}{2e})^2 = \frac{1}{8e^2}$.

Now suppose an algorithm $\mathcal{A}$ (ε, δ)-estimates the utilities of all monotone bidding strategies with $t \leq \frac{n}{80c_1^2\varepsilon^2}$ samples $\mathcal{S}$. Define $\mathcal{H} : \mathbb{R}_+^{n \times t} \times \mathbb{N} \to 2^{[n-1]}$ be a function that outputs, among all $T \subseteq [n-1]$ of size k, the one that maximizes bidder n's utility when other bidders bid according to strategy $\boldsymbol{\sigma}_T$. Formally,

$$\mathcal{H}(\mathcal{S}, k) = \underset{T \subseteq [n-1], |T| = k}{\arg\max} \mathcal{A}\left(\mathcal{S}, n, 1, (\boldsymbol{\sigma}_T, \sigma_n)\right),$$

By Definition 4.3, for any S with $|S| = \lfloor n/2 \rfloor$, for samples drawn from $\boldsymbol{D}_S$, with probability at least $1 - \delta$,

$$\mathcal{A}(\mathcal{S}, n, 1, (\boldsymbol{\sigma}_{[n-1] \setminus S}, \sigma_n)) \geq u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_{[n-1] \setminus S}\right) - \varepsilon;$$

and for any $T \subseteq [n-1]$ with $|T| = \lceil n/2 \rceil$,

$$\mathcal{A}(\mathcal{S}, n, 1, (\boldsymbol{\sigma}_T, \sigma_n)) \leq u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_T\right) + \varepsilon.$$

Therefore, for $W = \mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$,

$$u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_W\right) \geq u_n\left(1, \frac{1}{2}, \boldsymbol{\sigma}_{[n-1] \setminus S}\right) - 2\varepsilon.$$

Since $|W| = [n-1] \setminus S = \lceil n/2 \rceil$, by (4.13),

$$\Big(\lceil \frac{n}{2} \rceil - |W \setminus S|\Big) \cdot \frac{c_1 \varepsilon}{(n-1)4e^2} \leq 2\varepsilon.$$

So

$$|W \cap S| \leq (n-1) \cdot \frac{8e^2}{c_1}.$$

In other words, with probability at least $1-\delta$, $\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$ is the complement of S except for at most $\frac{8e^2}{c_1}$ fraction of the coordinates in $[n-1]$.

Similarly, for S of cardinality $\lceil n/2 \rceil$,

$$|\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil) \cap S| \leq (n-1) \cdot \frac{8e^2}{c_1} + 1.$$

Take c_2 to be $\frac{8e^2}{c_1}$. We have $c_2 < \frac{1}{20}$. For all large enough n and all S of size $\lfloor n/2 \rfloor$ or $\lceil n/2 \rceil$, with probability at least $1-\delta$, $\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$ correctly outputs the elements not in S with an exception of at most c_2 fraction of coordinates.

Let $\mathbf{S}$ be the set of all subsets of $[n-1]$ of size either $\lceil n/2 \rceil$ or $\lfloor n/2 \rfloor$. Consider any $S \in \mathbf{S}$. Let $\theta(S) \subseteq [n-1]$ denote the set of coordinates whose memberships in S are correctly predicted by $\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$ with probability at least 2/3; that is, $i \in \theta(S)$ iff with probability at least 2/3, $\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$ is correct about whether $i \in S$. Let the cardinality of $|\theta(S)|$ be $z(n-1)$. Suppose we draw coordinate i uniformly at random from $[n-1]$, and independently draw t samples $\mathcal{S}$ from $\boldsymbol{D}_S$, then the probability that $\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil)$ is correct about whether $i \in S$ satisfies:

$$\begin{aligned}\Pr_{i,\mathcal{S}}\Big[\mathcal{H}(\mathcal{S}, \lceil n/2 \rceil) \text{ is correct about whether } i \in S\Big] &\geq (1-c_2)(1-\delta) \\ &\geq 0.9,\end{aligned}$$

and

$$\Pr_{i,\mathcal{S}}\left[\mathcal{H}(\mathcal{S},\lceil n/2\rceil) \text{ is correct about whether } i \in S\right] \le \Pr_i\left[i \in \theta(S)\right]\cdot 1 + \Pr_i\left[i \notin \theta(S)\right]\cdot \frac{2}{3}$$
$$= z\cdot 1 + (1-z)\cdot \frac{2}{3},$$

which implies $z > 0.6$. If a pair of sets S and S' differ in only one coordinate i, and $i \in \theta(S)\cap\theta(S')$, then $\mathcal{H}(\cdot)$ serves as an algorithm that tells apart $\boldsymbol{D}_S$ and $\boldsymbol{D}_{S'}$, contradicting Corollary 4.3. We now show, with a counting argument, that such a pair of S and S' must exist.

Since for each $S \in \mathbf{S}$, $|\theta(S)| \ge 0.6(n-1)$, there exists a coordinate $i \in [n-1]$ and $\mathcal{T} \subseteq \mathbf{S}$, with $|\mathcal{T}| \ge 0.6|\mathbf{S}|$, such that for each $S \in \mathcal{T}$, $i \in \theta(S)$. But $\mathbf{S}$ can be decomposed into $|\mathbf{S}|/2$ pairs of sets, such that within each pair, the two sets differ by one in size, and precisely one of them contains coordinate i. Therefore among these pairs there must exist one (S, S') with $S, S' \in \mathcal{T}$, i.e., $i \in \theta(S)$ and $i \in \theta(S')$. Using $\mathcal{H}$, which is induced by $\mathcal{A}$, we can tell apart $\boldsymbol{D}_S$ and $\boldsymbol{D}_{S'}$ with probability at least $2/3$, which is a contradiction to Corollary 4.3. This completes the proof of Theorem 4.3. □

Part II

Incentive Design for Learning Agents

Chapter 5

General Principal-Agent Problems with a Learning Agent

joint work with Yiling Chen [LC25]

This part of my dissertation focuses on incentive design for learning agents. This chapter studies a general class of principal-agent problems with a single learning agent, while the next chapter will study a specific problem but with multiple learning agents.

5.1 Introduction

Classic economic models of principal-agent interactions, including auction design, contract design, and Bayesian persuasion, often assume that the agent is able to best respond to the strategy committed by the principal. For example, in Bayesian persuasion, the agent (receiver) needs to compute the posterior belief about the state of the world after receiving some information from the principal (sender) and take an optimal action based on the posterior belief; this requires the receiver accurately knowing the prior of the state as well as the signaling scheme used by the sender. In contract design, where a principal specifies an outcome-dependent payment scheme to incentivize the agent to take certain actions, the agent has to know the action-dependent outcome distribution in order to best respond to the contract. Requiring strong rationality assumptions, the best-responding behavior is often observed to be violated in practice [Cam98, Ben19].

In this work, using Bayesian persuasion as the main example, we study general principal-agent problems under an alternative behavioral model for the agent: *learning*. The use of learning as a behavioral model dates back to early economic literature on learning in games [Bro51, FL98] and has been actively studied by computer scientists and operations researchers in recent years [NST15, BMSW18, DSS19, MMSS22, CWWZ24, LLZW23, RZ24, GKS+24, SCB+24, D'A23]. A learning agent no longer has perfect knowledge of the parameter of the game or the principal's strategy. Instead of best responding, which is no longer possible or well-defined, the agent chooses his action based on past interactions with the principal. We focus on *no-regret* learning, which requires the agent to not suffer a large average regret at the end of repeated interactions with the principal, for not taking the optimal action at hindsight. This is a mild requirement satisfied by many natural learning algorithms (e.g., ε-greedy, MWU, UCB, EXP-3) and can reasonably serve as a possible behavioral assumption for real-world agents.

With a learning agent, can the principal achieve a better outcome than that in the classic model with a best-responding agent? Previous works on playing against learning agents [DSS19, GKS+24] showed that, in Stackelberg games and contract design, the leader/principal can obtain utility $U^* - o(1)$ against a no-regret learning follower/agent, where U^* is the Stackelberg value, defined to be the principal's optimal utility in the classic model with a best-responding agent. On the other hand, if the agent does a stronger version of no-regret learning, called no-swap-regret learning [HMC00, BM07], then the principal cannot obtain utility more than the Stackelberg value $U^* + o(1)$. Interestingly, the conclusion that no-swap-regret learning can cap the principal's utility at $U^* + o(1)$ does not hold when the agent has private information, such as in auctions [BMSW18] and Bayesian Stackelberg games [MMSS22]: the principal can sometimes exploit a no-swap-regret learning agent with private information to do much better than U^* in those games.

Three natural questions then arise: (1) What is the largest class of principal-agent problems under which the agent's no-swap-regret learning can cap the principal's utility at the Stackelberg value $U^* + o(1)$? (2) In cases where the principal's optimal utility against a learning agent is bounded by $[U^* - o(1), U^* + o(1)]$, what is the exact magnitude of the $o(1)$ terms? (3) Instead of analyzing games like Stackelberg games and contract design separately, can we analyze all principal-agent problems with learning agents in a unified way?

Our Contributions. Our work defines a general model of principal-agent problems with a learning agent, answering all questions (1) - (3). For (1), we show that the principal's utility is bounded around the Stackelberg value U^* in all generalized principal-agent problems where the agent does not have private information but the principal can be privately informed. In particular, this includes complete-information games like Stackelberg games and contract design, as well as Bayesian persuasion where the sender/principal privately observes the state of the world.

For (2) and (3), we provide a unified analytical framework to derive tight bounds on the principal's achievable utility against a no-regret or no-swap-regret learning agent in all generalized principal-agent problems where the agent does not have private information. Specifically, we explicitly characterize the $o(1)$ difference between the principal's utility and U^* in terms of the agent's regret.

Result 1 (from Theorems 5.1, 5.4, 5.5). *Against any no-regret learning agent with regret* $\mathrm{Reg}(T)$ *in* T *periods, the principal can obtain an average utility of at least* $U^* - O\big(\sqrt{\frac{\mathrm{Reg}(T)}{T}}\big)$. *The principal can do this using a fixed strategy in all* T *periods and only knowing the agent's regret bound but not the exact learning algorithm.*

Result 2 (from Proposition 5.2 and Example 5.2). *There exists a Bayesian persuasion instance where, for any strategy of the principal, there is a no-regret learning algorithm for the agent under which the principal's utility is at most* $U^* - \Omega\big(\sqrt{\frac{\mathrm{Reg}(T)}{T}}\big)$. *The same holds for no-swap-regret learning algorithms.*

Results 1 and 2 together characterize the best utility the principal can achieve against the *worst-case* learning algorithm of the agent: $U^* - \Theta\big(\sqrt{\frac{\mathrm{Reg}(T)}{T}}\big)$. Notably, this term has a squared root, instead of linear, dependency on the agent's average regret $\frac{\mathrm{Reg}(T)}{T}$.

We then consider the best utility the principal can achieve against the *best-case* learning algorithm of the agent. We show that, if the agent does no-swap-regret learning, then the principal cannot achieve more than $U^* + O\big(\frac{\mathrm{SReg}(T)}{T}\big)$, and this bound is tight:

Result 3 (from Theorems 5.2, 5.4, 5.5). *Against any no-swap-regret learning agent with swap-regret* $\mathrm{SReg}(T)$ *in* T *periods, the principal cannot obtain average utility larger than* $U^* + O\big(\frac{\mathrm{SReg}(T)}{T}\big)$. *This holds even if the principal knows the agent's learning algorithm and uses adaptive strategies.*

Result 4 (from Proposition 5.3 and Example 5.2). *There exists a Bayesian persuasion instance and a no-swap-regret learning algorithm for the agent, such that the principal can achieve average utility* $U^* + \Omega\big(\frac{\mathrm{SReg}(T)}{T}\big)$.

Interestingly, the squared root bound $U^* - \Theta\big(\sqrt{\frac{\mathrm{Reg}(T)}{T}}\big)$ from Results 1, 2 and the linear bound $U^* + \Theta\big(\frac{\mathrm{SReg}(T)}{T}\big)$ from Result 3, 4 are not symmetric. Since we show that these bounds are tight, this asymmetry is not because our analysis is tight. Instead, this asymmetric is intrinsic.

Results 1 to 4 characterize the range of utility achievable by the principal against a no-swap-regret learning agent: $[U^* - \Theta(\sqrt{\frac{\mathrm{SReg}(T)}{T}}), U^* + O(\frac{\mathrm{SReg}(T)}{T})]$. As $T \to \infty$, the range converges to U^*, which means that the agent's no-swap-regret learning behavior is essentially equivalent to best responding behavior. This justifies the classical economic notion that the equilibria of games are results of repeated interactions between learning players.

However, for no-regret but not necessarily no-swap-regret algorithms, the upper bound result $U^* + O(\frac{\mathrm{Reg}(T)}{T})$ does not hold. The repeated interaction between a principal and a no-regret learning agent does not always lead to the Stackelberg equilibrium outcome U^*:

Result 5 (Theorem 5.3). *There exists a Bayesian persuasion instance where, against a no-regret but not no-swap-regret learning agent (in particular, mean-based learning agent), the principal can do significantly better than the Stackelberg value U^*.*

In summary, our Results 1 to 4 exactly characterize the principal's optimal utility in principal-agent problems with a no-swap-regret agent, which not only refines previous works on playing against learning agents in specific games (e.g., Stackelberg games [DSS19] and contract design [GKS+24]) but also generalizes to all principal-agent problems where the agent does not have private information. In particular, when applied to Bayesian persuasion, our results imply that the sender cannot exploit a no-swap-regret learning receiver even if the sender possesses informational advantage over the receiver.

Some Intuitions. As alluded above, the main intuition behind our Results 1 to 4 is that the agent's learning behavior is closely related to *approximately best response.* A no-regret learning agent makes sub-optimal decisions with the sub-optimality measured by the regret. When the sub-optimality/regret is small, the principal-agent problem with a no-regret agent (or approximately best responding agent) converges to the problem with an exactly best responding agent. This explains why the principal against a no-regret learning agent can obtain a payoff that is close to the optimal payoff U^* against a best responding agent.

However, there are *two subtleties* behind the above intuition.

First, the intuition that a no-regret learning agent is approximately best responding is true only when the principal uses a *fixed* strategy throughout the interactions with the agent. If the principal uses *adaptive* strategies, then a no-regret agent is not necessarily approximately best responding to the average strategy of the principal across T periods, while a no-swap-regret agent is still approximately best responding. This is because a no-swap-regret algorithm ensures that, whenever the algorithm recommends some action a^t to the agent at a period t, it is almost optimal for the agent to take the recommended action a^t. But a no-regret algorithm only ensures the agent to not regret when comparing to taking any *fixed* action in all T periods. The agent could have done better by deviating to different actions given different recommendations from the algorithm. This means that a no-regret agent is not approximately best responding when the principal's strategy changes over time, which explains why the principal can exploit a no-regret agent sometimes (our Result 5).

Second, what is the reason for the asymmetry between the worst-case utility $U^* - \Theta(\sqrt{\frac{\mathrm{SReg}(T)}{T}})$ and the best-case utility $U^* + O(\frac{\mathrm{SReg}(T)}{T})$ that the principal can obtain against a no-swap-regret learning agent? Roughly speaking, a no-swap-regret learning agent is approximately best responding to the principal's average strategy over all T periods, with the degree of approximate best response measured by the average regret $\frac{\mathrm{SReg}(T)}{T} = \delta$. However, because no-swap-regret learning algorithms are randomized[1], they correspond to randomized approximately best responding strategies of the agent that are worse than the best responding strategy by a margin of δ *in expectation*. This means that the agent might take $\sqrt{\delta}$-sub-optimal actions with probability $\sqrt{\delta}$, which can cause a loss of 1 to the principal's utility with probability $\sqrt{\delta}$. So, the principal's expected utility can be decreased to $U^* - \sqrt{\delta} = U^* - \sqrt{\frac{\mathrm{SReg}(T)}{T}}$ in the worst case. On the other hand, when considering the principal's best-case utility, we care about the δ-approximately-best-responding strategy of the agent that maximizes the principal's utility. That strategy turns out to be equivalent to a deterministic strategy that gives the principal a utility of at most $U^* + O(\delta) = U^* + O(\frac{\mathrm{SReg}(T)}{T})$. This explains the asymmetry between the worst-case and best-case bounds.

Structure of this Chapter. We define our model of generalized principal-agent problems with a learning agent in Section 5.2. Since Bayesian persuasion is the main mo-

[1] It is well known that deterministic algorithms cannot satsify the no-regret property (see, e.g., **(author?)** [Rou16]).

tivation of our work, we also present the specific model of persuasion with a learning agent in Section 5.2.3. We develop our main results in Sections 5.3 and 5.4, by first reducing the generalized principal-agent problem with a learning agent to the problem with approximate best response, then characterizing the problem with approximate best response. Section 5.5 applies our general results to three specific principal-agent problems: Bayesian persuasion, Stackelberg games, and contract design. Section 5.6 offers additional discussions.

5.1.1 Related Works

Learning agents have been studied in principal-agent problems like auctions [BMSW18, CWWZ24, RZ24, KSS24], bimatrix Stackelberg games [DSS19, MMSS22, ACS24], contract design [GKS$^+$24, SCB$^+$24], and Bayesian persuasion [LLZW23, JP24]. These problems belong to the class of *generalized principal-agent problems* [Mye82, GHWX24]. We thus propose a general framework of generalized principal-agent problem with a learning agent, which encompasses several previous models, refines previous results, and provides new results.

The work by [CHJ20] also proposes a general framework of principal-agent problems with learning players, but has two key differences with ours: (1) They drop the common prior assumption while we still keep it. This assumption allows us to compare the principal's utility in the learning model with the classic model with common prior. (2) Their principal has commitment power, which is reasonable in, e.g., auction design, but less realistic in information design where the principal's strategy is a signaling scheme. Our principal does not commit.

For bimatrix Stackelberg games specifically, [DSS19] show that the follower's no-swap-regret learning can cap the leader's utility at $U^* + o(1)$. We find that this conclusion holds for all generalized principal-agent problems where the agent does not have private information. This conclusion does not hold when the agent is privately informed, as shown by [MMSS22] in Bayesian Stackelberg games. We view our work as characterizing the largest class of games under which this conclusion holds.

The literature on information design (Bayesian persuasion) has investigated various relaxations of the strong rationality assumptions in the classic models. For the sender, known prior (as we discussed and studied in Chapter 2) and known utility [KMZL17, BTCXZ21, CCMG20, FTX22, BSC$^+$24] are relaxed. For the receiver, the receiver may make mistakes in Bayesian updates [dCZ22], be risk-conscious [AIL23], do quantal response [FHT24] or approximate best response [YZ24]. Independently and concurrently of us, [JP24] also study Bayesian persuasion with a learning agent. Their work has a few differences with us: (1) Their model is a general Bayesian persuasion model with imperfect and non-stationary dynamics for the state of the world. Our model assumes a perfect and stationary environment, but generalizes Bayesian persuasion in another direction, namely, generalized principal-agent problems. (2) Their results are qualitatively similar to our Results 1 and 5, while our results are more quantitative and precise. (3) We additionally show that no-swap-regret learning can cap the sender's utility (Result 3).

As our problem reduces to generalized principal-agent problems with approximate best response, our work is also related to recent works on approximately-best-responding agents in Stackelberg games [GHWX23] and Bayesian persuasion [YZ24]. We focus on the range of payoff that can be obtained by a computationally-unbounded principal, ignoring the computational aspect considered by [GHWX23, YZ24]. Besides the "maxmin/robust" objective, we also study the "maxmax" objective where the agent approximately best responds *in favor of* the principal, which is usually not studied in the literature.

5.2 Generalized Principal-Agent Problem with a Learning Agent

This section defines our model, *generalized principal-agent problem with a learning agent.* This model includes Stackelberg games, contract design, and Bayesian persuasion with learning agents.

5.2.1 Generalized Principal-Agent Problem

Generalized principal-agent problem, proposed and developed by [Mye82, GHWX24], is a general model that includes auction design, contract design, Stackelberg games, and Bayesian persuasion. While [Mye82] and [GHWX24] allow the agent to have private information, our model assumes an agent with no private information. There are two players in a generalized principal-agent problem: a principal and an agent. The principal has a convex, compact decision space $\mathcal{X}$ and the agent has a finite action set A. The principal and the agent have utility functions $u, v : \mathcal{X} \times A \to \mathbb{R}$. We assume that $u(x, a)$, $v(x, a)$ are linear in $x \in \mathcal{X}$, which is satisfied by all the examples of generalized principal-agent problems we will consider (Bayesian persuasion, Stackelberg games, contract design). There is a signal/message set S. Signals are usually interpreted as recommendations of actions for the agent, where $S = A$, but we allow any signal set of size $|S| \geq |A|$. A strategy of the principal is a distribution $\pi \in \Delta(\mathcal{X} \times S)$ over pairs of decision and signal. When the utility functions u, v are linear, it is without loss of generality to assume that the principal does not randomize over multiple decisions for one signal [GHWX24], namely, the principal chooses a distribution over signals and a unique decision x_s associated with each signal $s \in S$. So, we can write a principal strategy as $\pi = \{(\pi_s, x_s)\}_{s \in S}$ where $\pi_s \geq 0$ is the probability of signal $s \in S$, $\sum_{s \in S} \pi_s = 1$, and $x_s \in \mathcal{X}$. There are two variants of generalized principal-agent problems:

- *Unconstrained* [Mye82]: there is no restriction on the principal's strategy π.

- *Constrained* [GHWX24]: the principal's strategy π has to satisfy a constraint in the form of $\sum_{s\in S}\pi_s x_s \in \mathcal{C}$ where $\mathcal{C} \subseteq \mathcal{X}$ is some convex set.

Unconstrained generalized principal-agent problems include contract design and Stackelberg games. Constrained generalized principal-agent problems include Bayesian persuasion (see Section 5.2.3).

In a one-shot generalized principal-agent problem where the principal has commitment power, the principal first commits to a strategy $\pi = \{(\pi_s, x_s)\}_{s\in S}$, then nature draws a signal $s \in S$ according to the distribution $\{\pi_s\}_{s\in S}$ and sends s to the agent (note: due to the commitment assumption, this is equivalent to revealing the pair (s, x_s) to the agent), then the agent takes an action $a_s \in \arg\max_{a\in A} v(x_s, a)$ that maximizes its utility (breaking ties in favor of the principal), and the principal obtains utility $u(x_s, a_s)$. The principal aims to maximize its expected utility $\mathbb{E}_{s\sim\pi}[u(x_s, a_s)]$ by choosing the strategy π. Denote the maximal expected utility that the principal can obtain by U^*:

$$U^* = \max_{\pi} \sum_{s\in S} \pi_s \max_{a_s \in \arg\max_{a\in A} v(x_s, a)} u(x_s, a_s). \tag{5.1}$$

U^* is called the Stackelberg value in the literature.

5.2.2 Learning Agent

Now we define the model of generalized principal-agent problem with a learning agent. The game is repeated for T rounds. Unlike the static model above, the principal now does not commit to its strategy π^t every round. The agent does not know the principal's strategy π^t or decision x^t at each round. Instead, the agent uses some adaptive algorithm to learn from history which action to take in response to each possible signal. We allow the agent's strategy to be randomized.

Generalized Principal-Agent Problem with a Learning Agent
In each round $t = 1, \ldots, T$:

(1) Using some algorithm that learns from history (including signals, actions, and utility feedback in the past, described in details later), the agent chooses a strategy $\rho^t : S \to \Delta(A)$ that maps each possible signal $s \in S$ to a distribution over actions $\rho^t(s) \in \Delta(A)$.

(2) The principal chooses a strategy $\pi^t = \{(\pi^t_s, x^t_s)\}_{s \in S}$, which is a distribution over signals S and a decision $x^t_s \in \mathcal{X}$ associated with each signal.

(3) Nature draws signal $s^t \sim \pi^t$ and reveals it. The principal makes decision $x^t = x^t_{s^t}$. The agent draws action $a^t \sim \rho^t(s^t)$.

(4) The principal and the agent obtain utility $u^t = u(x^t, a^t)$ and $v^t = v(x^t, a^t)$. The agent observes some feedback (e.g., $v^t(x^t, a^t)$ or x^t).

We assume that the principal knows the utility functions u and v of both players, and has some knowledge about the agent's learning algorithm (which will be specified later). The principal's goal is to maximize the expected average utility $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big]$.

Compared with the static model in Section 5.2.1 where the principal moves before the agent, we flip the decision-making order of the principal and the agent in the learning model: the agent moves first by choosing ρ^t, then the principal chooses π^t. This gives the principal an opportunity to "exploit" the agent by choosing a π^t that best responds to ρ^t, hence potentially do much better than the Stackelberg value U^* where the principal moves first. However, one of our main results (Result 2 in the Introduction) will show that the principal cannot do much better than U^* if the agent uses a particular type of learning algorithm, called contextual no-swap-regret algorithm, which we define below.

Agent's learning problem. The agent's learning problem can be regarded as a *contextual multi-armed bandit problem* [TDM10] where A is the set of arms, and a signal $s^t \in S$ serves as a context that affects the utility of each arm $a \in A$. The agent picks an arm to pull based on the current context s^t and the historical information about each arm under different contexts, adjusting its strategy over time based on the feedback collected after each round.

What feedback can the agent observe after each round? One may assume that the agent sees the principal's decision x^t after each round (this is call *full-information* feedback in the multi-armed bandit literature), or the utility $v^t = v(x^t, a^t)$ obtained in that round (this is called *bandit feedback*) but not the x^t, or some unbiased estimate of $v(x^t, a^t)$. We do not make specific assumptions on the feedback. All we need is that the feedback is sufficient for the agent to achieve contextual no-regret or contextual no-swap-regret, which are defined below:

Definition 5.1. *The agent's learning algorithm is said to satisfy:*

- contextual no-regret *if: there is a function* $\mathrm{CReg}(T) = o(T)$ *such that for any deviation function* $d : S \to A$*, the regret of the agent not deviating according to* d *is at most* $\mathrm{CReg}(T)$*:*[a]

$$\mathbb{E}\Big[\sum_{t=1}^{T} \big(v(x^t, d(s^t)) - v(x^t, a^t)\big)\Big] \le \mathrm{CReg}(T).$$

- contextual no-swap-regret *if: there is a function* $\mathrm{CSReg}(T) = o(T)$ *such that for any deviation function* $d : S \times A \to A$*, the regret of the receiver not deviating according to* d *is at most* $\mathrm{CSReg}(T)$*:*

$$\mathbb{E}\Big[\sum_{t=1}^{T} \big(v(x^t, d(s^t, a^t)) - v(x^t, a^t)\big)\Big] \le \mathrm{CSReg}(T).$$

We call CReg(T) *and* CSReg(T) *the* contextual regret *and* contextual swap-regret *of the agent.*

[a] A function $f(T) = o(T)$ means $\frac{f(T)}{T} \to 0$ as $T \to +\infty$. So, the average regret $\frac{\text{CReg}(T)}{T} \to 0$.

Contextual no-regret is implied by contextual no-swap-regret because the latter has a larger set of deviation functions. Contextual no-(swap-)regret algorithms are known to exist under bandit feedback. In fact, they can be easily constructed by running an ordinary no-(swap-)regret algorithm for each context independently. Formally:

Proposition 5.1. *There exist learning algorithms with contextual regret* $\text{CReg}(T) = O(\sqrt{|A||S|T})$ *and contextual swap-regret* $\text{CSReg}(T) = O(|A|\sqrt{|S|T})$. *They can be constructed by running an ordinary no-(swap-)regret multi-armed bandit algorithm for each context independently.*

See Section 5.7.1 for a proof of this Proposition.

5.2.3 Special Case: Bayesian Persuasion with a Learning Agent

We show that *Bayesian persuasion* [KG11] is a special case of constrained generalized principal-agent problems. We will also show that Bayesian persuasion is in fact equivalent to *cheap talk* [CS82] under our learning agent model.

Bayesian persuasion as a generalized principal-agent problem. There are two players in Bayesian persuasion: a sender (principal) and a receiver (agent). There are a finite set Ω of states of the world, a signal set S, an action set A, a prior distribution $\mu_0 \in \Delta(\Omega)$ over the states, and utility functions $u, v : \Omega \times A \to \mathbb{R}$ for the sender and the receiver. When the state is $\omega \in \Omega$ and the receiver takes action $a \in A$, the sender and the receiver obtain utility $u(\omega, a)$, $v(\omega, a)$, respectively. Both players know μ_0, but only the sender has access to the realized state $\omega \sim \mu_0$. The sender commits to some signaling scheme $\pi : \Omega \to \Delta(S)$, mapping any state to a probability distribution over signals, to partially reveal information about the state w to the receiver. In the classic model, after receiving a signal $s \in S$, the receiver will form the posterior belief $\mu_s \in \Delta(\Omega)$ about the state: $\mu_s(\omega) = \frac{\mu_0(\omega)\pi(s|\omega)}{\pi_s}$, where $\pi_s = \sum_{\omega \in \Omega} \mu_0(\omega)\pi(s|\omega)$ is the total probability that signal s is sent, and take an optimal action with respect to μ_s, i.e., $a_s \in \arg\max_{a \in A} \sum_{\omega \in \Omega} \mu_s(\omega)v(\omega, a)$. The sender aims to find a signaling scheme to maximizde its expected utility $\mathbb{E}[u(\omega, a_s)]$.

It is well-known [KG11] that a signaling scheme $\pi : \Omega \to \Delta(S)$ decomposes the prior μ_0 into a distribution over posteriors whose average is equal to the prior μ_0:

$$\sum_{s \in S} \pi_s \mu_s = \mu_0 \in \{\mu_0\} =: \mathcal{C}, \quad \sum_{s \in S} \pi_s = 1. \tag{5.2}$$

Equation (5.2) is called the *Bayes plausibility* condition. Conversely, any distribution over posteriors $\{(p_s, \mu_s)\}_{s \in S}$ satisfying Bayes plausibility $\sum_{s \in S} p_s \mu_s = \mu_0$ can be converted into a signaling scheme that sends signal s with probability p_s. Thus, we can use a distribution over posteriors $\{(\pi_s, \mu_s)\}_{s \in S}$ satisfying Bayes plausibility to represent a signaling scheme. Then, let's equate the posterior belief μ_s in Bayesian persuasion to the principal's decision x_s in the generalized principal-agent problem, so the principal/sender's decision space becomes $\mathcal{X} = \Delta(\Omega)$. The Bayes plausibility condition (5.2) becomes the constraint in the constrained generalized principal-agent problem. When the agent/receiver takes action a, the principal/sender's (expected) utility under decision/posterior $x_s = \mu_s$ is $u(x_s, a) = \mathbb{E}_{\omega \sim \mu_s} u(\omega, a) = \sum_{\omega \in \Omega} \mu_s(\omega)u(\omega, a)$. Suppose the agent takes action a_s given signal $s \in S$. Then we see that the sender's utility of using signaling scheme π in Bayesian persuasion (left) is equal to the principal's utility of using strategy π in the generalized principal-agent problem (right):

$$\sum_{\omega \in \Omega} \mu_0(\omega) \sum_{s \in S} \pi(s|\omega)u(\omega, a_s) = \sum_{s \in S} \pi_s \sum_{\omega \in \Omega} \mu_s(\omega)u(\omega, a_s) = \sum_{s \in S} \pi_s u(x_s, a_s) = \mathbb{E}_{s \sim \pi}[u(x_s, a)].$$

Similarly, the agent/receiver's utilities in the two problems are equal. The utility functions $u(x, a)$, $v(x, a)$ are linear in the principal's decision $x \in \mathcal{X}$, satisfying our assumption.

Persuasion (or cheap talk) with a learning agent When specialized to Bayesian persuasion, the generalized principal-agent problem with a learning agent becomes the following:

Persuasion (or Cheap Talk) with a Learning Receiver
In each round $t = 1, \ldots, T$, the following events happen:

(1) Using some algorithm that learns from history, the receiver chooses a strategy $\rho^t : S \to \Delta(A)$ that maps each signal $s \in S$ to a distribution over actions $\rho^t(s) \in \Delta(A)$.

(2) The sender chooses a signaling scheme $\pi^t : \Omega \to \Delta(S)$.

(3) A state of the world $\omega^t \sim \mu_0$ is realized, observed by the sender but not the receiver. The sender sends signal $s^t \sim \pi^t(\omega^t)$ to the receiver. The receiver draws action $a^t \sim \rho^t(s)$.

(4) The sender obtains utility $u^t = u(\omega^t, a^t)$ and the receiver obtains utility $v^t = v(\omega^t, a^t)$.[a]

[a] The definition of utility here, $u(\omega^t, a^t), v(\omega^t, a^t)$, is slightly different from the definition in Section 5.2.2, which was the expected utility on decision/posterior x^t, $u(x^t, a^t), v(x^t, a^t)$. Because we eventually only care about the sender's utility and the receiver's regret in expectation, this difference does not matter.

The receiver does not need to know the prior μ_0 if its learning algorithm does not make use of μ_0. And same as the model in Section 5.2.2, the receiver chooses ρ^t without knowing the sender's signaling scheme π^t, and the sender does not commit. In the classical *cheap talk* model [CS82], the sender does not have commitment power and the two players move simultaneously. So, under our learning receiver model, cheap talk and Bayesian persuasion are equivalent. Our "persuasion with a learning receiver" model can also be called "cheap talk with a learning receiver".

5.3 Reduction from Learning to Approximate Best Response

In this section, we reduce the generalized principal-agent problem with a learning agent to the problem with an approximately-best-responding agent. We show that, if the agent uses contextual no-regret learning algorithms, then the principal can obtain an average utility that is at least the "maxmin" approximate-best-response objective $\underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\mathrm{CReg}(T)/T\big)$ (to be defined below). On the other hand, if the agent does contextual no-swap-regret learning, then the principal cannot do better than the "maxmax" approximate-best-response objective $\overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\mathrm{CSReg}(T)/T\big)$. In addition, if the agent uses some learning algorithms that are no-regret but not no-swap-regret, the principal can sometimes do better than the "maxmax" objective $\overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\mathrm{CSReg}(T)/T\big)$.

5.3.1 Definitions of Approximate Best Response

We first define the generalized principal-agent problem with an *approximately-best-responding* agent. The classic generalized principal-agent problem (Section 5.2.1) assumes that, after receiving a signal $s \in S$ (and observing the principal's decision $x_s \in \mathcal{X}$), the agent will take an optimal action with respect to x_s. This means that the agent uses a strategy ρ^* that *best responds* to the principal's strategy π:

$$\rho^*(s) \in \operatorname*{arg\,max}_{a\in A} v(x_s, a), \quad \forall s \in S \qquad \Longrightarrow \qquad \rho^* \in \operatorname*{arg\,max}_{\rho: S\to\Delta(A)} V(\pi, \rho). \tag{5.3}$$

Here, $V(\pi, \rho) = \sum_{s\in S} \pi_s \sum_{a\in A} \rho(a|s) v(x_s, a)$ denotes the expected utility of the agent when the principal uses strategy π and the agent uses randomized strategy $\rho : S \to \Delta(A)$.

Here, we allow the agent to *approximately* best respond. Let $\delta \ge 0$ be a parameter. We define two types of δ-best-responding strategies for the agent: deterministic and randomized.

- A deterministic strategy ρ: for each signal $s \in S$, the agent takes an action a that is δ-optimal for x_s. Denote this set of strategies by $\mathcal{D}_\delta(\pi)$:

$$\mathcal{D}_\delta(\pi) = \big\{\rho : S \to A \mid v(x_s, \rho(s)) \ge v(x_s, a') - \delta, \ \forall a' \in A\big\}. \tag{5.4}$$

- A randomized strategy ρ: for each signals s, the agent can take a randomized action. The expected utility of ρ is at most δ-worst than the best strategy ρ^*.

$$\mathcal{R}_\delta(\pi) = \big\{\rho : S \to \Delta(A) \mid V(\pi, \rho) \ge V(\pi, \rho^*) - \delta\big\}. \tag{5.5}$$

Equivalently, $\mathcal{R}_\delta(\pi) = \big\{\rho : S \to \Delta(A) \mid V(\pi, \rho) \ge V(\pi, \rho') - \delta, \ \forall \rho' : S \to A\big\}$.

Our model of approximately-best-responding agent includes, for example, two other models in the Bayesian persuasion literature that also relax the agent's Bayesian rationality assumption: the quantal response model (proposed by [MP95] in normal-form games and studied by [FHT24] in Bayesian persuasion) and a model where the agent makes mistakes in Bayesian update [dCZ22].

Example 5.1. *Assume that the receiver's utility is in* $[0, 1]$*. In Bayesian persuasion, the following strategies of the receiver are* δ*-best-responding (see Section 5.8.1 for a proof):*

- Quantal response: *given signal* $s \in S$*, the agent chooses action* $a \in A$ *with probability* $\frac{\exp(\lambda v(\mu_s, a))}{\sum_{a' \in A} \exp(\lambda v(\mu_s, a'))}$*, with* $\lambda > 0$*. This strategy belongs to* $\mathcal{R}_\delta(\pi)$ *with* $\delta = \frac{1 + \log(|A|\lambda)}{\lambda}$.

- Inaccurate belief: *given signal $s \in S$, the agent forms some posterior μ'_s that is different yet close to the true posterior μ_s in total variation distance $d_{\mathrm{TV}}(\mu'_s, \mu_s) \le \varepsilon$. The agent picks an optimal action for μ'_s. This strategy belongs to $\mathcal{D}_{\delta=2\varepsilon}(\pi)$.*

Principal's objectives. With an approximately-best-responding agent, we will study two types of objectives for the principal. The first type is the maximal utility that the principal can obtain if the agent approximately best responds in the *worst* way for the principal: for $X \in \{\mathcal{D}, \mathcal{R}\}$, define

$$\underline{\mathrm{OBJ}}^X(\delta) = \sup_{\pi} \min_{\rho \in X_\delta(\pi)} U(\pi, \rho), \tag{5.6}$$

where $U(\pi, \rho) = \sum_{s \in S} \pi_s \sum_{a \in A} \rho(a|s) u(x_s, a)$ is the principal's expected utility when the principal uses strategy π and the agent uses strategy ρ. We used "sup" in (5.6) because the maximizer does not necessarily exist. $\underline{\mathrm{OBJ}}^X(\delta)$ is a "maxmin" objective and can be regarded as the objective of a "robust generalized principal-agent problem".

The second type of objectives is the maximal utility that the principal can obtain if the agent approximately best responds in the *best* way:

$$\overline{\mathrm{OBJ}}^X(\delta) = \max_{\pi} \max_{\rho \in X_\delta(\pi)} U(\pi, \rho). \tag{5.7}$$

This is a "maxmax" objective that quantifies the maximal extent to which the principal can exploit the agent's irrational behavior.

Clearly, $\underline{\mathrm{OBJ}}^X(\delta) \le \underline{\mathrm{OBJ}}^X(0) \le \overline{\mathrm{OBJ}}^X(0) \le \overline{\mathrm{OBJ}}^X(\delta)$. And we note that $\overline{\mathrm{OBJ}}^X(0) = \overline{\mathrm{OBJ}}(0)$ is independent of X and equal to the Stackelberg value U^* defined in (5.1):

$$\overline{\mathrm{OBJ}}(0) = \max_{\pi} \max_{\rho\text{: best-response to } \pi} U(\pi, \rho) = U^*. \tag{5.8}$$

Finally, we note that, because $\mathcal{D}_0(\pi) \subseteq \mathcal{D}_\delta(\pi) \subseteq \mathcal{R}_\delta(\pi)$, the chain of inequalities $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \le$

$\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \le U^* \le \overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \le \overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ hold.

5.3.2 Agent's No-Regret Learning: Lower Bound on Principal's Utility

Theorem 5.1. *Suppose the agent uses a learning algorithm with a contextual regret upper bounded by* $\mathrm{CReg}(T)$*. The principal knows* $\mathrm{CReg}(T)$ *but not the exact algorithm of the agent. By using some fixed strategy* $\pi^t = \pi$ *for all* T *rounds, the principal can obtain an average utility* $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big]$ *that is arbitrarily close to* $\underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big)$.

To prove Theorem 5.1, we provide a lemma to relate the agent's regret and the principal's utility in the learning model to those in the static model. We define some notations. Let the principal use some fixed strategy $\pi^t = \pi$ and the agent use some learning algorithm. Let $p^t_{a|s} = \Pr[a^t = a \,|\, s^t = s]$ be the probability that the agent's algorithm chooses action a conditioning on signal s being sent in round t. Let $\rho : S \to \Delta(A)$ be a randomized agent strategy that, given signal s, chooses each action $a \in A$ with probability $\rho(a|s) = \frac{\sum_{t=1}^{T} p^t_{a|s}}{T}$.

Lemma 5.1. *When the principal uses a fixed strategy* $\pi^t = \pi$ *in all* T *rounds, the regret of the agent not deviating according to* $d : S \to A$ *is equal to* $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T}\big(v(x^t, d(s^t)) - v(x^t, a^t)\big)\big] = V(\pi, d) - V(\pi, \rho)$*, and the average utility of the principal* $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big]$ *is equal to* $U(\pi, \rho)$*.*

Proof. Since $\pi^t = \pi$ is fixed, we have $\pi^t_s = \pi_s$ and $x^t_s = x_s$, $\forall s \in S$. The regret of the agent not deviating according to d is:

$$\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}\Big(v(x^t, d(s^t)) - v(x^t, a^t)\Big)\Big] = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\pi_s^t\sum_{a\in A}p_{a|s}^t\Big(v(x_s^t, d(s)) - v(x_s^t, a)\Big)$$
$$= \sum_{s\in S}\pi_s\sum_{a\in A}\frac{\sum_{t=1}^T p_{a|s}^t}{T}\Big(v(x_s, d(s)) - v(x_s, a)\Big)$$
$$= \sum_{s\in S}\pi_s v(x_s, d(s)) \ - \ \sum_{s\in S}\pi_s\sum_{a\in A}\rho(a|s)v(x_s, a) \ = \ V(\pi, d) - V(\pi, \rho).$$

Here, d is interpreted as an agent strategy that deterministically takes action $d(s)$ for signal s.

By a similar derivation, we see that the principal's expected utility is equal to $\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^T u(x^t, a^t)\Big] = \sum_{s\in S}\pi_s\sum_{a\in A}\frac{\sum_{t=1}^T p_{a|s}^t}{T}u(x_s, a) = U(\pi, \rho)$, which proves the lemma.

□

Proof of Theorem 5.1. By Lemma 5.1 and the no-regret condition that the agent's regret $\mathbb{E}\big[\sum_{t=1}^T\big(v(x^t, d(s^t)) - v(x^t, a^t)\big)\big] \le \mathrm{CReg}(T)$, we have

$$V(\pi, d) - V(\pi, \rho) \ = \ \frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}\Big(v(x^t, d(s^t)) - v(x^t, a^t)\Big)\Big] \ \le \ \frac{\mathrm{CReg}(T)}{T}, \quad \forall d : S \to A.$$

This means that the agent's randomized strategy ρ is a $\delta = \frac{\mathrm{CReg}(T)}{T}$-best-response to the principal's fixed signaling scheme π, namely $\rho \in \mathcal{R}_{\delta=\frac{\mathrm{CReg}(T)}{T}}(\pi)$. This holds for any π. In particular, if for any $\varepsilon > 0$ the principal uses a signaling scheme π^ε that obtains an objective that is ε-close to $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_\pi \min_{\rho\in\mathcal{R}_\delta(\pi)} U(\pi, \rho)$, then by Lemma 5.1, the principal's expected utility in the learning model is at least

$$\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}u(a^t, \omega^t)\Big] \ = \ U(\pi^\varepsilon, \rho) \ \ge \min_{\rho\in\mathcal{R}_\delta(\pi^\varepsilon)} U(\pi^\varepsilon, \rho) \ \ge \ \underline{\mathrm{OBJ}}^{\mathcal{R}}\Big(\delta = \frac{\mathrm{CReg}(T)}{T}\Big) - \varepsilon.$$

Letting $\varepsilon \to 0$ proves the theorem. □

We then show that the result in Theorem 5.1 is tight: there exist cases where the principal cannot do better than $\underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big)$ even using adaptive strategies:

Proposition 5.2. *For any strategy of the principal that depends on history but not on ρ^t, there exists an agent's learning algorithm with contextual regret at most* $\mathrm{CReg}(T)$ *under which the principal's average utility is no more than* $\underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big)$. *The same holds for agent's learning algorithms with contextual swap regret at most* $\mathrm{CSReg}(T)$.

Proof. Fix the principal's strategy $\sigma = (\sigma^t)_{t=1}^T$, where each σ^t is a mapping from the history $h^{t-1} = (s^i, a^i)_{i=1}^{t-1}$ (including past signals and actions) to the strategy π^t for round t. Given any function $\mathrm{CReg}(T)$, let $\delta = \frac{\mathrm{CReg}(T)}{T}$. Consider the following algorithm for the agent: at each round t, given history $h^{t-1} = (s^i, a^i)_{i=1}^{t-1}$, compute the principal's strategy $\pi^t = \sigma^t(h^{t-1})$, then play a strategy $\rho^t \in \arg\min_{\rho \in \mathcal{R}_\delta(\pi^t)} U(\pi^t, \rho)$, namely, play a randomized δ-best-responding strategy that minimizes the principal's utility. Since the agent δ-best responds to the principal's strategy at every round, the agent's total regret is at most $T\delta = \mathrm{CReg}(T)$. The principal's average utility is

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[u(x^t, a^t)] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{h^{t-1}}\Big[U(\pi^t, \rho^t)\Big] \\
&\le \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{h^{t-1}}\Big[\sup_{\pi}\min_{\rho\in\mathcal{R}_\delta(\pi)} U(\pi, \rho)\Big] \qquad \text{because } \rho^t \in \arg\min_{\rho\in\mathcal{R}_\delta(\pi^t)} U(\pi^t, \rho) \\
&= \underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\tfrac{\mathrm{CReg}(T)}{T}\big).
\end{aligned}$$

The same proofs works for contextual no-swap-regret learning algorithms. □

5.3.3 Agent's No-Swap-Regret Learning: Upper Bound on Principal's Utility

As we mentioned in Section 5.2.2, the fact that the principal moves after the learning agent in each round gives the principal a possibility to exploit the agent, so as to do better than U^*. However, exploiting the agent in a single round may cause the agent to learn a bad strategy for the principal in later rounds. It turns out that, if the agent's learning algorithm satisfies the contextual no-swap-regret property, then the principal cannot exploit the agent in the long run. Formally:

Theorem 5.2. *Against a contextual no-swap-regret learning agent, the principal cannot obtain utility more than* $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \le \overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CSReg}(T)}{T}\big)$, *even if the principal knows the agent's learning algorithm and chooses* π^t *based on* ρ^t.

Before presenting the full proof of this theorem, we give the high level idea of the proof. The key idea is to think of the signal $s^t \sim \pi^t$ from the principal and the action $a^t \sim \rho^t(s^t)$ recommended by the agent's learning algorithm together as a joint signal (s^t, a^t) from some hypothetical signaling scheme π'. In response to π', the agent takes the action a^t recommended by the algorithm, namely using the mapping $(s^t, a^t) \mapsto a^t$ as his strategy. A contextual no-swap-regret algorithm guarantees that the agent is at most $\frac{\mathrm{CSReg}(T)}{T}$ worse compared to using the strategy $d^* : S \times A \to A$ that best responds to π'. So, the agent's overall strategy is a $\frac{\mathrm{CSReg}(T)}{T}$-approximate best response to π'. This limits the principal's overall utility to be at most $\overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CSReg}(T)}{T}\big)$. See details below:

Proof. Let $p_s^t = \Pr[s^t = s] = \mathbb{E}\big[\mathbb{1}[s^t = s]\big] = \mathbb{E}[\pi_s^t]$ be the (unconditional) probability that signal s is sent in round t. Let $p_{a|s}^t = \Pr[a^t = a \,|\, s^t = s]$ be the probability that the agent's algorithm takes action a conditioning on signal s being sent in round t. Let $d : S \times A \to A$ be any deviation function for the agent. The agent's utility gain by deviation is upper bounded by the contextual swap regret:

$$
\begin{aligned}
\frac{\mathrm{CSReg}(T)}{T} &\geq \frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T}\Big(v(x^t, d(s^t, a^t)) - v(x^t, a^t)\Big)\Big] \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S} p_s^t \sum_{a\in A} p_{a|s}^t \mathbb{E}\Big[v(x_s^t, d(s,a)) - v(x_s^t, a) \;\Big|\; s^t = s, a^t = a\Big] \\
&= \sum_{s\in S}\sum_{a\in A} \frac{\sum_{j=1}^{T} p_s^j p_{a|s}^j}{T} \frac{1}{\sum_{j=1}^{T} p_s^j p_{a|s}^j} \sum_{t=1}^{T} p_s^t p_{a|s}^t \mathbb{E}\Big[v(x_s^t, d(s,a)) - v(x_s^t, a) \,\Big|\, s^t = s, a^t = a\Big] \\
&= \sum_{s\in S}\sum_{a\in A} \frac{\sum_{j=1}^{T} p_s^j p_{a|s}^j}{T} \Bigg[v\Big(\tfrac{\sum_{t=1}^{T} p_s^t p_{a|s}^t \mathbb{E}[x_s^t|s^t=s,a^t=a]}{\sum_{j=1}^{T} p_s^j p_{a|s}^j}, d(s,a)\Big) - v\Big(\tfrac{\sum_{t=1}^{T} p_s^t p_{a|s}^t \mathbb{E}[x_s^t|s^t=s,a^t=a]}{\sum_{j=1}^{T} p_s^j p_{a|s}^j}, a\Big)\Bigg],
\end{aligned}
\tag{5.9}
$$

where the last line is because of linearity of $v(\cdot, a)$. Define

$$
q_{s,a} = \frac{\sum_{j=1}^{T} p_s^j p_{a|s}^j}{T} \quad \text{and} \quad y_{s,a} = \frac{\sum_{t=1}^{T} p_s^t p_{a|s}^t \mathbb{E}[x_s^t|s^t = s, a^t = a]}{\sum_{j=1}^{T} p_s^j p_{a|s}^j} \in \mathcal{X}.
$$

Then (5.9) becomes

$$
\frac{\mathrm{CSReg}(T)}{T} \;\geq\; \sum_{s\in S}\sum_{a\in A} q_{s,a}\Big[v(y_{s,a}, d(s,a)) - v(y_{s,a}, a)\Big]. \tag{5.10}
$$

We note that $\sum_{s\in S}\sum_{a\in A} q_{s,a} = \frac{\sum_{j=1}^{T}\sum_{s\in S}\sum_{a\in A} p_s^j p_{a|s}^j}{T} = 1$, so q is a probability distribution over $S \times A$. And note that

$$
\begin{aligned}
\sum_{s,a\in S\times A} q_{s,a} y_{s,a} \;&=\; \sum_{s,a\in S\times A} \frac{1}{T}\sum_{t=1}^{T} p_s^t p_{a|s}^t \mathbb{E}[x_s^t|s^t = s, a^t = a] \;=\; \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S} p_s^t \mathbb{E}[x_s^t|s^t = s] \\
&=\; \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S}\mathbb{E}\big[\mathbb{1}[s^t = s]x_s^t\big] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\sum_{s\in S}\mathbb{1}[s^t = s]x_s^t\big] \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[x^t\big] \\
&=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\sum_{s\in S}\pi_s^t x_s^t\big] \in \mathcal{C} \qquad \text{because } \sum_{s\in S}\pi_s^t x_s^t \in \mathcal{C}.
\end{aligned}
$$

This means that $\pi' = \{(q_{s,a}, y_{s,a})\}_{(s,a)\in S\times A}$ defines a valid principal strategy with the larger signal space $S \times A$.[a] Then, we note that the right side of (5.10) is the difference between the agent's expected utility under principal strategy π' when responding by strategy $d : S \times A \to A$ and responding by the strategy that maps signal (s, a) to action a. So, (5.10) becomes

$$\frac{\mathrm{CSReg}(T)}{T} \;\geq\; V(\pi', d) - V(\pi', (s,a) \mapsto a), \quad \forall d : S \times A \to A. \tag{5.11}$$

In particular, this holds when d is the agent's best-responding strategy. This means that the agent strategy $(s, a) \mapsto a$ is a $(\frac{\mathrm{CSReg}(T)}{T})$-best-response to π'. So, the principal's expected utility is upper bounded by the utility in the approximate-best-response model:

$$\begin{aligned}\frac{1}{T}\mathbb{E}\Big[\sum_{t=1}^{T} u(x^t, a^t)\Big] &= \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S} p_s^t \sum_{a\in A} p_{a|s}^t \mathbb{E}\big[u(x_s^t, a) \mid s^t = s, a^t = a\big] \\ &= \sum_{s\in S}\sum_{a\in A} q_{s,a} u(y_{s,a}, a) \;=\; U(\pi', (s,a)\mapsto a) \;\leq\; \overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\tfrac{\mathrm{CSReg}(T)}{T}\big).\end{aligned}$$

□

[a] As long as $|S| \geq |A|$, enlarging the signal space from S to $S \times A$ will not change the optimal objective for the principal, because the optimal strategy of the principal only needs to use $|A|$ signals by the revelation principle.

Similar to Proposition 5.2, we can show that the result in Theorem 5.2 is tight: there exist cases where the principal can achieve $\overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CSReg}(T)}{T}\big)$. The proof is straightforward and hence omitted.

Proposition 5.3. *There exists a fixed strategy π for the principal and a learning algorithm for the agent with contextual swap regret at most* CSReg(T) *under which the principal can achieve average utility* $\overline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CSReg}(T)}{T}\big)$.

5.3.4 Agent's Mean-Based Learning: Exploitable by Principal

Many no-regret (but not no-swap-regret) learning algorithms (e.g., MWU, FTPL, EXP-3) satisfy the following *contextual mean-based* property:

Definition 5.2 ([BMSW18]). *Let* $\sigma_s^t(a) = \sum_{j\in[t]:s^j=s} v(\omega^j, a)$ *be the sum of historical utilities of the agent in the first* t *rounds if he takes action* a *when the signal/context is* s*. An algorithm is called* γ-mean-based *if: whenever there exists* a' *such that* $\sigma_s^{t-1}(a) < \sigma_s^{t-1}(a') - \gamma T$*, the probability that the algorithm chooses action* a *at round* t *if the context is* s *is* $\Pr[a^t = a \mid s^t = s] < \gamma$*, with* $\gamma = o(1)$*.*

It is known that a mean-based learning agent can sometimes be exploited by the principal in Stackelberg games [DSS19]. We show that this also holds in Bayesian persuasion:

Theorem 5.3. *There exists a Bayesian persuasion instance where, as long as the receiver does* γ*-mean-based learning, the sender can obtain a utility significantly larger than* $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\gamma)$ *and* U^**.*

To prove Theorem 5.3, consider the following instance: There are 2 states (A, B), 3 actions (L, M, R), with uniform prior $\mu_0(\mathrm{A}) = \mu_0(\mathrm{B}) = 0.5$ and the following utility matrices (left for sender, right for receiver):

$u(\omega, a)$	L	M	R
A	0	-2	-2
B	0	0	2

$v(\omega, a)$	L	M	R
A	$\sqrt{\gamma}$	-1	0
B	-1	1	0

Claim 5.1. *In this instance, the optimal sender utility* U^* *in the classic BP model is* 0*, and the approximate-best-response objective* $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\gamma) = O(\gamma)$*.*

Proof. Recall that any signaling scheme decomposes the prior μ_0 into multiple posteriors $\{\mu_s\}_{s\in S}$. If a posterior μ_s puts probability > 0.5 to state B, then the receiver will take action M, which gives the sender a utility ≤ 0; if the posterior μ_s puts probability ≤ 0.5 to state B, then no matter what action the receiver takes, the sender's expected utility on μ_s cannot be greater than 0. So, the sender's expected utility is ≤ 0 under any signaling scheme. An optimal signaling scheme is to reveal no information (keep $\mu_s = \mu_0$); the receiver takes R and the sender gets utility 0.

This instance satisfies the assumptions of Theorem 5.5, so $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\gamma) \leq U^* + O(\gamma) = O(\gamma)$. □

Claim 5.2. *By doing the following, the sender can obtain utility* $\approx \frac{1}{2} - O(\sqrt{\gamma})$ *if the receiver is* γ*-mean-based learning:*

- *in the first* $T/2$ *rounds: if the state is* A*, send signal 1; if the state is* B*, send 2.*
- *in the remaining* $T/2$ *rounds, switch the scheme: if the state is* A*, send 2; if state is* B*, send 1.*

Proof. In the first $T/2$ rounds, the receiver finds that signal 1 corresponds to state A so he will take action L with high probability when signal 1 is sent; signal 2 corresponds to B so he will take action M with high probability. In this phase, the sender obtains utility ≈ 0 per round. At the end of this phase, for signal 1, the receiver accumulates utility $\approx \frac{T}{2}\frac{1}{2}\sqrt{\gamma} = \frac{T}{4}\sqrt{\gamma}$ for action L. For signal 2, the receiver accumulates utility $\approx \frac{T}{2}\frac{1}{2}\cdot 1 = \frac{T}{4}$ for action M.

In the remaining $T/2$ rounds, the following will happen:

- For signal 1, the receiver finds that the state is now B, so the utility of action L decreases by 1 every time signal 1 is sent. Because the utility of L accumulated in the first phase was $\approx \frac{T}{4}\sqrt{\gamma}$, after $\approx \frac{T}{4}\sqrt{\gamma}$ rounds in second phase the utility of

L should decrease to below 0, and the receiver will no longer play L (with high probability) at signal 1. The receiver will not play M at signal 1 in most of the second phase either, because there are more A states than B states at signal 1 historically. So, the receiver will play action R most times, roughly $\frac{T}{4} - \frac{T}{4}\sqrt{\gamma}$ rounds. This gives the sender a total utility of $\approx (\frac{T}{4} - \frac{T}{4}\sqrt{\gamma}) \cdot 2 = \frac{T}{2} - O(T\sqrt{\gamma})$.

- For signal 2, the state is now A. But the receiver will continue to play action M in most times. This because: R has utility 0; L accumulated $\approx -\frac{T}{4}$ utility in the first phase, and only increases by $\sqrt{\gamma}$ per round in the second phase, so its accumulated utility is always negative; instead, M has accumulated $\frac{T}{4}$ utility in the first phase, and decreases by 1 every time signal 2 is sent in the second phase, so its utility is positive until near the end. So, the receiver will play M. This gives the sender utility 0.

Summing up, the sender obtains total utility $\approx \frac{T}{2} - O(T\sqrt{\gamma})$ in these two phases, which is $\frac{1}{2} - O(\sqrt{\gamma}) > 0$ per round in average. □

The above two claims together prove the Theorem 5.3.

5.4 Generalized Principal-Agent Problems with Approximate Best Response

After presenting the reduction from learning to approximate best response, we now study generalized principal-agent problems with approximate best response. We will show that both the maxmin objectives $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$, $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ and the maxmax objectives $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$, $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ are close to the optimal principal objective U^* in the best-response model when the degree δ of the agent's approximate best response is small, under some natural assumptions described below.

Assumptions and notations. We make some innocuous assumptions. First, the agent has no weakly dominated action:

Assumption 5.1 (No Dominated Action)**.** *An action $a_0 \in A$ of the agent is* weakly dominated *if there exists a mixed action $\alpha' \in \Delta(A \setminus \{a_0\})$ such that $v(x, \alpha') = \mathbb{E}_{a \sim \alpha'}[v(x, a)] \geq v(x, a_0)$ for all $x \in \mathcal{X}$. We assume that the agent has no weakly dominated action.*

Claim 5.3. *Assumption 5.1 implies: there exists a constant $G > 0$ such that, for any agent action $a \in A$, there exists a principal decision $x \in \mathcal{X}$ such that $v(x, a) - v(x, a') \geq G$ for every $a' \in A \setminus \{a\}$.*

The proof of this claim is in Section 5.9.1. The constant $G > 0$ in Claim 5.3 is analogous to the concept of "inducibility gap" in Stackelberg games [VSZ04, GHWX23]. In fact, [GHWX23] show that, if the inducibility gap $G > \delta$, then the maximin approximate-best-response objective satisfies $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \geq U^* - \frac{\delta}{G}$ in Stackelberg games. Our results will significantly generalize theirs to any generalized principal-agent problem, to randomized agent strategies, and to the maximax objectives $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$, $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$.

To present our results, we need to introduce a few more notions and assumptions. Let

$$\mathrm{diam}(\mathcal{X}; \|\cdot\|) = \max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|$$

be the diameter of the space $\mathcal{X}$, where $\|\cdot\|$ is some norm. For convenience we assume $\mathcal{X} \subseteq \mathbb{R}^d$ and use the ℓ_1-norm $\|x\|_1 = \sum_{i=1}^d |x_{(i)}|$ or the ℓ_∞-norm $\|x\|_\infty = \max_{i=1}^d |x_{(i)}|$. For a generalized principal-agent problem with constraint $\sum_{s \in S} \pi_s x_s \in \mathcal{C}$, let $\partial \mathcal{X}$ be the boundary of $\mathcal{X}$ and let $D(\mathcal{C}, \partial \mathcal{X}) = \min_{c \in \mathcal{C}, x \in \partial X} \|c - x\|$ be the distance from $\mathcal{C}$ to the boundary of $\mathcal{X}$. We assume that $\mathcal{C}$ is away from the boundary of $\mathcal{X}$:

Assumption 5.2 ($\mathcal{C}$ is in the interior of $\mathcal{X}$)**.** $D(\mathcal{C}, \partial\mathcal{X}) > 0$.

Assumption 5.3 (Bounded and Lipschitz utility)**.** *The principal's utility function is bounded: $|u(x, a)| \le B$, and L-Lipschitz in $x \in \mathcal{X}$: $|u(x_1, a) - u(x_2, a)| \le L\|x_1 - x_2\|$.*

5.4.1 Main Results

We now present the main results of this section: lower bounds on $\underline{\mathrm{OBJ}}^X(\delta)$ and upper bounds on $\overline{\mathrm{OBJ}}^X(\delta)$ in generalized principal-agent problems without and with constraints.

Theorem 5.4 (Without constraint)**.** *For an unconstrained generalized principal-agent problem, under Assumptions 5.1 and 5.3, for $0 \le \delta < G$, we have*

- $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \ge U^* - \mathrm{diam}(\mathcal{X})L\frac{\delta}{G}$.
- $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \ge U^* - 2\sqrt{\frac{2BL}{G}\mathrm{diam}(\mathcal{X})\delta}$ *for* $\delta < \frac{\mathrm{diam}(\mathcal{X})GL}{2B}$.
- $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \le \overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \le U^* + \mathrm{diam}(\mathcal{X})L\frac{\delta}{G}$.

Theorem 5.5 (With constraint)**.** *For a generalized principal-agent problem with the constraint $\sum_{s\in S}\pi_s x_s \in \mathcal{C}$, under Assumptions 5.1, 5.2 and 5.3, for $0 \le \delta < \frac{D(\mathcal{C},\partial\mathcal{X})}{\mathrm{diam}(\mathcal{X})}G$, we have*

- $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \ge U^* - \big(\mathrm{diam}(\mathcal{X})L + 2B\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\frac{\delta}{G}$.
- $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \ge U^* - 2\sqrt{\frac{2B}{G}\big(\mathrm{diam}(\mathcal{X})L + 2B\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\delta}$.
- $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \le \overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \le U^* + \big(\mathrm{diam}(\mathcal{X})L + 2B\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\frac{\delta}{G}$.

The expression "$\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\delta$" suggests that $\frac{1}{D(\mathcal{C},\partial\mathcal{X})}$ is similar to a "condition number" [Ren94] that quantifies the "stability" of the principal-agent problem against the agent's approximate-best-responding behavior. When $D(\mathcal{C}, \partial\mathcal{X})$ is larger ($\mathcal{C}$ is further away from the boundary of $\mathcal{X}$), the condition number is smaller, the problem is more stable, and the δ-best-response objectives $\underline{\mathrm{OBJ}}^X(\delta)$ and $\overline{\mathrm{OBJ}}^X(\delta)$ are closer to the best-response objective U^*.

High-level idea: perturbation. The high level idea to prove Theorems 5.4 and 5.5 is a perturbation argument. Consider proving the upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ for example. Let (π, ρ) be any pair of principal's strategy and agent's δ-best-responding strategy. We perturb the principal's strategy π slightly to be a strategy π' such that ρ is *exactly* best-responding to π' (such a perturbation is possible due to Assumption 5.1). Since ρ is best-responding to π', the pair (π', ρ) cannot give the principal a higher utility than U^* (which is the optimal principal utility under the best-response model). This means that the original pair (π, ρ) cannot give the principal a utility much higher than U^*, thus implying an upper bound on $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$. Extra care is needed when dealing with randomized strategies of the agent. See details in Section 5.9.3.

The bound $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq U^* - O(\sqrt{\delta})$ is tight. We note that, in Theorems 5.4 and 5.5, the maxmin objective with randomized agent strategies is bounded by $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq U^* - O(\sqrt{\delta})$, while the objective with deterministic agent strategies is bounded by $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \geq U^* - O(\delta)$. This is *not* because our analysis is not tight. In fact, the squared root bound $U^* - \Theta(\sqrt{\delta})$ for $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ is tight. We prove this by giving an example where $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \leq U^* - \Omega(\sqrt{\delta})$. Consider the following classical Bayesian persuasion example:

Example 5.2. *There are 2 states* $\Omega = \{\mathrm{Good}, \mathrm{Bad}\}$*, 2 actions* $A = \{a, b\}$*, with the following utility matrices*

sender	*a*	*b*
Good	1	0
Bad	1	0

receiver	*a*	*b*
Good	1	0
Bad	-1	0

The prior probability of Good *state is* $0 < \mu_0 < \frac{1}{2}$*, so the receiver takes action* b *by default. In this example, we have* $\mathrm{diam}(\mathcal{X}) = 1$*,* $D(\mathcal{C}, \partial\mathcal{X}) = \mu_0$*,* $U^* = 2\mu_0$*, and:*

- *for* $\delta < \frac{\mu_0}{2}$*,* $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \le U^* - 2\sqrt{2\mu_0\delta} + \delta = U^* - \Omega(\sqrt{\delta})$*.*
- *for* $\delta < 1 - 2\mu_0$*,* $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \ge U^* + \delta$*.*

See Section 5.9.2 for a proof.

5.5 Applications: Specific Principal-Agent Problems

We apply the general results in Section 5.3 and 5.4 to derive concrete results for three specific principal-agent problems: Bayesian persuasion, Stackelberg games, and contract design.

5.5.1 Bayesian Persuasion

As noted in Section 5.2, Bayesian persuasion is a generalized principal-agent problem with constraint $\sum_{s\in S} \pi_s x_s \in \mathcal{C} = \{\mu_0\}$ where each $x_s = \mu_s = (\mu_s(\omega))_{\omega\in\Omega} \in \mathcal{X} = \Delta(\Omega)$ is a posterior belief. Suppose the principal's utility is bounded: $|u(\omega, a)| \le B$. Then, the principal's utility function $u(\mu_s, a) = \sum_{\omega\in\Omega} \mu_s(\omega)u(\omega, a)$ is $(L = B)$-Lipschitz in μ_s (under ℓ_1-norm), so Assumption 5.3 is satisfied. Suppose the prior μ_0 has positive probability for every $\omega \in \Omega$, and let $p_0 = \min_{\omega\in\Omega} \mu_0(\omega) > 0$. Then, we have the distance

$$D(\mathcal{C}, \partial X) = \min\left\{\|\mu_0 - \mu\|_1 : \mu \in \Delta(\Omega) \text{ s.t. } \mu(\omega) = 0 \text{ for some } \omega \in \Omega\right\} \ge p_0 > 0,$$

so Assumption 5.2 is satisfied. The diameter satisfies

$$\mathrm{diam}(\mathcal{X}; \ell_1) = \max_{\mu_1, \mu_2 \in \Delta(\Omega)} \|\mu_1 - \mu_2\|_1 \le 2.$$

Finally, we assume Assumption 5.1 (no dominated action for the agent). Then, Theorem 5.5 gives bounds on the approximate-best-response objectives in Bayesian persuasion:

Corollary 5.1 (Bayesian persuasion with approximate best response)**.** *For* $0 \le \delta < \frac{Gp_0}{2}$,

- $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \ge U^* - 2B(1 + \frac{2}{p_0})\frac{\delta}{G}$.
- $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \ge U^* - 4B\sqrt{(1 + \frac{2}{p_0})\frac{\delta}{G}}$.
- $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \le \overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \le U^* + 2B(1 + \frac{2}{p_0})\frac{\delta}{G}$.

Further applying Theorem 5.1 and 5.2, we obtain the central result for our motivating problem, persuasion with a learning agent:

Corollary 5.2 (Persuasion with a learning agent)**.** *Suppose* T *is sufficiently large such that* $\frac{\mathrm{CReg}(T)}{T} < \frac{Gp_0}{2}$ *and* $\frac{\mathrm{CSReg}(T)}{T} < \frac{Gp_0}{2}$*, then*

- *with a contextual no-regret learning agent, the principal can obtain utility at least*

  $$\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \ge \underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\tfrac{\mathrm{CReg}(T)}{T}\big) \ge U^* - 4B\sqrt{(1 + \tfrac{2}{p_0})\tfrac{1}{G}}\sqrt{\tfrac{\mathrm{CReg}(T)}{T}} \quad (5.12)$$

  using a fixed signaling scheme in all rounds.

- *with a contextual no-swap-regret learning agent, the principal's obtainable utility* *is at most*

$$\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \;\le\; \overline{\mathrm{OBJ}}^{\mathcal{D}}\big(\tfrac{\mathrm{CSReg}(T)}{T}\big) \;\le\; U^* + 2B(1+\tfrac{2}{p_0})\tfrac{1}{G}\tfrac{\mathrm{CSReg}(T)}{T} \qquad (5.13)$$

even knowing the receiver's learning algorithm and using adaptive signaling schemes.

Result (5.13) is interesting because it shows that the sender cannot exploit a no-swap-regret learning receiver beyond $U^* + o(1)$ even if the sender has informational advantage (knowing the state ω) and knows the receiver's algorithm or strategy ρ^t before choosing the signaling scheme. Result (5.12) is interesting because it shows that the sender can achieve the Bayesian persuasion optimal objective (which is U^*) in the problem of cheap talk with a learning agent (recall from Section 5.2.3 that persuasion and cheap talk are equivalent in our model).

5.5.2 Stackelberg Games

In a Stackelberg game, the principal (leader), having a finite action set B, first commits to a mixed strategy $x = (x_{(b)})_{b\in B} \in \Delta(B)$, which is a distribution over actions. So the principal's decision space $\mathcal{X}$ is $\Delta(B)$. The agent (follower) then takes an action $a \in A$ in response to x. The (expected) utilities for the two players are $u(x,a) = \sum_{b\in B} x_{(b)} u(b,a)$ and $v(x,a) = \sum_{b\in B} x_{(b)} u(b,a)$. The signal s can (but not necessarily) be an action that the principal recommends the agent to take.

Assume bounded utility $|u(b,a)| \le B$. Then, the principal's utility function $u(x,a)$ is bounded in $[-B,B]$ and $(L = B)$-Lipschitz in x. The diameter

$$\mathrm{diam}(\mathcal{X}) = \max_{x_1,x_2\in\Delta(B)} \|x_1 - x_2\|_1 \le 2.$$

Applying the theorem for unconstrained generalized principal-agent problems (Theorem 5.4) and the theorems for learning agent (Theorem 5.1 and 5.2), we obtain:

Corollary 5.3 (Stackelberg game with a learning agent)**.** *Suppose T is sufficiently large such that $\frac{\mathrm{CReg}(T)}{T} < G$ and $\frac{\mathrm{CSReg}(T)}{T} < G$, then:*

- *with a contextual no-regret learning agent, the principal can obtain utility $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \geq \underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big) \geq U^* - \frac{4B}{\sqrt{G}}\sqrt{\frac{\mathrm{CReg}(T)}{T}}$.*
- *with a contextual no-swap-regret learning agent, the principal cannot obtain utility more than $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \leq \overline{\mathrm{OBJ}}^{\mathcal{D}}\big(\frac{\mathrm{CSReg}(T)}{T}\big) \leq U^* + \frac{2B}{G}\frac{\mathrm{CSReg}(T)}{T}$.*

The conclusion that the principal can obtain utility at least $U^* - o(1)$ against a no-regret learning agent and no more than $U^* + o(1)$ against a no-swap-regret agent in Stackelberg games was proved by [DSS19]. Our Corollary 5.3 reproduces this conclusion and moreover provides bounds on the $o(1)$ terms, namely, $U^* - O(\sqrt{\frac{\mathrm{CReg}(T)}{T}})$ and $U^* + O(\frac{\mathrm{CSReg}(T)}{T})$. This demonstrates the generality and usefulness of our framework.

5.5.3 Contract Design

In contract design, there is a finite outcome space $O = \{r_1, \dots, r_d\}$ where each $r_i \in \mathbb{R}$ is a monetary reward to the principal. When the agent takes action $a \in A$, outcome r_i will happen with probability $p_{ai} \geq 0$, $\sum_{i=1}^{d} p_{ai} = 1$. The principal cannot observe the action taken by the agent but can observe the realized outcome. The principal's decision space $\mathcal{X}$ is the set of contracts, where a contract $x = (x_{(i)})_{i=1}^{d} \in [0, +\infty]^d$ is a vector that specifies the payment to the agent for each possible outcome. So, if the agent takes action a under contract x, the principal obtains expected utility

$$u(x, a) = \sum_{i=1}^{d} p_{ai}(r_i - x_{(i)})$$

and the agent obtains $v(x, a) = \sum_{i=1}^{d} p_{ai} x_{(i)} - c_a$, where $c_a \geq 0$ is the cost of action $a \in A$ for the agent. The signal s can (but not necessarily) be an action that the principal recommends the agent to take. The principal's decision space $\mathcal{X} \subseteq [0, +\infty]^d$ in contract design, however, may be unbounded and violate the requirement of bounded diameter $\mathrm{diam}(\mathcal{X})$ that we need. We have two remedies for this.

The first remedy is to require the principal's payment to the agent be upper bounded by some constant $P < +\infty$, so $0 \leq x_{(i)} \leq P$ and $\mathcal{X} = [0, P]^d$. Under this requirement and the assumption of bounded reward $|r_i| \leq R$, the principal's utility becomes bounded by $|u(x, a)| \leq \sum_{i=1}^{d} p_{ai}(R + P) = R + P = B$ and $(L = 1)$-Lipschitz under ℓ_∞-norm:

$$|u(x_1, a) - u(x_2, a)| = \big| \sum_{i=1}^{d} p_{ai}(x_{1(i)} - x_{2(i)}) \big| \leq \max_{i=1}^{d} |x_{1(i)} - x_{2(i)}| \sum_{i=1}^{d} p_{ai} = \|x_1 - x_2\|_\infty.$$

And the diameter of $\mathcal{X}$ is bounded by (under ℓ_∞-norm)

$$\mathrm{diam}(\mathcal{X}; \ell_\infty) = \max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|_\infty = \max_{x_1, x_2 \in [0,P]^d} \max_{i=1}^{d} |x_{1(i)} - x_{2(i)}| \leq P.$$

Now, we can apply the theorem for unconstrained generalized principal-agent problems (Theorem 5.4) and the theorems for learning agent (Theorem 5.1 and Theorem 5.2) to obtain:

Corollary 5.4 (Contract design (with bounded payment) with a learning agent**).** *Suppose T is sufficiently large such that $\frac{\mathrm{CReg}(T)}{T} < \frac{PG}{2(R+P)}$ and $\frac{\mathrm{CSReg}(T)}{T} < G$, then:*

- *with a contextual no-regret learning agent, the principal can obtain utility at least $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \geq \underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big) \geq U^* - 2\sqrt{\frac{2(R+P)P}{G}}\sqrt{\frac{\mathrm{CReg}(T)}{T}}$.*
- *with contextual a no-swap-regret learning agent, the principal cannot obtain utility more than $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \leq \overline{\mathrm{OBJ}}^{\mathcal{D}}\big(\frac{\mathrm{CSReg}(T)}{T}\big) \leq U^* + \frac{P}{G}\frac{\mathrm{CSReg}(T)}{T}$.*

The second remedy is to write contract design as a generalized principal-agent problem in another way. Let $\tilde{x} = (\tilde{x}_{(a)})_{a \in A} \in [0, +\infty]^{|A|}$ be a vector recording the *expected payment* from the principal to the agent for each action $a \in A$:

$$\tilde{x}_{(a)} = \sum_{i=1}^{d} p_{ai} x_{(i)}.$$

And let $\tilde{r}_{(a)}$ be the expected reward of action a, $\tilde{r}_{(a)} = \sum_{i=1}^{d} p_{ai} r_i$. Then, the principal and the agent's utility can be rewritten as functions of $\tilde{x}$ and a:

$$u(\tilde{x}, a) = \tilde{r}_{(a)} - \tilde{x}_{(a)}, \qquad v(\tilde{x}, a) = \tilde{x}_{(a)} - c_a,$$

which are linear (strictly speaking, affine) in $\tilde{x} \in \tilde{\mathcal{X}}$. Assuming bounded reward $|\tilde{r}_{(a)}| \le R$, we can without loss of generality assume that the expected payment $\tilde{x}_{(a)}$ is bounded by R as well, because otherwise the principal will get negative utility. So, the principal's decision space can be restricted to

$$\tilde{\mathcal{X}} = \big\{ \tilde{x} \mid \exists\, x \in [0, +\infty]^d \text{ such that } \tilde{x}_{(a)} = \sum_{i=1}^{d} p_{ai} x_{(i)} \text{ for every } a \in A \big\} \cap [0, R]^{|A|},$$

which is convex and has bounded diameter (under ℓ_∞ norm)

$$\operatorname{diam}(\tilde{\mathcal{X}}; \ell_\infty) \le \operatorname{diam}([0, R]^{|A|}; \ell_\infty) = R.$$

The utility function $u(\tilde{x}, a)$ is bounded by $2R$ and $(L = 1)$-Lipschitz (under ℓ_∞ norm):

$$|u(\tilde{x}_1, a) - u(\tilde{x}_2, a)| = |\tilde{x}_{1(a)} - \tilde{x}_{2(a)}| \le \max_{a \in A} |\tilde{x}_{1(a)} - \tilde{x}_{2(a)}| = \|\tilde{x}_1 - \tilde{x}_2\|_\infty.$$

Thus, we can apply the theorem for unconstrained generalized principal-agent problems (Theorem 5.4) and the theorems for learning agent (Theorem 5.1 and Theorem 5.2) to obtain:

Corollary 5.5 (Contract design with a learning agent)**.** *Suppose T is sufficiently large such that $\frac{\mathrm{CReg}(T)}{T} < \frac{G}{2}$ and $\frac{\mathrm{CSReg}(T)}{T} < G$, then:*

- *with a contextual no-regret learning agent, the principal can obtain utility at least $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \ge \underline{\mathrm{OBJ}}^{\mathcal{R}}\big(\frac{\mathrm{CReg}(T)}{T}\big) \ge U^* - \frac{4R}{\sqrt{G}}\sqrt{\frac{\mathrm{CReg}(T)}{T}}$.*
- *with a contextual no-swap-regret learning agent, the principal cannot obtain utility more than $\frac{1}{T}\mathbb{E}\big[\sum_{t=1}^{T} u(x^t, a^t)\big] \le \overline{\mathrm{OBJ}}^{\mathcal{D}}\big(\frac{\mathrm{CSReg}(T)}{T}\big) \le U^* + \frac{R}{G}\frac{\mathrm{CSReg}(T)}{T}$.*

Providing the quantitative lower and upper bounds, the above results refine the result in [GKS+24] that the principal can obtain utility at least $U^* - o(1)$ against a no-regret learning agent and no more than $U^* + o(1)$ against a no-swap-regret agent. This again demonstrates the versatility of our general framework.

5.6 Discussion

In summary, our work provides an explicit characterization of the principal's achievable utility in generalized principal-agent problems with a contextual no-swap-regret learning agent. It is an asymmetric range $\big[U^* - O(\sqrt{\frac{\mathrm{CSReg}(T)}{T}}), U^* + O(\frac{\mathrm{CSReg}(T)}{T})\big]$. We show that this conclusion holds in all generalized principal-agent problems where the agent does not have private information, in particular including Bayesian persuasion where the principal is privately informed. As we mentioned in the Introduction, the upper bound $U^* + O(\frac{\mathrm{CSReg}(T)}{T})$ does not hold when the agent has private information or does certain types of no-regret but not no-swap-regret learning. Deriving the exact upper bound in the latter cases is an interesting direction for future work.

Other directions for future work include, for example, relaxing the assumption that the principal has perfect knowledge of the environment – what if both principal and agent are learning players? And what if the environment is non-stationary, like a Markovian environment [JP24] or an adversarial dynamic environment [CHJ20]? In unknown or non-stationary environments, the benchmark U^* needs to be redefined, and a joint design of both players' learning algorithms might be interesting.

5.7 Omitted Proofs in Section 5.2

5.7.1 Details about Contextual No-(Swap-)Regret Algorithms: Proof of Proposition 5.1

Let Alg be an arbitrary no-regret (no-swap-regret) learning algorithm for a multi-armed bandit (MAB) problem with $|A|$ arms. There exist such algorithms with regret $O(\sqrt{T|A|\log|A|})$ (variants of Exp3 [ACBFS02]) and even $O(\sqrt{T|A|})$ (doubling trick + polyINF [AB10]) for any time horizon $T > 0$. By swap-to-external regret reductions, they can be converted to multi-armed bandit algorithms with swap regret $O(\sqrt{T|A|^3\log|A|})$ [BM07] and $O(|A|\sqrt{T})$ [Ito20]. We then convert Alg into a contextual no-regret (contextual no-swap-regret) algorithm, in the following way:

Algorithm 5.1: Convert any MAB algorithm to a contextual MAB algorithm

```
Input: MAB algortihm Alg. Arm set $A$. Context set $S$.
Instantiate $|S|$ copies $\mathrm{Alg}_1, \dots, \mathrm{Alg}_{|S|}$ of Alg, and initialize their round number by
    $t_1 = \cdots = t_{|S|} = 0$.
for round $t = 1, 2, \dots$ do
    Receive context $s^t$. Call $\mathrm{Alg}_{s^t}$ to obtain an action $a^t$.
    Play $a^t$ and obtain feedback (which includes the reward $v^t(a^t)$ of action $a^t$).
    Feed the feedback to $\mathrm{Alg}_{s^t}$. Increase its round number $t_{s^t}$ by 1.
```

Proposition 5.4. *The contextual regret of Algorithm 6 is at most*

$$\mathrm{CReg}(T) \le \max\Big\{ \sum_{s=1}^{|S|} \mathrm{Reg}(T_s) \;\Big|\; T_1 + \cdots + T_{|S|} = T \Big\},$$

where $\mathrm{Reg}(T_s)$ *is the regret of* Alg *for time horizon* T_s*.*

The contextual swap-regret of Algorithm 6 is at most

$$\mathrm{CSReg}(T) \le \max\Big\{ \sum_{s=1}^{|S|} \mathrm{SReg}(T_s) \;\Big|\; T_1 + \cdots + T_{|S|} = T \Big\},$$

where $\mathrm{SReg}(T_s)$ *is the swap-regret of* Alg *for time horizon* T_s*.*

When plugging in $\mathrm{Reg}(T_s) = O(\sqrt{|A|T_s})$*, we obtain* $\mathrm{CReg}(T) \le O(\sqrt{|A||S|T})$*.*

When plugging in $\mathrm{SReg}(T_s) = O(|A|\sqrt{T_s})$*, we obtain* $\mathrm{CSReg}(T) \le O(|A|\sqrt{|S|T})$*.*

Proof. The contextual regret of Algorithm 6 is

$$\begin{aligned}
\mathrm{CReg}(T) &= \max_{d:S\to A} \mathbb{E}\Big[\sum_{t=1}^{T} \big(v^t(d(s^t)) - v^t(a^t)\big)\Big] \\
&= \max_{d:S\to A} \mathbb{E}\Big[\sum_{s=1}^{|S|} \sum_{t:s^t=s} \big(v^t(d(s)) - v^t(a^t)\big)\Big] \\
&\le \sum_{s=1}^{|S|} \max_{a'\in A} \mathbb{E}\Big[\sum_{t:s^t=s} \big(v^t(a') - v^t(a^t)\big)\Big] \\
&\le \sum_{s=1}^{|S|} \mathbb{E}_{T_s}\big[\mathrm{Reg}(T_s)\big] \quad \text{where } T_s \text{ is the number of rounds where } s^t = s \\
&\le \max\Big\{ \sum_{s=1}^{|S|} \mathrm{Reg}(T_s) \;\Big|\; T_1 + \cdots + T_{|S|} = T \Big\}.
\end{aligned}$$

When $\mathrm{Reg}(T_s) = O(\sqrt{|A|T_s})$, by Jensen's inequality we obtain

$$\mathrm{CReg}(T) \leq \sum_{s=1}^{|S|} O(\sqrt{|A|T_s}) \leq O(\sqrt{|A|})\sqrt{|S|}\sqrt{\sum_{s=1}^{|S|} T_s} = O(\sqrt{|A||S|T}).$$

The argument for contextual swap-regret is similar:

$$\begin{aligned}
\mathrm{CSReg}(T) &= \max_{d:S\times A\to A} \mathbb{E}\Big[\sum_{t=1}^{T} \big(v^t(d(s^t,a^t)) - v^t(a^t)\big)\Big] \\
&= \max_{d:S\times A\to A} \mathbb{E}\Big[\sum_{s=1}^{|S|} \sum_{t:s^t=s} \big(v^t(d(s,a^t)) - v^t(a^t)\big)\Big] \\
&\leq \sum_{s=1}^{|S|} \max_{d':A\to A} \mathbb{E}\Big[\sum_{t:s^t=s} \big(v^t(d'(a^t)) - v^t(a^t)\big)\Big] \\
&\leq \sum_{s=1}^{|S|} \mathbb{E}_{T_s}\big[\mathrm{SReg}(T_s)\big] \quad \text{where } T_s \text{ is the number of rounds where } s^t = s \\
&\leq \max\Big\{ \sum_{s=1}^{|S|} \mathrm{SReg}(T_s) \;\Big|\; T_1 + \cdots + T_{|S|} = T \Big\}.
\end{aligned}$$

When $\mathrm{SReg}(T_s) = O(|A|\sqrt{T_s})$, by Jensen's inequality we obtain

$$\mathrm{CSReg}(T) \leq \sum_{s=1}^{|S|} O(|A|\sqrt{T_s}) \leq O(|A|)\sqrt{|S|}\sqrt{\sum_{s=1}^{|S|} T_s} = O(|A|\sqrt{|S|T}).$$

□

5.8 Omitted Proofs in Section 5.3

5.8.1 Proof of Example 5.1

Consider the quantal response model. Let $\gamma = \frac{\log(|A|\lambda)}{\lambda}$. Given signal s, with posterior μ_s, we say an action $a \in A$ is *not* γ-optimal for posterior μ_s if

$$v(\mu_s, a_s^*) - v(\mu_s, a) \geq \gamma$$

where a_s^* is an optimal action for μ_s. The probability that the receiver chooses not γ-optimal action a is at most:

$$\begin{aligned} \frac{\exp(\lambda v(\mu_s, a))}{\sum_{a \in A} \exp(\lambda v(\mu_s, a))} \leq \frac{\exp(\lambda v(\mu_s, a))}{\exp(\lambda v(\mu_s, a_s^*))} &= \exp\Big(- \lambda \big[v(\mu_s, a_s^*) - v(\mu_s, a) \big] \Big) \\ &\leq \exp(-\lambda\gamma) = \frac{1}{|A|\lambda}. \end{aligned}$$

By a union bound, the probability that the receiver chooses any not γ-approxiamtely optimal action is at most $\frac{1}{\lambda}$. So, the expected loss of utility of the receiver due to not taking the optimal action is at most

$$(1 - \frac{1}{\lambda}) \cdot \gamma + \frac{1}{\lambda} \cdot 1 \leq \frac{\log(|A|\lambda) + 1}{\lambda}$$

This means that the quantal response strategy is a $\frac{\log(|A|\lambda)+1}{\lambda}$-best-responding randomized strategy.

Consider inaccurate belief. Given signal s, the receiver has belief μ_s' with total variation distance $d_{\mathrm{TV}}(\mu_s', \mu_s) \leq \varepsilon$ to the true posterior μ_s. For any action $a \in A$, the difference of expected utility of action a under beliefs μ'_s and μ_s is at most ε:

$$\left|\mathbb{E}_{\omega\sim\mu'_s}[v(\omega,a)] - \mathbb{E}_{\omega\sim\mu_s}[v(\omega,a)]\right| \le d_{\mathrm{TV}}(\mu'_s,\mu_s) \le \varepsilon.$$

So, the optimal action for μ'_s is a 2ε-optimal action for μ_s. This means that the receiver strategy is a deterministic 2ε-best-responding strategy.

5.9 Omitted Proofs in Section 5.4

5.9.1 Proof of Claim 5.3

If no $G > 0$ satisfies the claim, then there must exist an $a_0 \in A$ such that for all $x \in \mathcal{X}$, $v(a_0,\mu) - v(a',\mu) \le 0$ for some $a' \in A \setminus \{a_0\}$. Namely,

$$\max_{x\in\mathcal{X}} \min_{a'\in A\setminus\{a_0\}} \left\{v(x,a_0) - v(x,a')\right\} \le 0.$$

Then, by the minimax theorem, we have

$$\min_{\alpha'\in\Delta(A\setminus\{a_0\})} \max_{x\in\mathcal{X}} \left\{v(x,a_0) - v(x,\alpha')\right\} = \max_{x\in\mathcal{X}} \min_{a'\in A\setminus\{a_0\}} \left\{v(x,a_0) - v(x,a')\right\} \le 0.$$

This means that a_0 is weakly dominated by some mixed action $\alpha' \in \Delta(A\setminus\{a_0\})$, violating Assumption 5.1.

5.9.2 Proof of Example 5.2

We use the probability $\mu \in [0,1]$ of the Good state to represent a belief (so the probability of Bad state is $1-\mu$).

First, the sender's optimal utility when the receiver exactly best responds is $2\mu_0$:

$$U^* = 2\mu_0.$$

This is achieved by the signaling scheme π^* that decomposes the prior μ_0 into two posteriors $\mu_a = \frac{1}{2}$ and $\mu_b = 0$ with probability $2\mu_0$ and $1 - 2\mu_0$ respectively, with the receiver taking action a under posterior μ_a and b under μ_b.

Under signaling scheme π^*, suppose the receiver takes action a with probability $\frac{\delta}{1-2\mu_0}$ under posterior μ_b. Compared to best-responding, the receiver loses utility $(1 - 2\mu_0) \cdot \frac{\delta}{1-2\mu_0} \cdot 1 = \delta$ in expectation, so the receiver is δ-approximate best responding. The sender obtains $(1 - 2\mu_0) \cdot \frac{\delta}{1-2\mu_0} \cdot 1 = \delta$ more utility in this case. So, we have proven the first claim $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq U^* + \delta$.

In the remaining, we will prove the second claim $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_\pi \min_{\rho \in \mathcal{R}_\delta(\pi)} U(\pi, \rho) \leq U^* - 2\sqrt{2\mu_0\delta}$. Consider any signaling scheme of the sender, $\pi = \{(\pi_s, \mu_s)\}_{s \in S}$, which is a decomposition of the prior μ_0 into $|S|$ posteriors $\mu_s \in [0, 1]$ such that $\sum_{s \in S} \pi_s \mu_s = \mu_0$. Let $\rho : S \to \Delta(A)$ be a randomized strategy of the receiver, where $\rho(a|s)$ and $\rho(b|s)$ denote the probability that the receiver takes actions a and b under signal s. The sender's expected utility under π and ρ is:

$$U(\pi, \rho) = \sum_{s \in S} \pi_s \big[\rho(a|s) \cdot 1 + \rho(b|s) \cdot 0\big] = \sum_{s \in S} \pi_s \rho(a|s). \tag{5.14}$$

The receiver's utility when taking action a at posterior μ_s is $\mu_s \cdot 1 + (1 - \mu_s) \cdot (-1) = 2\mu_s - 1$. So, the receiver's expected utility under π and ρ is

$$V(\pi, \rho) = \sum_{s \in S} \pi_s \big[\rho(a|s) \cdot (2\mu_s - 1) + \rho(b|s) \cdot 0\big] = \sum_{s \in S} \pi_s \rho(a|s)(2\mu_s - 1). \tag{5.15}$$

Clearly, the receiver's best response ρ^* is to take action a with certainty if and only if

$\mu_s > \frac{1}{2}$, with expected utility

$$V(\pi, \rho^*) = \sum_{s:\mu_s > \frac{1}{2}} \pi_s(2\mu_s - 1). \tag{5.16}$$

To find $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_\pi \min_{\rho \in \mathcal{R}_\delta(\pi)} U(\pi, \rho)$, we fix any π and solve the inner optimization problem (minimizing the sender's utility) regarding ρ:

$$\begin{aligned}
\min_\rho \quad & U(\pi, \rho) = \sum_{s \in S} \pi_s \rho(a|s) \\
\text{s.t.} \quad & \rho \in \mathcal{R}_\delta(\pi) \quad \iff \quad \delta \geq V(\pi, \rho^*) - V(\pi, \rho) \\
& \qquad\qquad\qquad\qquad = \sum_{s:\mu_s > \frac{1}{2}} \pi_s(2\mu_s - 1) - \sum_{s \in S} \pi_s \rho(a|s)(2\mu_s - 1).
\end{aligned}$$

Without loss of generality, we can assume that the solution ρ satisfies $\rho(a|s) = 0$ whenever $\mu_s \leq \frac{1}{2}$ (if $\rho(a|s) > 0$ for some $\mu_s \leq \frac{1}{2}$, then making $\rho(a|s)$ to be 0 can decrease the objective $\sum_{s \in S} \pi_s \rho(a|s)$ while still satisfying the constraint). So, the optimization problem can be simplified to:

$$\begin{aligned}
\min_\rho \quad & U(\pi, \rho) = \sum_{s:\mu_s > \frac{1}{2}} \pi_s \rho(a|s) \\
\text{s.t.} \quad & \delta \geq \sum_{s:\mu_s > \frac{1}{2}} \pi_s(2\mu_s - 1) - \sum_{s:\mu_s > \frac{1}{2}} \pi_s \rho(a|s)(2\mu_s - 1) \\
& \quad = \sum_{s:\mu_s > \frac{1}{2}} \pi_s(2\mu_s - 1)(1 - \rho(a|s)), \\
& \rho(a|s) \in [0, 1], \quad \forall s \in S : \mu_s > \tfrac{1}{2}.
\end{aligned}$$

We note that this is a fractional knapsack linear program, which has a greedy solution (e.g., [KV12]): sort the signals with $\mu_s > \frac{1}{2}$ in increasing order of $2\mu_s - 1$ (equivalently, increasing order of μ_s); label those signals by $s = 1, \ldots, n$; find the first position k for which $\sum_{s=1}^{k} \pi_s(2\mu_s - 1) > \delta$:

$$k = \min\big\{j : \sum_{s=1}^{j} \pi_s(2\mu_s - 1) > \delta\big\};$$

then, an optimal solution ρ is given by:

$$\begin{cases} \rho(a|s) = 0 & \text{for } s = 1, \dots, k-1; \\ \rho(a|k) = 1 - \frac{\delta - \sum_{s=1}^{k-1} \pi_s(2\mu_s - 1)}{\pi_k(2\mu_k - 1)} & \text{for } s = k; \\ \rho(a|s) = 1 & \text{for } s = k+1, \dots, n. \end{cases}$$

The objective value (sender's expected utility) of the above solution ρ is

$$\begin{aligned} U(\pi, \rho) &= \sum_{s:\mu_s > \frac{1}{2}} \pi_s \rho(a|s) \\ &= \pi_k\Big(1 - \frac{\delta - \sum_{s=1}^{k-1} \pi_s(2\mu_s - 1)}{\pi_k(2\mu_k - 1)}\Big) + \sum_{s=k+1}^{n} \pi_s \\ &= \sum_{s=k}^{n} \pi_s - \frac{\delta}{2\mu_k - 1} + \sum_{s=1}^{k-1} \frac{\pi_s(2\mu_s - 1)}{2\mu_k - 1}. \end{aligned}$$

Since the signaling scheme π must satisfy $\sum_{s\in S} \pi_s \mu_s = \mu_0$, we have

$$\begin{aligned} &\mu_0 = \sum_{s\in S} \pi_s\mu_s \geq \sum_{s=1}^{n} \pi_s\mu_s = \sum_{s=1}^{k-1} \pi_s\mu_s + \sum_{s=k}^{n} \pi_s\mu_s \geq \sum_{s=1}^{k-1} \pi_s\mu_s + \sum_{s=k}^{n} \pi_s\mu_k \\ \implies \quad &\sum_{s=k}^{n} \pi_s \leq \frac{\mu_0 - \sum_{s=1}^{k-1} \pi_s\mu_s}{\mu_k}. \end{aligned}$$

So,

$$U(\pi,\rho) \le \frac{\mu_0 - \sum_{s=1}^{k-1}\pi_s\mu_s}{\mu_k} - \frac{\delta}{2\mu_k - 1} + \sum_{s=1}^{k-1}\frac{\pi_s(2\mu_s-1)}{2\mu_k-1}$$
$$= \frac{\mu_0}{\mu_k} - \frac{\delta}{2\mu_k-1} + \sum_{s=1}^{k-1}\pi_s\Big(\frac{2\mu_s-1}{2\mu_k-1} - \frac{\mu_s}{\mu_k}\Big).$$

Since $\frac{2\mu_s-1}{2\mu_k-1} - \frac{\mu_s}{\mu_k} = \frac{\mu_s-\mu_k}{(2\mu_s-1)\mu_k} \le 0$ for any $s \le k-1$, we get

$$U(\pi,\rho) \le \frac{\mu_0}{\mu_k} - \frac{\delta}{2\mu_k-1} = f(\mu_k).$$

We find the maximal value of $f(\mu_k) = \frac{\mu_0}{\mu_k} - \frac{\delta}{2\mu_k-1}$. Take its derivative:

$$f'(\mu_k) = -\frac{\mu_0}{\mu_k^2} + \frac{2\delta}{(2\mu_k-1)^2} = \frac{\big[(\sqrt{2\delta}+2\sqrt{\mu_0})\mu_k - \sqrt{\mu}_0\big]\cdot\big[(\sqrt{2\delta}-2\sqrt{\mu_0})\mu_k + \sqrt{\mu}_0\big]}{\mu_k^2(2\mu_k-1)^2},$$

which has two roots $\frac{\sqrt{\mu}_0}{\sqrt{2\delta}+2\sqrt{\mu_0}} < \frac{1}{2}$ and $\frac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}} \in (\frac{1}{2}, 1)$ when $0 < \delta < \frac{\mu_0}{2}$. So, $f(x)$ is increasing in $[\frac{1}{2}, \frac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}})$ and decreasing in $(\frac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}}, 1]$. Since $\mu_k > \frac{1}{2}$, $f(\mu_k)$ is maximized at $\mu_k = \frac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}}$. This implies

$$U(\pi,\rho) \le f\big(\tfrac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}}\big) = \frac{\mu_0}{\sqrt{\mu}_0}(2\sqrt{\mu}_0 - \sqrt{2\delta}) - \frac{\delta}{2\frac{\sqrt{\mu}_0}{2\sqrt{\mu_0}-\sqrt{2\delta}} - 1} = 2\mu_0 - 2\sqrt{2\mu_0\delta} + \delta.$$

This holds for any π. So, $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_\pi \min_{\rho\in\mathcal{R}_\delta(\pi)} U(\pi,\rho) \le U^* - 2\sqrt{2\mu_0\delta} + \delta = U^* - \Omega(\sqrt{\delta})$.

5.9.3 Proof of Theorems 5.4 and 5.5

Lower bounds on $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ and upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$. First, we prove the lower bounds on $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ and the upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ in Theorems 5.4 and 5.5, given by the following two lemmas:

Lemma 5.2. *In an unconstrained generalized principal-agent problem,* $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \geq U^* - \mathrm{diam}(\mathcal{X}) L \frac{\delta}{G}$.

With constraint $\sum_{s\in S} \pi_s x_s \in \mathcal{C}$, $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \geq U^* - \big(\mathrm{diam}(\mathcal{X}) L + 2B \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})}\big) \frac{\delta}{G}$.

Lemma 5.3. *In an unconstrained generalized principal-agent problem,* $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \leq U^* + \mathrm{diam}(\mathcal{X}) L \frac{\delta}{G}$.

With constraint $\sum_{s\in S} \pi_s x_s \in \mathcal{C}$, $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \leq U^* + \big(\mathrm{diam}(\mathcal{X}) L + 2B \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})}\big) \frac{\delta}{G}$.

The proofs of Lemmas 5.2 and 5.3 are similar and given in Section 5.9.4 and 5.9.5. The main idea to prove Lemma 5.3 is the following. Let (π, ρ) be any pair of principal's strategy and agent's δ-best-responding strategy. We perturb the principal's strategy π slightly to be a strategy π' for which ρ is *exactly* best-responding (such a perturbation is possible due to Assumption 5.1). Since ρ is best-responding to π', the pair (π', ρ) cannot give the principal a higher utility than U^* (which is the optimal principal utility under the best-response model). This means that the original pair (π, ρ) cannot give the principal a utility much higher than U^*, implying an upper bound on $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$.

Upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ imply upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$. Then, because $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \leq \overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$, we immediately obtain the upper bounds on $\overline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ in the two theorems.

Lower bounds for $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ imply lower bounds for $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ Finally, we show that the lower bounds for $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta)$ imply the lower bounds for $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$, using the following lemma:

Lemma 5.4. *For any* $\delta \geq 0, \Delta > 0$, $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq \underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta) - \frac{2B\delta}{\Delta}$.

The proof of this lemma is in Section 5.9.6.

Using Lemma 5.4 with $\Delta = \sqrt{\frac{2BG\delta}{\mathrm{diam}(\mathcal{X})L}}$ and the lower bound for $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta)$ in Lemma 5.2 for the unconstrained case, we obtain:

$$\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq \underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta) - \tfrac{2B\delta}{\Delta} \geq U^* - \mathrm{diam}(\mathcal{X})L\tfrac{\Delta}{G} - \tfrac{2B\delta}{\Delta} = U^* - 2\sqrt{\tfrac{2BL}{G}\mathrm{diam}(\mathcal{X})\delta},$$

which gives the lower bound for $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ in Theorem 5.4.

Using Lemma 5.4 with $\Delta = \sqrt{\frac{2BG\delta}{L\mathrm{diam}(\mathcal{X})+2B\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}}}$ and the lower bound for $\underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta)$ in Lemma 5.2 for the constrained case, we obtain:

$$\begin{aligned}\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \geq \underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta) - \tfrac{2B\delta}{\Delta} &\geq U^* - \big(\mathrm{diam}(\mathcal{X})L + 2B\tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\tfrac{\Delta}{G} - \tfrac{2B\delta}{\Delta} \\ &= U^* - 2\sqrt{\tfrac{2B}{G}\big(\mathrm{diam}(\mathcal{X})L + 2B\tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\delta}.\end{aligned}$$

This proves the lower bound for $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta)$ in Theorem 5.5.

5.9.4 Proof of Lemma 5.2

Let (π, ρ) be a pair of principal strategy and agent strategy that achieves the optimal principal utility with an exactly-best-responding agent, namely, $U(\pi, \rho) = U^*$. Without loss of generality ρ can be assumed to be deterministic, $\rho : S \to A$. The strategy π consists of pairs $\{(\pi_s, x_s)\}_{s\in S}$ that satisfy

$$\sum_{s\in S} \pi_s x_s =: \mu_0 \in \mathcal{C}, \tag{5.17}$$

and the action $a = \rho(s)$ is optimal for the agent with respect to x_s. We will construct another principal strategy π' such that, even if the agent chooses the worst δ-best-responding strategy to π', the principal can still obtain utility arbitrarily close to $U^* - \big(L\mathrm{diam}(\mathcal{X}; \ell_1) + 2B\frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C},\partial\mathcal{X})}\big)\frac{\delta}{G}$.

To construct π' we do the following: For each signal $s \in S$, with corresponding action

$a = \rho(s)$, by Claim 5.3 there exists $y_a \in \mathcal{X}$ such that $v(y_a, a) - v(y_a, a') \geq G$ for any $a' \neq a$. Let $\theta = \frac{\delta}{G} + \varepsilon \in [0, 1]$ for arbitrarily small $\varepsilon > 0$, and let $\tilde{x}_s$ be the convex combination of x_s and $y_{\rho(s)}$ with weights $1 - \theta, \theta$:

$$\tilde{x}_s = (1 - \theta)x_s + \theta y_{\rho(s)}. \tag{5.18}$$

We note that $a = \rho(s)$ is the agent's optimal action for $\tilde{x}_s$ and moreover it is better than any other action $a' \neq a$ by more than δ:

$$\begin{aligned} v(\tilde{x}_s, a) - v(\tilde{x}_s, a') &= (1 - \theta)\big[\underbrace{v(x_s, a) \ - \ v(x_s, a')}_{\geq 0 \text{ because } a = \rho(s) \text{ is optimal for } x_s} \big] \ + \ \theta\big[\underbrace{v(y_a, a) - v(y_a, a')}_{\geq G \text{ by our choice of } y_a} \big] \\ &\geq 0 + \theta G > \tfrac{\delta}{G} G = \delta. \end{aligned} \tag{5.19}$$

Let μ' be the convex combination of $\{\tilde{x}_s\}_{s \in S}$ with weights $\{\pi_s\}_{s \in S}$:

$$\mu' = \sum_{s \in S} \pi_s \tilde{x}_s. \tag{5.20}$$

Note that μ' might not satisfy the constraint $\mu' \in \mathcal{C}$. So, we want to find another vector $z \in \mathcal{X}$ and a coefficient $\eta \in [0, 1]$ such that

$$(1 - \eta)\mu' + \eta z \in \mathcal{C}. \tag{5.21}$$

(If μ' already satisfies $\mu' \in \mathcal{C}$, then let $\eta = 0$.) To do this, we consider the ray starting from μ' pointing towards μ_0: $\{\mu' + t(\mu_0 - \mu') \mid t \geq 0\}$. Let z be the intersection of the ray with the boundary of $\mathcal{X}$:

$$z = \mu' + t^*(\mu_0 - \mu'), \qquad t^* = \arg\max\{t \geq 0 \mid \mu' + t(\mu_0 - \mu') \in \mathcal{X}\}.$$

Then, rearranging $z = \mu' + t^*(\mu_0 - \mu')$, we get

$$\tfrac{1}{t^*}(z - \mu') = \mu_0 - \mu' \quad\iff\quad (1 - \tfrac{1}{t^*})\mu' + \tfrac{1}{t^*}z = \mu_0 \in \mathcal{C},$$

which satisfies (5.21) with $\eta = \frac{1}{t^*}$. We then give an upper bound on $\eta = \frac{1}{t*}$:

Claim 5.4. $\eta = \frac{1}{t^*} \le \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})}\theta$.

Proof. On the one hand,

$$\begin{aligned}\|\mu_0 - \mu'\| &= \Big\|\sum_{s\in S}\pi_s x_s - \sum_{s\in S}\pi_s\tilde{x}_s\Big\| = \Big\|\sum_{s\in S}\pi_s\theta(y_{\rho(s)} - x_s)\Big\| \\ &\le \theta\sum_{s\in S}\pi_s\big\|y_{\rho(s)} - x_s\big\| \le \theta\sum_{s\in S}\pi_s\cdot\mathrm{diam}(\mathcal{X}) = \theta\cdot\mathrm{diam}(\mathcal{X}).\end{aligned}$$

On the other hand, because $z - \mu'$ and $\mu_0 - \mu'$ are in the same direction, we have

$$\|z - \mu'\| = \|z - \mu_0\| + \|\mu_0 - \mu'\| \ge \|z - \mu_0\| \ge D(\mathcal{C}, \partial\mathcal{X})$$

because μ_0 is in $\mathcal{C}$ and z is on the boundary of $\mathcal{X}$. Therefore, $\eta = \frac{1}{t^*} = \frac{\|\mu_0 - \mu'\|}{\|z - \mu'\|} \le \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})}\theta$. □

The convex combinations (5.21) (5.20) define a new principal strategy π' with $|S| + 1$ signals, consisting of $\tilde{x}_s$ with probability $(1 - \eta)\pi_s$ and z with probability η, satisfying $\sum_{s\in S}(1 - \eta)\pi_s\tilde{x}_s + \eta z = \mu_0 \in \mathcal{C}$. Consider the agent's worst (for the principal) δ-best-responding strategies ρ' to π':

$$\rho' \in \operatorname*{arg\,min}_{\rho\in\mathcal{D}_\delta(\pi')} U(\pi', \rho).$$

We note that $\rho'(\tilde{x}_s)$ must be equal to $\rho(s)$ for each $s \in S$. This is because $a = \rho(s)$ is strictly better than any other action $a' \neq a$ by a margin of δ (5.19), so a is the only

δ-optimal action for $\tilde{x}_s$.

Then, the principal's expected utility under π' and ρ' is

$$\begin{aligned}
U(\pi', \rho') &\overset{(5.21),(5.20)}{=} (1-\eta) \sum_{s \in S} \pi_s u(\tilde{x}_s, \rho'(\tilde{x}_s)) \ + \ \eta u(z, \rho'(z)) \\
&\geq (1-\eta) \sum_{s \in S} \pi_s u(\tilde{x}_s, \rho(s)) \ - \ \eta B \\
&\geq (1-\eta) \sum_{s \in S} \pi_s \Big(u(x_s, \rho(s)) - L \underbrace{\|\tilde{x}_s - x_s\|}_{= \|\theta(y_{\rho(s)} - x_s)\| \leq \theta \mathrm{diam}(\mathcal{X})} \Big) \ - \ \eta B \\
&\geq (1-\eta) U(\pi, \rho) - L\theta \mathrm{diam}(\mathcal{X}) - \eta B \\
&\geq U(\pi, \rho) - L\theta \mathrm{diam}(\mathcal{X}) - 2\eta B \\
\text{by Claim 5.4} &\geq U(\pi, \rho) - L\theta \mathrm{diam}(\mathcal{X}) - 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \theta \\
&= U(\pi, \rho) - \big(L \mathrm{diam}(\mathcal{X}) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \big(\tfrac{\delta}{G} + \varepsilon \big) \\
&= U^* - \big(L \mathrm{diam}(\mathcal{X}) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \tfrac{\delta}{G} - O(\varepsilon).
\end{aligned}$$

So, we conclude that

$$\begin{aligned}
\underline{\mathrm{OBJ}}^{\mathcal{D}}(\delta) \ = \ \sup_{\pi} \min_{\rho \in \mathcal{D}_\delta(\pi)} U(\pi, \rho) \ &\geq \ \min_{\rho \in \mathcal{D}_\delta(\pi')} U(\pi', \rho) \\
&= \ U(\pi', \rho') \ \geq \ U^* - \big(L \mathrm{diam}(\mathcal{X}) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \tfrac{\delta}{G} - O(\varepsilon).
\end{aligned}$$

Letting $\varepsilon \to 0$ finishes the proof for the case with the constraint $\sum_{s \in S} \pi_s x_s \in \mathcal{C}$.

The case without $\sum_{s \in S} \pi_s x_s \in \mathcal{C}$ is proved by letting $\eta = 0$ in the above argument.

5.9.5 Proof of Lemma 5.3

Let π be a principal strategy and $\rho \in \mathcal{R}_\delta(\pi)$ be a δ-best-responding randomized strategy of the agent. The principal strategy π consists of pairs $\{(\pi_s, x_s)\}_{s\in S}$ with

$$\sum_{s\in S} \pi_s x_s =: \mu_0 \in \mathcal{C}. \tag{5.22}$$

At signal s, the agent takes action a with probability $\rho(a|s)$. Let $\delta_{s,a}$ be the "suboptimality" of action a with respect to x_s:

$$\delta_{s,a} = \max_{a'\in A} \big\{ v(x_s, a') - v(x_s, a) \big\}. \tag{5.23}$$

By Claim 5.3, for action a there exists $y_a \in \mathcal{X}$ such that $v(y_a, a) - v(y_a, a') \geq G$ for any $a' \neq a$. Let $\theta_{s,a} = \frac{\delta_{s,a}}{G+\delta_{s,a}} \in [0,1]$ and let $\tilde{x}_{s,a}$ be the convex combination of x_s and y_a with weights $1-\theta_{s,a}$ and $\theta_{s,a}$:

$$\tilde{x}_{s,a} = (1-\theta_{s,a})x_s + \theta_{s,a} y_a. \tag{5.24}$$

Claim 5.5. *We have two useful claims regarding $\tilde{x}_{s,a}$ and $\theta_{s,a}$:*

(1) a is an optimal action for the agent with respect to $\tilde{x}_{s,a}$: $v(\tilde{x}_{s,a}, a) - v(\tilde{x}_{s,a}, a') \geq 0, \forall a' \in A$.

(2) $\sum_{s\in S}\sum_{a\in A} \pi_s \rho(a|s)\theta_{s,a} \leq \frac{\delta}{G}$.

Proof. (1) For any $a' \neq a$, by the definition of $\tilde{x}_{s,a}$ and $\theta_{s,a}$,

$$\begin{aligned} v(\tilde{x}_{s,a}, a) - v(\tilde{x}_{s,a}, a') \; &= \; (1-\theta_{s,a})\big[v(x_s, a) - v(x_s, a')\big] \; + \; \theta_{s,a}\big[v(y_a, a) - v(y_a, a')\big] \\ &\geq \; (1-\theta_{s,a})(-\delta_{s,a}) \; + \; \theta_{s,a} G \; = \; \tfrac{G}{G+\delta_{s,a}}(-\delta_{s,a}) + \tfrac{\delta_{s,a}}{G+\delta_{s,a}} G \; = \; 0. \end{aligned}$$

(2) By the condition that ρ is a δ-best-response to π, we have

$$\begin{aligned} \delta \;\geq\; \max_{\rho^*: S \to A} V(\pi, \rho^*) - V(\pi, \rho) \;&=\; \sum_{s \in S} \pi_s \Big(\max_{a' \in A} \big\{ v(x_s, a') \big\} - \sum_{a \in A} \rho(a|s) v(x_s, a) \Big) \\ = \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) \max_{a' \in A} \big\{ v(x_s, a') - v(x_s, a) \big\} \;&=\; \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) \delta_{s,a}. \end{aligned}$$

So, $\sum_{s\in S}\sum_{a\in A}\pi_s\rho(a|s)\theta_{s,a} = \sum_{s\in S}\sum_{a\in A}\pi_s\rho(a|s)\frac{\delta_{s,a}}{G+\delta_{s,a}} \leq \sum_{s\in S}\sum_{a\in A}\pi_s\rho(a|s)\frac{\delta_{s,a}}{G} \leq \frac{\delta}{G}$. □

We let μ' be the convex combination of $\{\tilde{x}_{s,a}\}_{s,a\in S\times A}$ with weights $\{\pi_s\rho(a|s)\}_{s,a\in S\times A}$:

$$\mu' = \sum_{s,a \in S \times A} \pi_s \rho(a|s) \tilde{x}_{s,a}. \tag{5.25}$$

Note that μ' might not satisfy the constraint $\mu' \in \mathcal{C}$. So, we want to find another vector $z \in \mathcal{X}$ and a coefficient $\eta \in [0, 1]$ such that

$$(1 - \eta)\mu' + \eta z \in \mathcal{C}. \tag{5.26}$$

(If μ' already satisfies $\mu' \in \mathcal{C}$, then let $\eta = 0$.) To do this, we consider the ray pointing from μ' to μ_0: $\{\mu' + t(\mu_0 - \mu') \mid t \geq 0\}$. Let z be the intersection of the ray with the boundary of $\mathcal{X}$:

$$z = \mu' + t^*(\mu_0 - \mu'), \qquad t^* = \arg\max\{t \geq 0 \mid \mu' + t(\mu_0 - \mu') \in \mathcal{X}\}.$$

Then, rearranging $z = \mu' + t^*(\mu_0 - \mu')$, we get

$$\tfrac{1}{t^*}(z - \mu') = \mu_0 - \mu' \quad\iff\quad (1 - \tfrac{1}{t^*})\mu' + \tfrac{1}{t^*} z = \mu_0 \in \mathcal{C},$$

which satisfies (5.26) with $\eta = \frac{1}{t^*}$. We then give an upper bound on $\eta = \frac{1}{t*}$:

Claim 5.6. $\eta = \frac{1}{t^*} \le \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})} \frac{\delta}{G}$.

Proof. On the one hand,

$$\begin{aligned}\|\mu_0 - \mu'\| &= \Big\| \sum_{s\in S} \pi_s x_s - \sum_{s\in S}\sum_{a\in A} \pi_s \rho(a|s) \tilde{x}_{s,a} \Big\| = \Big\| \sum_{s\in S}\sum_{a\in A} \pi_s \rho(a|s) \theta_{s,a} (y_a - x_s) \Big\| \\ &\le \sum_{s\in S}\sum_{a\in A} \pi_s \rho(a|s) \theta_{s,a} \big\| y_a - x_s \big\| \le \sum_{s\in S}\sum_{a\in A} \pi_s \rho(a|s) \theta_{s,a} \mathrm{diam}(\mathcal{X}) \overset{\text{Claim 5.5}}{\le} \mathrm{diam}(\mathcal{X}) \tfrac{\delta}{G}.\end{aligned}$$

On the other hand, because $z - \mu'$ and $\mu_0 - \mu'$ are in the same direction, we have

$$\|z - \mu'\| = \|z - \mu_0\| + \|\mu_0 - \mu'\| \ge \|z - \mu_0\| \ge D(\mathcal{C}, \partial\mathcal{X})$$

because μ_0 is in $\mathcal{C}$ and z is on the boundary of $\mathcal{X}$. Therefore, $\eta = \frac{1}{t^*} = \frac{\|\mu_0 - \mu'\|}{\|z - \mu'\|} \le \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial\mathcal{X})} \frac{\delta}{G}$.

□

The convex combinations (5.26) (5.25) define a new principal strategy π' (with $|S| \times |A| + 1$ signals) consisting of $\tilde{x}_{s,a}$ with probability $(1-\eta)\pi_s\rho(a|s)$ and z with probability η. Consider the following deterministic agent strategy ρ' in response to π': for $\tilde{x}_{s,a}$, take action $\rho'(\tilde{x}_{s,a}) = a$; for z, take any action that is optimal for z. We note that ρ' is a best-response to π', $\rho' \in \mathcal{R}_0(\pi')$, because, according to Claim 5.5, a is an optimal action with respect to $\tilde{x}_{s,a}$.

Then, consider the principal's utility under π' and ρ':

$$
\begin{aligned}
U(\pi', \rho') &\overset{(5.26),(5.25)}{=} (1-\eta) \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) u(\tilde{x}_{s,a}, \rho'(\tilde{x}_{s,a})) \ + \ \eta u(z, \rho'(z)) \\
&\geq \ (1-\eta) \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) u(\tilde{x}_{s,a}, a) \ - \ \eta B \\
&\geq \ (1-\eta) \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) \Big(u(x_s, a) - L \underbrace{\|\tilde{x}_s - x_s\|}_{= \|\theta_{s,a}(y_a - x_s)\| \leq \theta_{s,a} \mathrm{diam}(\mathcal{X})} \Big) \ - \ \eta B \\
&\geq \ (1-\eta) U(\pi, \rho) \ - \ L \mathrm{diam}(\mathcal{X}) \sum_{s \in S} \sum_{a \in A} \pi_s \rho(a|s) \theta_{s,a} \ - \ \eta B \\
\text{(Claim 5.5)} &\geq \ U(\pi, \rho) - L \mathrm{diam}(\mathcal{X}) \tfrac{\delta}{G} - 2\eta B \\
\text{(Claim 5.6)} &\geq \ U(\pi, \rho) - \big(L \mathrm{diam}(\mathcal{X}) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \tfrac{\delta}{G}.
\end{aligned}
$$

Rearranging, $U(\pi, \rho) \leq U(\pi', \rho') + \big(L \mathrm{diam}(\mathcal{X}) + 2B \frac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \frac{\delta}{G}$. Note that this argument holds for any pair (π, ρ) that satisfies $\rho \in \mathcal{R}_\delta(\pi)$. And recall that $\rho' \in \mathcal{R}_0(\pi')$. So, we conclude that

$$
\begin{aligned}
\overline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) \ = \ \max_{\pi} \max_{\rho \in \mathcal{R}_\delta(\pi)} U(\pi, \rho) \ &\leq \ \max_{\pi'} \max_{\rho' \in \mathcal{R}_0(\pi)} U(\pi', \rho') + \big(L \mathrm{diam}(\mathcal{X}; \ell_1) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \tfrac{\delta}{G} \\
&= \ U^* + \big(L \mathrm{diam}(\mathcal{X}; \ell_1) + 2B \tfrac{\mathrm{diam}(\mathcal{X})}{D(\mathcal{C}, \partial \mathcal{X})} \big) \tfrac{\delta}{G}.
\end{aligned}
$$

This proves the case with the constraint $\sum_{s \in S} \pi_s x_s \in \mathcal{C}$.

The case without $\sum_{s \in S} \pi_s x_s \in \mathcal{C}$ is proved by letting $\eta = 0$ in the above argument.

5.9.6 Proof of Lemma 5.4

Let $A_\Delta(x) = \big\{ a \in A \mid v(x, a) \geq v(x, a') - \Delta, \forall a' \in A \big\}$ be the set of Δ-optimal actions of the agent in response to principal decision $x \in \mathcal{X}$. The proof of Lemma 5.4 uses another lemma that relates the principal utility under a randomized δ-best-responding agent strategy $\rho \in \mathcal{R}_\delta(\pi)$ and that under an agent strategy ρ' that only randomizes over

$A_\Delta(x_s)$.

Lemma 5.5. *Let $\pi = \{(\pi_s, x_s)\}_{s\in S}$ be a principal strategy and $\rho \in \mathcal{R}_\delta(\pi)$ be a randomized δ-best-response to π. For any $\Delta > 0$, there exists an agent strategy $\rho' : s \mapsto \Delta(A_\Delta(x_s))$ that randomizes over Δ-optimal actions only for each x_s, such that the principal's utility under ρ' and ρ satisfies: $\big|U(\pi,\rho') - U(\pi,\rho)\big| \le \frac{2B\delta}{\Delta}$.*

Proof. Let $a_s^* = \max_{a\in A} v(x_s, a)$ be the agent's optimal action for x_s. Let $\overline{A_\Delta(x_s)} = A \setminus A_\Delta(x_s)$ be the set of actions that are not Δ-optimal for x_s. By the definition that $\rho \in \mathcal{R}_\delta(\pi)$ is a δ-best-response to π, we have

$$\begin{aligned}
\delta &\ge \sum_{s\in S} \pi_s \Big[v(x_s, a_s^*) - \sum_{a\in A} \rho(a|s) v(x_s, a) \Big] \\
&= \sum_{s\in S} \pi_s \Big(\sum_{a\in A_\Delta(x_s)} \rho(a|s) \big[\underbrace{v(x_s, a_s^*) - v(x_s, a)}_{\ge 0} \big] + \sum_{a\in \overline{A_\Delta(x_s)}} \rho(a|s) \big[\underbrace{v(x_s, a_s^*) - v(x_s, a)}_{>\Delta} \big] \Big) \\
&\ge 0 + \Delta \sum_{s\in S} \pi_s \sum_{a\in \overline{A_\Delta(x_s)}} \rho(a|s) \\
&= \Delta \sum_{s\in S} \pi_s \rho(\overline{A_\Delta(x_s)} \,|\, s).
\end{aligned}$$

Rearranging,

$$\sum_{s\in S} \pi_s \rho(\overline{A_\Delta(x_s)} \,|\, s) \le \frac{\delta}{\Delta}. \tag{5.27}$$

Then, we consider the randomized strategy ρ' that, for each s, chooses each action $a \in A_\Delta(x_s)$ with the conditional probability that ρ chooses a given $a \in A_\Delta(x_s)$:

$$\rho'(a \,|\, s) = \frac{\rho(a \,|\, s)}{\rho(A_\Delta(x_s) \,|\, s)}.$$

The sender's utility under ρ' is:

$$U(\pi, \rho') = \sum_{s \in S} \pi_s \sum_{a \in A_\Delta(x_s)} \frac{\rho(a \mid s)}{\rho(A_\Delta(x_s) \mid s)} u(x_s, a).$$

The sender's utility under ρ is

$$U(\pi, \rho) = \sum_{s \in S} \pi_s \sum_{a \in A_\Delta(x_s)} \rho(a \mid s) u(x_s, a) \; + \; \sum_{s \in S} \pi_s \sum_{a \in \overline{A_\Delta(x_s)}} \rho(a \mid s) u(x_s, a)$$

Taking the difference between the two utilities, we get

$$\begin{aligned}
&\left| U(\pi, \rho') - U(\pi, \rho) \right| \\
&\leq \Big| \sum_{s \in S} \pi_s \Big(\frac{1}{\rho(A_\Delta(x_s) \mid s)} - 1 \Big) \sum_{a \in A_\Delta(x_s)} \rho(a \mid s) u(x_s, a) \Big| \; + \; \Big| \sum_{s \in S} \pi_s \sum_{a \in \overline{A_\Delta(x_s)}} \rho(a \mid s) u(x_s, a) \Big| \\
&= \Big| \sum_{s \in S} \pi_s \frac{1 - \rho(A_\Delta(x_s) \mid s)}{\rho(A_\Delta(x_s) \mid s)} \sum_{a \in A_\Delta(x_s)} \rho(a \mid s) u(x_s, a) \Big| \; + \; \Big| \sum_{s \in S} \pi_s \sum_{a \in \overline{A_\Delta(x_s)}} \rho(a \mid s) u(x_s, a) \Big| \\
&\leq \sum_{s \in S} \pi_s \frac{1 - \rho(A_\Delta(x_s) \mid s)}{\rho(A_\Delta(x_s) \mid s)} \sum_{a \in A_\Delta(x_s)} \rho(a \mid s) \cdot B \; + \; \sum_{s \in S} \pi_s \sum_{a \in \overline{A_\Delta(\mu_s)}} \rho(a \mid s) \cdot B \\
&= B \sum_{s \in S} \pi_s \frac{\rho(\overline{A_\Delta(x_s)} \mid s)}{\rho(A_\Delta(x_s) \mid s)} \rho(A_\Delta(x_s) \mid s) \; + \; B \sum_{s \in S} \pi_s \rho(\overline{A_\Delta(x_s)} \mid s) \\
&= 2B \sum_{s \in S} \pi_s \rho(\overline{A_\Delta(x_s)} \mid s) \overset{(5.27)}{\leq} \frac{2B\delta}{\Delta}.
\end{aligned}$$

This proves the lemma. □

We now prove Lemma 5.4.

Proof of Lemma 5.4. Consider the objective $\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_\pi \min_{\rho \in \mathcal{R}_\delta(\pi)} U(\pi, \rho)$. By Lemma 5.5, for any (π, ρ) there exists an agent strategy $\rho' : s \mapsto \Delta(A_\Delta(x_s))$ that only randomizes over Δ-optimal actions such that $\left| U(\pi, \rho') - U(\pi, \rho) \right| \leq \frac{2B\delta}{\Delta}$. Because minimizing over $\Delta(A_\Delta(x_s))$ is equivalent to minimizing over $A_\Delta(x_s)$, which corresponds to deterministic Δ-best-responding strategies, we get:

$$\begin{aligned}
\underline{\mathrm{OBJ}}^{\mathcal{R}}(\delta) = \sup_{\pi} \min_{\rho \in \mathcal{R}_\delta(\pi)} U(\pi, \rho) &\geq \sup_{\pi} \min_{\rho' : s \mapsto \Delta(A_\Delta(x_s))} U(\pi, \rho') - \tfrac{2B\delta}{\Delta} \\
&= \sup_{\pi} \min_{\rho' : s \mapsto A_\Delta(x_s)} U(\pi, \rho') - \tfrac{2B\delta}{\Delta} \\
&= \underline{\mathrm{OBJ}}^{\mathcal{D}}(\Delta) - \tfrac{2B\delta}{\Delta}.
\end{aligned}$$

□

Chapter 6

Multi-Agent Learning in Auctions

joint work with

Xiaotie Deng, Xinyan Hu, Weiqiang Zheng [DHLZ22]

6.1 Introduction

First price auctions are the current trend in online advertising auctions. A major example is Google Ad Exchange's switch from second price auctions to first price auctions in 2019 [PLST20, GWMS22].

Compared to second price auctions, first price auctions are non-truthful: bidders need to reason about other bidders' private values and bidding strategies and choose their own bids accordingly to maximize their utilities. Finding a good bidding strategy used to be a difficult task due to each bidder's lack of information of other bidders. But given the repeated nature of online advertising auctions and with the advance of computing technology, nowadays' bidders are able to learn to bid using automated bidding algorithms. As one bidder adjusts bidding strategies using a learning algorithm, other bidders' utilities are affected and thus they will adjust their strategies as well. Then, a natural question follows: *if all bidders in a repeated first price auction use some learning algorithms to adjust bidding strategies at the same time, will they converge to a Nash equilibrium of the auction?*

A partial answer to this question is given by [HSMS98] who show that, in a repeated first price auction where bidders have fixed values for the item, a Nash equilibrium may or may not be learned by the *Fictitious Play* algorithm, where in each round of auctions every bidder best responds to the empirical distributions of other bidders' bids in history. Fictitious Play, however, is a deterministic algorithm that does not have the *no-regret* property — a desideratum for learning algorithms in adversarial environments. The no-regret property can only be obtained by randomized algorithms [Rou16]. As observed by [NST15] that bidders' behavior on Bing's advertising system is consistent with no-regret learning, it is hence important, from both theoretical and practical points of view, to understand the convergence property of no-regret algorithms in repeated first price auctions. This motivates our work.

Our Contributions. Focusing on repeated first price auctions where bidders have fixed values, we completely characterize the Nash convergence property of a wide class of randomized online learning algorithms called "mean-based algorithms" [BMSW18]. This class contains most of popular no-regret algorithms, including Multiplicative Weights Update (MWU), Follow the Perturbed Leader (FTPL), etc..

We systematically analyze two notions of Nash convergence: (1) *time-average*: the fraction of rounds where bidders play a Nash equilibrium approaches 1 in the limit; (2) *last-iterate*: the mixed strategy profile of bidders approaches a Nash equilibrium in the limit. Specifically, the results depend on the number of bidders with the highest value:

- If the number is at least three, the bidding dynamics of mean-based algorithms almost surely converges to Nash equilibrium, both in time-average and in last-iterate.

- If the number is two, the bidding dynamics almost surely converges to Nash equilibrium in time-average but not necessarily in last-iterate.

- If the number is one, the bidding dynamics may not converge to Nash equilibrium in time-average nor in last-iterate.

For the last case, the non-convergence result is proved for the Follow the Leader algorithm, which is a mean-based algorithm that is not necessarily no-regret. We also show by experiments that no-regret mean-based algorithms such as MWU and ε_t-Greedy may not last-iterate converge to a Nash equilibrium.

Intuitions and Techniques. The intuition behind our convergence results (the first two cases above) relates to the notion of "iterated elimination of dominated strategies" in game theory. Suppose there are three bidders all having a same integer value v for the item and choosing bids from the set $\{0, 1, \ldots, v-1\}$. The unique Nash equilibrium is all bidders bidding $v-1$. The elimination of dominated bids is as follows: firstly, bidding 0 is dominated by bidding 1 for each of the three bidders no matter what other bidders bid, so bidders will learn to bid 1 or higher instead of bidding 0 at the beginning; then, given that no bidders bid 0, bidding 1 is dominated by bidding 2, so all bidders learn to bid at least 2; ...; in this way all bidders learn to bid $v-1$.[1]

The above intuition is only high-level. In particular, since bidders use mean-based algorithms which may pick a dominated bid with a small but positive probability, additional argument is needed to show that bidders will finally converge to $v-1$ with high probability. To do this, we borrow and generalize a technique (which is a combination of time-partitioning and Azuma's inequality) from [FGL$^+$21] who show that bidders in a second price auction with multiple Nash equilibria converge to the truthful equilibrium if they use mean-based algorithms with an initial uniform exploration stage. Their argument relies on the fact that, in a second price auction, all bidders learn the truthful Nash equilibrium with high probability during the uniform exploration stage. In contrast, we allow any mean-based algorithms without an initial uniform exploration stage.

[1] This logic has been implicitly spelled out by [HSMS98]. But their formal argument only works for deterministic algorithms like Fictitious Play.

6.1.1 Discussion

The Fixed Value Assumption. Our work assumes that each bidder has a fixed value for the item sold throughout the repeated auction. Seemingly restrictive, this assumption is in fact quite common in the literature on repeated auctions, in various contexts including value inference [NST15], dynamic pricing [ARS13, DPS15, ILPT17], and bidding equilibrium [HSMS98, IJS14, KN22, BS22]. An exception is the work by [FGL$^+$21] who study repeated first price auctions under the Bayesian assumption that bidders' values are i.i.d. samples from a distribution. However, their result is restricted to a 2-symmetric-bidder setting with the Uniform$[0, 1]$ distribution where the Bayesian Nash equilibrium (BNE) is simply every bidder bidding half of their values. For general asymmetric distributions there is no explicit characterization of the BNE [Leb96, Leb99, MR00] despite the existence of (inefficient) numerical approximations [FG03, EMR09, WSZ20, CP23]. No algorithms are known to be able to compute BNE efficiently for all asymmetric distributions, let alone a simple, generic learning algorithm (see the Related Work section in Chapter 4).

Second, as we will show, even with the seemingly innocuous assumption of fixed values, the learning dynamics of mean-based algorithms already exhibit complicated behaviors: it may converge to different equilibria in different runs or not converge at all. One can envision more unpredictable behaviors when values are not fixed.

Learning in General Games. Our work is related to a fundamental question in the field of *learning in games* [FL98, CBL06, NRTV07]: if players in a repeated game employ online learning algorithms to adjust strategies, will they converge to an equilibrium? And what kinds of equilibrium? Classical results include the convergence of no-regret learning algorithms to a coarse correlated equilibrium and no-internal-regret algorithms to a correlated equilibrium in any game [FV97, HMC00]. But since (coarse) correlated equilibria are weaker than the archetypical solution concept of a Nash equilibrium, a more appealing and challenging question is the convergence to Nash equilibrium. Positive answers to this question are only known for some special cases of algorithms and games: e.g., no-regret algorithms converge to Nash equilibria in zero-sum games, 2×2 games, and routing games [FL98, CBL06, NRTV07]. In contrast, several works give non-convergence examples: e.g., the non-convergence of MWU in a 3×3 game [DFP+10] and Regularized Learning Dynamics in zero-sum games [MPP18]. In this work we study the Nash equilibrium convergence property in first price auctions for a large class of learning algorithms, namely the mean-based algorithms, and provide both positive and negative results.

Last v.s. Average Iterate Convergence. We note that previous results on convergence of learning dynamics to Nash equilibria in games are mostly attained in an average sense, i.e., the empirical distributions of players' actions converge. Our notion of time-average convergence, which requires players play a Nash equilibrium in almost every round, is different from the convergence of empirical distributions; in fact, ours is stronger if the Nash equilibrium is unique. Nevertheless, time-average convergence fails to capture the full picture of the dynamics since players' last-iterate strategy profile may not converge. Existing results about last-iterate convergence show that many learning dynamics actually diverge or cycle even in a simple 3×3 game [DFP+10] or zero-sum games [MPP18], except for a few convergence examples like optimistic gradient descent in two-player zero-sum games or monotone games [DP18, WLZL21, COZ22]. Our results and techniques, regarding the convergence of any mean-based algorithm in first price auctions, shed light on further study of last-iterate convergence in more general settings.

6.1.2 Additional Related Works

Online Learning in Auctions. A large fraction of existing works on online learning in repeated auctions are from the *seller*'s perspective, i.e., studying how a seller can maximize revenue by adaptively changing the rules of the auction (e.g., reservation price) over time (e.g., [BH05, ARS13, MM14, CBGM15, BMSW18, HLW18, ACK+19, KN19, DLL+20, GJM21]). We focus on the *bidders'* learning problem. Some previous works from bidders' perspective are about "learning to bid", studying on how to design no-regret algorithms for a bidder to bid in various formats of repeated auctions, including first price auctions [BGM+19, BFG23, HWZ25], second price auctions [IJS14, WPR16], and more general auctions [FPS18, KSLK20]. Those works consider a single bidder, instead of the interaction among multiple bidders learning to bid at the same time. We instead study the consequence of such interaction, showing that the learning dynamics of multiple bidders may or may not converge to the Nash equilibrium of the auction.

Multi-Agent Learning in First Price Auctions In addition to the aforementioned works by [FGL+21] and [HSMS98], other works on multi-agent learning in first price auctions include, e.g., several empirical works by [BFH+21, GWMS22, BS22], who observe convergence results for some learning algorithms experimentally, and a theoretical work by [KN22]. [KN22] prove that in repeated first price auctions with two mean-based learning bidders, *if* the dynamics converge to some limit, then this limit must be a CCE in which the bidder with the higher value submits bids that are close to the lower value. However, they do not provide the condition under which the dynamics converge. We prove that the dynamics converge if the two bidders have the same value and in fact converge to the stronger notion of a Nash equilibrium. This result complements [KN22] and supports the aforementioned empirical findings.

Learning to Iteratively Eliminate Dominated Strategies. To our knowledge, we are the first to prove that mean-based learning algorithms are able to iteratively eliminate dominated strategies in repeated games. Although this result seems intuitive, the proof is involved due to the randomness of mean-based algorithms. As mentioned in the Introduction, we generalize a "time-partitioning" technique in [FGL+21] to overcome this difficulty and give a formal proof for this result. Some recent works on multi-agent learning in other games [WXY22, FYC22] also observed but did not formally prove this result. Moreover, building on our work, [BDO24] prove that mean-based algorithms can iteratively eliminate dominated strategies in more general games.

Interestingly, though, [WXY22] note that mean-based algorithms needs an *exponential* time to iteratively eliminate all dominated strategies in some special games, while [WKBJ23] develop a polynomial-time algorithm that is no mean-based. We do not know the exact convergence rate of mean-based algorithms in first price auction games.

6.2 Model and Preliminaries

Repeated First Price Auctions. We consider repeated first-price sealed-bid auctions where a single seller sells a good to a set of $N \geq 2$ players (bidders) $\mathcal{N} = \{1, 2, ..., N\}$ for infinitely many rounds. Each player $i \in \mathcal{N}$ has a fixed private value v^i for the good throughout. See Section 5.6 for a discussion on this assumption. We assume that v^i is a positive integer in some range $\{1, \dots, V\}$ where V is an upper bound on v^i. Assume $V \geq 3$. No player knows other players' values. Without loss of generality, assume $v^1 \geq v^2 \geq \cdots \geq v^N$.

At each round $t \geq 1$ of the repeated auctions, each bidder i submits a bid $b_t^i \in \{0, 1, \dots, V\}$ to compete for the good. A discrete set of bids captures the reality that the minimum unit of money is a cent. The bidder with the highest bid wins the good. If there are more than one highest bidders, the good is allocated to one of them uniformly at random. The bidder who wins the good pays her bid b_t^i, obtaining utility $v^i - b_t^i$; other bidders obtain utility 0. Let $u^i(b_t^i, \boldsymbol{b}_t^{-i})$ denote bidder i's (expected) utility when i bids b_t^i while other bidders bid $\boldsymbol{b}_t^{-i} = (b_t^1, \dots, b_t^{i-1}, b_t^{i+1}, \dots, b_t^N)$, i.e., $u^i(b_t^i, \boldsymbol{b}_t^{-i}) = (v^i - b_t^i)\mathbb{I}[b_t^i = \max_{j \in \mathcal{N}} b_t^j] \frac{1}{|\arg\max_{j \in \mathcal{N}} b_t^j|}$.

We assume that bidders never bid above or equal to their values since that brings them negative or zero utility, which is clearly dominated by bidding 0. We denote the set of possible bids of each bidder i by $\mathcal{B}^i = \{0, 1, \ldots, v^i - 1\}$.

Multi-Agent Online Learning. We assume that each bidder $i \in \mathcal{N}$ chooses her bids at every round using an online learning algorithm. Specifically, we regard the set of possible bids $\mathcal{B}^i$ as a set of actions (or arms). At each round t, the algorithm picks (possibly in a random way) an action $b_t^i \in \mathcal{B}^i$ to play, and then receives some feedback. The feedback may include the rewards (i.e., utilities) of all possible actions in $\mathcal{B}^i$ (in the *experts* setting) or only the reward of the chosen action b_t^i (in the *multi-arm bandit* setting). With feedback, the algorithm updates its choice of actions in future rounds. We do not assume a specific feedback model in this work. Our analysis will apply to all online learning algorithms that satisfy the following property, called "mean-based" [BMSW18, FGL+21], which roughly says that the algorithm picks actions with low average historical rewards with low probabilities.

Definition 6.1 (mean-based algorithm)**.** *Let $\alpha_t^i(b)$ be the average reward of action b in the first t rounds: $\alpha_t^i(b) = \frac{1}{t}\sum_{s=1}^{t} u^i(b, \boldsymbol{b}_s^{-i})$. An algorithm is γ_t-*mean-based *if, for any $b \in \mathcal{B}^i$, whenever there exists $b' \in \mathcal{B}^i$ such that $\alpha_{t-1}^i(b') - \alpha_{t-1}^i(b) > V\gamma_t$, the probability that the algorithm picks b at round t is at most γ_t. An algorithm is* mean-based *if it is γ_t-mean-based for some decreasing sequence $(\gamma_t)_{t=1}^{\infty}$ such that $\gamma_t \to 0$ as $t \to \infty$.*

In this work, we assume that the online learning algorithm can run for infinitely many rounds. This captures the scenario where bidders do not know how long they will be in the auction and hence use learning algorithms that work for an arbitrarily long time. Infinite-round mean-based algorithms can be obtained by modifying classical finite-round mean-based algorithms (e.g., MWU) with constant learning rates to have decreasing learning rates, as shown below:

Example 6.1. *Let* $(\varepsilon_t)_{t=1}^{\infty}$ *be a decreasing sequence approaching* 0. *The following algorithms are mean-based:*

- Follow the Leader *(also called* Greedy*): at each round* $t \geq 1$*, each player* $i \in \mathcal{N}$ *chooses an action* $b \in \arg\max_{b \in \mathcal{B}^i}\{\alpha_{t-1}^i(b)\}$ *(with a tie-breaking rule specified by the algorithm).*
- ε_t-Greedy*: at each round* $t \geq 1$*, each player* $i \in \mathcal{N}$ *with probability* $1-\varepsilon_t$ *chooses* $b \in \arg\max_{b \in \mathcal{B}^i}\{\alpha_{t-1}^i(b)\}$*, with probability* ε_t *chooses an action in* $\mathcal{B}^i$ *uniformly at random.*
- Multiplicative Weights Update *(MWU, also called* Hedge*): at each round* $t \geq 1$*, each player* $i \in \mathcal{N}$ *chooses each action* $b \in \mathcal{B}^i$ *with probability* $\frac{w_{t-1}(b)}{\sum_{b' \in \mathcal{B}^i} w_{t-1}(b')}$*, where* $w_t(b) = \exp(\varepsilon_t \sum_{s=1}^t u^i(b, \boldsymbol{b}_s^{-i}))$*. This MWU algorithm is different from another MWU algorithm where the weight is defined by* $w_t(b) = w_{t-1}(b) \cdot \exp(\varepsilon_t u^i(b, \boldsymbol{b}_t^{-i})) = \exp(\sum_{s=1}^t \varepsilon_s u^i(b, \boldsymbol{b}_s^{-i}))$*. The latter algorithm is not mean-based because the rewards* $u^i(b, \boldsymbol{b}_s^{-i})$ *in earlier rounds matter more than rewards in later rounds given decreasing* ε_s*. The algorithm we define here treat rewards at different rounds equally and is hence mean-based.*

Clearly, Follow the Leader is $(\gamma_t = 0)$-mean-based and ε_t-Greedy is ε_t-mean-based. One can see [BMSW18] for why MWU is mean-based. Additionally, MWU is no-regret when the sequence $(\varepsilon_t)_{t=1}^{\infty}$ is set to $\varepsilon_t = O(1/\sqrt{t})$ (see, e.g., Theorem 2.3 in [CBL06]).

Equilibria in First Price Auctions. Before presenting our main results, we characterize the set of all Nash equilibria in the single-round first price auction where bidders have fixed values $v^1 \geq v^2 \geq \cdots \geq v^N$. We focus on pure-strategy Nash equilibria. Recall that $u^i(b^i, \boldsymbol{b}^{-i})$ denotes the utility of bidder i when she bids b^i while others bid $\boldsymbol{b}^{-i} = (b^1, \ldots, b^{i-1}, b^{i+1}, \ldots, b^N)$. A bidding profile $\boldsymbol{b} = (b^1, \ldots, b^N) = (b^i, \boldsymbol{b}^{-i})$ is called a

Nash equilibrium if $u^i(\boldsymbol{b}) \geq u^i(b', \boldsymbol{b}^{-i})$ for any $b' \in \mathcal{B}^i$ and any $i \in \mathcal{N}$. Let M^i be the set of bidders who have the same value as bidder i, $M^i = \{j \in \mathcal{N} : v^j = v^i\}$. M^1 is the set of bidders with the highest value.

Proposition 6.1. *The set of (pure-strategy) Nash equilibria in the first price auction with fixed values $v^1 \geq v^2 \geq \cdots \geq v^N$ are bidding profiles $\boldsymbol{b} = (b^1, \ldots, b^N)$ that satisfy the following:*

- *The case of $|M^1| \geq 3$: $b^i = v^1 - 1$ for $i \in M^1$ and $b^j \leq v^1 - 2$ for $j \notin M^1$.*
- *The case of $|M^1| = 2$:*
  - *If $N = 2$ or $v^1 = v^2 > v^3 + 1$: there are two types of Nash equilibria: (1) $b^1 = b^2 = v^1 - 1$, with $b^j \leq v^1 - 3$ for $j \notin M^1$; (2) $b^1 = b^2 = v^1 - 2$, with $b^j \leq v^1 - 3$ for $j \notin M^1$.*
  - *If $N > 2$ and $v^1 = v^2 = v^3 + 1$: $b^1 = b^2 = v^1 - 1$ and $b^j \leq v^1 - 2$ for $j \notin M^1$.*
- *The case of $|M^1| = 1$:*
  - *Bidding profiles that satisfy the following are Nash equilibria: $b^1 = v^2$, at least one bidder in M^2 bids $v^2 - 1$, all other bidders bid $b^j \leq v^2 - 1$.*
  - *If $v^1 = v^2 + 1$ and $|M^2| = 1$, then there is another type of Nash equilibria: $b^1 = b^2 = v^2 - 1$, $b^j \leq v^2 - 2$ for $j \notin \{1, 2\}$.*

There are no other (pure-strategy) Nash equilibria.

The proof of this proposition is straightforward and omitted. Intuitively, whenever more than one bidder has the highest value ($|M^1| \geq 2$), they should compete with each other by bidding $v^1 - 1$ (or $v^1 - 2$ if $|M^1| = 2$ and no other bidders are able to compete with them). When $|M^1| = 1$, the unique highest-value bidder (bidder 1) competes with the second-highest bidders (M^2).

6.3 Convergence of Mean-Based Algorithms in First Price Auctions

We introduce some additional notations. Let $\boldsymbol{x}_t^i \in \Delta(\mathcal{B}^i)$ be the mixed strategy of player i in round t, where the b-th component of $\boldsymbol{x}_t^i$ is the probability that player i bids $b \in \mathcal{B}^i$ in round t. The sequence $(\boldsymbol{x}_t^i)_{t=1}^{\infty}$ is a stochastic process, where the randomness comes from the bidder's learning algorithms. Let $\mathbf{1}_b$ be the vector $(0, ..., 0, 1, 0, ..., 0)$ where 1 is in the b-th position.

Our main results about the convergence of mean-based algorithms in repeated first price auctions depend on the number of bidders with the highest value, $|M^1|$.

6.3.1 The case of $|M^1| \geq 3$

Theorem 6.1. *If $|M^1| \geq 3$ and every bidder follows a mean-based algorithm, then, with probability* 1*, both of the following events happen:*

- *Time-average convergence of bid sequence:*

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}\big[\forall i \in M^1, b_s^i = v^1 - 1\big] = 1.$$

- *Last-iterate convergence of mixed strategy profile:* $\forall i \in M^1,\ \lim_{t\to\infty} \boldsymbol{x}_t^i = \mathbf{1}_{v^1-1}$.

Theorem 6.1 can be interpreted as follows. According to Proposition 6.1, when $|M^1| \geq 3$, the bidding profile $\boldsymbol{b}_s$ at round s is a Nash equilibrium if and only if $\forall i \in M^1, b_s^i = v^1 - 1$, with bidders not in M^1 bidding $\leq v^1 - 2$ by assumption (note that the bidders not in M^1 can follow a mixed strategy and need not converge to a deterministic bid). Hence, the first result of Theorem 6.1 implies that the fraction of rounds where bidders play a Nash equilibrium approaches 1 in the limit. The second result shows that all bidders in M^1 bid $v^1 - 1$ with certainty eventually, achieving a Nash equilibrium. We will prove Theorem 6.1 in Section 6.4.

6.3.2 The case of $|M^1| = 2$

Theorem 6.2. *If $|M^1| = 2$ and every bidder follows a mean-based algorithm, then, with probability 1, one of the following two events happens:*

- $\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1,\ b_s^i = v^1 - 2] = 1$;
- $\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1,\ b_s^i = v^1 - 1] = 1$ *and* $\forall i \in M^1$, $\lim_{t\to\infty} \boldsymbol{x}_t^i = \mathbf{1}_{v^1-1}$.

Moreover, if $N > 2$ and $v^3 = v^1 - 1$ then only the second event happens.

For the case of $N = 2$ or $v^3 < v^1 - 1$, according to Proposition 6.1, $\boldsymbol{b}_s$ is a Nash equilibrium if and only if both bidders in M^1 play $v^1 - 1$ or $v^1 - 2$ at the same time, with other bidders bidding $\leq v^1 - 3$. Hence, Theorem 6.2 shows that the bidders eventually converge to one of the two possible types of equilibria. Interestingly, experiments show that some mean-based algorithms converge to the equilibrium of $v^1 - 1$ while some converge to $v^1 - 2$. Also, one algorithm may converge to different equilibria in different runs. See Section 6.5 for details.

In the case of time-average convergence to the equilibrium of $v^1 - 2$ (the first case of Theorem 6.2), the last-iterate convergence result ($\forall i \in M^1$, $\lim_{t\to\infty} \boldsymbol{x}_t^i = \mathbf{1}_{v^1-2}$) does not always hold. Consider an example with 2 bidders, with $v^1 = v^2 = 3$. We can construct a γ_t-mean-based algorithm with $\gamma_t = O(\frac{1}{t^{1/4}})$ such that, with constant probability, $\lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1,\ b_s^i = v^1 - 2] = 1$ holds but in infinitely many rounds we have $\boldsymbol{x}_t^i = \mathbf{1}_2 = \mathbf{1}_{v^1-1}$, so the algorithm does not converge in last iterate. The key idea is that, when $\alpha_t^i(1) - \alpha_t^i(2)$ is positive but lower than $V\gamma_t$ in some round t (which happens infinitely often), we can let the algorithm bid 2 with certainty in round $t+1$. This does not violate the γ_t-mean-based property. The algorithm is presented in Algorithm 6.1, and the full analysis is in Section 6.7.2.

Proposition 6.2. *There exist an example with $|M^1| = 2$ and a mean-based algorithm (Algorithm 6.1) such that, with constant probability, the bidders time-average converge to the Nash equilibrium of $v^1 - 2$ but do not last-iterate converge to a Nash equilibrium.*

Algorithm 6.1: A mean-based bidding algorithm

```
Input: Value $v = 3$
Let $T_0 = 10^{12}$. $T_k = 32^k T_0$ for $k \geq 0$. $\gamma_t = T_k^{-1/4}$ for $t \in [T_k + 1, T_{k+1}]$.
In the first $T_0$ rounds: Bid $b_t = 1$ for $t \leq T_0 - T_0^{2/3}$ and bid $b_t = 0$ for
  $T_0 - T_0^{2/3} + 1 \leq t \leq T_0$.
for $t \geq T_0 + 1$ do
    Find $k$ such that $32^k T_0 + 1 \leq t \leq 32^{k+1} T_0$.
    if $t = T_k + 1$, $\arg\max_b \alpha_{t-1}(b) = 1$, and $\alpha_{t-1}^i(1) - \alpha_{t-1}^i(2) < V\gamma_t$ then
        Bid $b_t = 2$.
    else
        Bid $b_t \in \arg\max_{b \in \{0,1,2\}} \alpha_{t-1}(b)$ with probability $1 - T_{k+1}^{-1/3}$ and 0 with
          probability $T_{k+1}^{-1/3}$.
```

6.3.3 The case of $|M^1| = 1$

In the case of $|M^1| = 1$, mean-based learning dynamics may not converge to a Nash equilibrium of the auction in time-average or last-iterate, as shown in the following example.

Example 6.2. *Let $v^1 = 10$, $v^2 = v^3 = 7$. Assume that players use the* Follow the Leader *algorithm with a specific tie-breaking rule. They may generate the following bid sequence $(b_t^1, b_t^2, b_t^3)_{t \geq 1}$: $(7,6,1), (7,1,6), (7,1,1), (7,6,1), (7,1,6), (7,1,1), \ldots,$ while satisfying 0-mean-based. Note that $(7,1,1)$ is not a Nash equilibrium according to Proposition 6.1 but it appears in $\frac{1}{3}$ fraction of rounds, which means that the dy-*

namics do not converge in the time-average sense or the last-iterate sense to a Nash equilibrium.

Proof. To prove this example, we need to verify that the players' algorithms are indeed 0-mean-based under the above bid sequence. Because players 2 and 3 always get zero utility no matter what they bid, we only need to verify the 0-mean-based property for player 1. Let q_t be the fraction of rounds in the first t rounds where one of players 2 and 3 bids 6 (in the other $1 - q_t$ fraction of rounds both players 2 and 3 bid 1); clearly, $q_t \geq \frac{2}{3}$ for any $t \geq 1$. For player 1, at each round t her average utility by bidding 7 is $\alpha^1_{t-1}(7) = 10 - 7 = 3$; by bidding 6, $\alpha^1_{t-1}(6) = (10-6)(\frac{1}{2}q_{t-1} + (1 - q_{t-1})) = 4(1 - \frac{q_{t-1}}{2}) \leq \frac{8}{3} < 3$; by bidding 2, $\alpha^1_{t-1}(2) = (10-2)(1-q_{t-1}) \leq \frac{8}{3} < 3$; and clearly $\alpha^1_{t-1}(b) < 3$ for any other bid. Hence, $7 = \arg\max_{b \in \mathcal{B}^1}\{\alpha^1_{t-1}(b)\}$, player 1 satisfies 0-mean-based. □

Example 6.2 also shows that, in the case of $|M^1| = 1$, the bidding dynamics generated by a mean-based algorithm may not converge to Nash equilibrium in the classical sense of "convergence of empirical distribution". Specifically, let $p^i_t = \frac{1}{t}\sum_{s=1}^{t} \mathbf{1}_{b^i_s} \in \Delta(\mathcal{B}^i)$ be the empirical distribution of player i's bids up to round t. "Convergence of empirical distribution" means that the players' empirical distributions $(p^1_t, p^2_t, p^3_t)_{t \geq 1}$ converge to a mixed-strategy Nash equilibrium in the limit. In Example 6.2, the players' empirical distributions converge to (p^1, p^2, p^3) where $p^1(7) = 1$ and for $i = 2, 3$, $p^i(6) = \frac{1}{3}$ and $p^i(1) = \frac{2}{3}$. Given bidders 2 and 3's strategies (p^2, p^3), bidder 1 can obtain utility $(10-2)(\frac{2}{3})^2 = \frac{32}{9}$ by bidding 2, which is larger than the utility of bidding 7, which is $10 - 7 = 3$. Thus, bidder 1's strategy p^1 is not a best response to (p^2, p^3), hence (p^1, p^2, p^3) is not a Nash equilibrium.

The mean-based algorithm in Example 6.2 is not no-regret. In Section 6.5 we show by experiments that such non-convergence results also hold for no-regret mean-based algorithms, e.g., MWU.

6.4 Proof of Theorem 6.1

The proof of Theorem 6.1 covers the main ideas and proof techniques of our convergence results, so we present it here. We first provide a proof sketch. Then in Section 6.4.1 we provide some properties of mean-based algorithms that will be used in the formal proof. Section 6.4.2 and Section 6.4.3 prove Theorem 6.1.

Proof sketch At a high level, the proof uses the idea of iterative elimination of dominated strategies in game theory. We first use an induction argument to show that bidders with the highest value (i.e., those in M^1) will gradually learn to eliminate bids $0, 1, \ldots, v^1 - 3$. Then we further prove that: if $|M^1| = 3$, they will eliminate $v^1 - 2$ and hence converge to $v^1 - 1$; if $|M^1| = 2$, the two bidders may end up playing $v^1 - 1$ or $v^1 - 2$.

To see why bidders in M^1 will learn to eliminate 0, suppose that there are two bidders in total and one of them (say, bidder i) bids b with probability $P(b)$ in history. For the other bidder (say, bidder j), if bidder j bids 0, she obtains utility $\alpha(0) = (v^1 - 0)\frac{P(0)}{2}$; if she bids 1, she obtains utility $\alpha(1) = (v^1 - 1)(P(0) + \frac{P(1)}{2})$. Since $\alpha(1) - \alpha(0) = \frac{v^1-2}{2}P(0) + (v^1 - 1)\frac{P(1)}{2} > 0$ (assuming $v^1 \geq 3$), bidding 1 is better than bidding 0 for bidder j. Given that bidder j is using a mean-based algorithm, she will play 0 with small probability (say, zero probability). The same argument applies to bidder i. Hence, both bidders will learn to not play 0. Then we take an inductive step: assuming that no bidders play $0, \ldots, k-1$, we have $\alpha(k+1) - \alpha(k) = \frac{v^1-k-2}{2}P(k) + \frac{v^1-k-1}{2}P(k+1) > 0$ for $k \leq v^1 - 3$, therefore $k+1$ is a better response than k and both players will avoid bidding k. An induction shows that they will finally learn to avoid $0, 1, \ldots, v^1 - 3$. Then, for the case of $|M^1| \geq 3$, we will use an additional lemma (Lemma 6.2) to show that, if bidders bid $0, 1, \ldots, v^1 - 3$ rarely in history, they will also avoid bidding $v^1 - 2$ in the future.

However, mean-based learning bidders may choose dominated bids with a small but positive probability. To prove a high-probability convergence result, we use and generalize a time-partitioning technique from [FGL+21]. We partition the time horizon into periods $1 < T_0 < T_1 < T_2 < \cdots$. If bidders bid $0, 1, \ldots, k-1$ with low frequency from round 1 to T_{k-1}, then using the mean-based properties in Lemma 6.1 and Lemma 6.2, we show that they will bid k with probability at most γ_t in each round from $T_{k-1}+1, T_{k-1}+2, \ldots,$ to T_k. A use of Azuma's inequality shows that the frequency of bid k in period $(T_{k-1}, T_k]$ is also low with high probability, which concludes the induction. Constructing an appropriate partition allows us to argue that the frequency of bids less than $v^1 - 1$ converges to 0 *with high probability*.

6.4.1 Iterative Elimination Properties of Mean-Based Algorithms

We define some notations for the proofs. Let $P_t^i(k)$ be the frequency of the highest bid submitted by bidders other than i being k during the first t rounds:

$$P_t^i(k) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\max_{j\neq i} b_s^j = k].$$

Let $P_t^i(0:k)$ be $\sum_{\ell=0}^{k} P_t^i(\ell)$. Let $P_t^i(0:-1)$ be 0. Let $Q_t^i(k)$ be the probability of bidder i winning the item with ties if she bids k in history:

$$Q_t^i(k) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\max_{j\neq i} b_s^j = k]\frac{1}{|\arg\max_{j\neq i} b_s^j| + 1}.$$

Clearly,

$$0 \;\leq\; \frac{1}{N}P_t^i(k) \;\leq\; Q_t^i(k) \;\leq\; \frac{1}{2}P_t^t(k) \;\leq\; \frac{1}{2}. \tag{6.1}$$

Recall that $\alpha_t^i(k)$ is bidder i's average utility by bidding k in the first t rounds. We can write $\alpha_t^i(k)$ using $P_t^i(0:k-1)$ and $Q_i^t(k)$:

$$\alpha_t^i(k) = (v^i - k)\big(P_t^i(0:k-1) + Q_t^i(k)\big). \tag{6.2}$$

We use H_t to denote the history of the first t rounds, which includes the realization of all randomness in the first t rounds. Bidders themselves do not necessarily observe the full history H_t. Given H_{t-1}, each bidder's mixed strategy $\boldsymbol{x}_t^i$ at round t is fully determined, and the k-th component of $\boldsymbol{x}_t^i$ is $\Pr[b_t^i = k \mid H_{t-1}]$. The following lemma shows that, if other bidders rarely bid 0 to $k-1$ in history, then bidder $i \in M^1$ will not bid k with large probability in round t, for $k \le v^1 - 3$.

Lemma 6.1. *Assume $v^1 \geq 3$. For any $i \in M^1$, any $k \in \{0, 1, \ldots, v^1 - 3\}$, any t such that $\gamma_t < \frac{1}{12NV}$, if the history H_{t-1} of the first $t-1$ rounds satisfies $P_{t-1}^i(0 : k-1) \leq \frac{1}{3NV}$, then $\Pr[b_t^i = k \mid H_{t-1}] \leq \gamma_t$.*

The intuition behind Lemma 6.1 is that, if other bidders never bid 0 to $k-1$, then bidder i's bid k will be dominated by a mixed strategy between bidding $k+1$ and $v^1 - 1$. However, Lemma 6.1 is stronger than the classical notion of iterative dominance because it requires bidder i's bid k to be dominated even when other bidders bid 0 to $k-1$ with a small constant probability $\frac{1}{3NV}$. This stronger property is crucial to our proof of Theorem 6.1.

Proof of Lemma 6.1. If $\alpha_{t-1}^i(k+1) - \alpha_{t-1}^i(k) > V\gamma_t$, then by the mean-based property, the conditional probability $\Pr[b_t^i = k \mid \alpha_{t-1}^i(k+1) - \alpha_{t-1}^i(k) > V\gamma_t, H_{t-1}]$ is at most γ_t, so the lemma holds. Then, we consider the case where $\alpha_{t-1}^i(k+1) - \alpha_{t-1}^i(k) \leq V\gamma_t$. Using (6.2) and (6.1),

$$\begin{aligned} V\gamma_t &\geq \alpha_{t-1}^i(k+1) - \alpha_{t-1}^i(k) \\ &\geq (v^1 - k - 1)P_{t-1}^i(k) - P_{t-1}^i(0 : k-1) \\ &\qquad - (v^1 - k)\tfrac{P_{t-1}^i(k)}{2}, \end{aligned}$$

which implies

$$P^i_{t-1}(k) \le \tfrac{2}{v^1-k-2}\big(V\gamma_t + P^i_{t-1}(0:k-1)\big). \tag{6.3}$$

We then upper bound $\alpha^i_{t-1}(k)$ as follows:

$$\begin{aligned}
\alpha^i_{t-1}(k) \ &\le\ (v^1-k)\big(P^i_{t-1}(0:k-1) + \tfrac{1}{2}P^i_{t-1}(k)\big) \\
&\le (v^1-k)P^i_{t-1}(0:k-1) \qquad\qquad \text{by (6.3)} \\
&\qquad + \tfrac{v^1-k}{v^1-k-2}\big(V\gamma_t + P^i_{t-1}(0:k-1)\big) \\
&= \tfrac{v^1-k}{v^1-k-2}V\gamma_t + \big(v^1-k+\tfrac{v^1-k}{v^1-k-2}\big)P^i_{t-1}(0:k-1) \\
&\le 3V\gamma_t + \big(v^1-k+3\big)P^i_{t-1}(0:k-1) \\
&\le 3V\gamma_t + 2VP^i_{t-1}(0:k-1).
\end{aligned}$$

By the condition of the lemma, we have $P^i_{t-1}(0:k-1) \le \frac{1}{3NV} < \frac{1}{2NV} - 2\gamma_t$. So

$$\alpha^i_{t-1}(k) < 3V\gamma_t + 2V\left(\tfrac{1}{2NV} - 2\gamma_t\right) = \tfrac{1}{N} - V\gamma_t.$$

Then, we note that $\alpha^i_{t-1}(v^1-1) = P^i_{t-1}(0:v^1-2) + Q^i_{t-1}(v^1-1) \ge \frac{1}{N}P^i_{t-1}(0:v^1-1) = \frac{1}{N}\cdot 1$ where the last equality holds because no bidder bids above v^1-1 by assumption. Therefore,

$$\alpha^i_{t-1}(v^1-1) - \alpha^i_{t-1}(k) > \tfrac{1}{N} - \big(\tfrac{1}{N} - V\gamma_t\big) = V\gamma_t.$$

From the mean-based property, we obtain $\Pr[b^i_t = k \mid \alpha^i_{t-1}(k{+}1) - \alpha^i_{t-1}(k) \le V\gamma_t, H_{t-1}] \le \gamma_t$. □

The following lemma is for $k = v^1-2$: if bidders rarely bid 0 to v^1-3 in history and $|M^1| \ge 3$, then bidder $i \in M^1$ will not bid v^1-2 with large probability in round t. Intuitively, this is because v^1-2 is dominated by v^1-1. See Appendix for details.

Lemma 6.2. *Suppose $|M^1| \geq 3$ and $v^1 \geq 2$. For any t such that $\gamma_t < \frac{1}{12NV}$, if the history H_{t-1} of the first $t-1$ rounds satisfies $\frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \leq \frac{1}{3NV}$, then, $\forall\ i \in M^1$,* $\Pr[b_t^i = v^1 - 2 \mid H_{t-1}] \leq \gamma_t$.

6.4.2 Iteratively Eliminating Bids $0, 1, \ldots, v^1 - 3$

In this subsection we will use an induction argument to prove that, after a sufficiently long time, bidders in M^1 will rarely bid $0, 1, \ldots, v^1 - 3$ (Corollary 6.1). We partition the time horizon into $v^1 - 3$ periods. Let constants $c = 1 + \frac{1}{12NV}$ and $d = \lceil \log_c(8NV) \rceil$. Let T_b be any (constant) integer such that $\gamma_{T_b} < \frac{1}{12N^2V^2}$ and $\exp\left(-\frac{(c-1)T_b}{1152N^2V^2}\right) \leq \frac{1}{2}$. Let $T_0 = 12NVT_b$ and $T_k = c^d T_{k-1} = c^{dk} T_0 \geq (8NV)^k T_0$ for $k \in \{1, 2, \ldots, v^1 - 3\}$. Let A_k be event

$$A_k = \left[\frac{1}{T_k}\sum_{t=1}^{T_k} \mathbb{I}[\exists i \in M^1, b_t^i \leq k] \leq \frac{1}{4NV}\right],$$

which says that bidders in M^1 bid $0, 1, \ldots, k$ not too often in the first T_k rounds. Our goal is to show that $\Pr[A_{v^1-3}]$ is high.

The **base case** is to show that A_0 happens with high probability.

Lemma 6.3. $\Pr[A_0] \geq 1 - \exp\left(-\frac{T_b}{24NV}\right)$.

Proof. Consider any round $t \geq T_b$. For any $i \in M^1$, given any history H_{t-1} of the first $t-1$ rounds, it holds that $P_{t-1}^i(0 : -1) = 0 \leq \frac{1}{3NV}$. Hence, by Lemma 6.1, we have $\Pr[b_t^i = 0 \mid H_{t-1}] \leq \gamma_t$. Using a union bound over $i \in M^1$,

$$\Pr[\exists i \in M^1, b_t^i = 0 \mid H_{t-1}] \leq |M^1|\gamma_t.$$

Let $Z_t = \mathbb{I}[\exists i \in M^1, b_t^i = 0] - |M^1|\gamma_t$ and let $X_t = \sum_{s=T_b+1}^{t} Z_s$. We have $\mathbb{E}[Z_t \mid H_{t-1}] \leq 0$. Therefore, the sequence $X_{T_b+1}, X_{T_b+2}, \ldots, X_{T_0}$ is a supermartingale (with respect to the sequence of history $H_{T_b}, H_{T_b+1}, \dots, H_{T_0-1}$). By Azuma's inequality, for any $\Delta > 0$, we have

$$\Pr\left[\sum_{t=T_b+1}^{T_0} Z_t \geq \Delta\right] \leq \exp\left(-\frac{\Delta^2}{2(T_0 - T_b)}\right).$$

Let $\Delta = T_b$. We have with probability at least $1 - \exp\left(-\frac{\Delta^2}{2(T_0-T_b)}\right) \geq 1 - \exp\left(-\frac{T_b}{24NV}\right)$, $\sum_{t=T_b+1}^{T_0} Z_t < T_b$ holds, namely, $\sum_{t=T_b+1}^{T_0} \mathbb{I}[\exists i \in M^1, b_t^i = 0] < T_b + \sum_{t=T_b+1}^{T_0} |M^1|\gamma_t$, which implies

$$\begin{aligned}
&\frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{I}[\exists i \in M^1, b_t^i = 0] \\
&\leq \frac{1}{T_0}\left(T_b + \sum_{t=T_b+1}^{T_0} \mathbb{I}[\exists i \in M^1, b_t^i = 0]\right) \\
&< \frac{1}{T_0}\left(2T_b + \sum_{t=T_b+1}^{T_0} |M^1|\gamma_t\right) \ \leq \ \frac{1}{4NV},
\end{aligned}$$

where the last inequality is because $\frac{T_b}{T_0} = \frac{1}{12NV}$ and $\frac{1}{T_0}\sum_{t=T_b+1}^{T_0} |M^1|\gamma_t \leq |M^1|\gamma_{T_b} \leq \frac{1}{12NV}$. □

Then, we use **induction** to show that, if bidders in M^1 seldom bid $0, 1, \dots, k$ in the first T_k rounds, then they will also seldom bid $0, 1, \dots, k, k+1$ in the first T_{k+1} rounds, with high probability.

Lemma 6.4. *Suppose* $|M^1| \geq 2$. *For every* $k \in [0, v^1 - 4]$, $\Pr[A_{k+1} \mid A_k] \geq 1 - \sum_{j=1}^{d} \exp\left(-\frac{|\Gamma_\ell^j|}{1152N^2V^2}\right)$.

To prove Lemma 6.4, we divide the rounds in $[T_k, T_{k+1}]$ to $d = \lceil \log_c(8NV) \rceil$ episodes such that $T_k = T_k^0 < T_k^1 < \cdots < T_k^d = T_{k+1}$ where $T_k^j = cT_k^{j-1}$. Let $\Gamma_k^j = [T_k^{j-1}+1, T_k^j]$, with $|\Gamma_k^j| = T_k^j - T_k^{j-1}$. We define a series of events B_k^j for $j \in [0, d]$. B_k^0 is the same as A_k. For $j \in [1, d]$,

$$B_k^j = \left[\sum_{t \in \Gamma_k^j} \mathbb{I}[\exists i \in M^1, b_t^i \leq k+1] \leq \frac{|\Gamma_k^j|}{8NV}\right].$$

Claim 6.1. *For every $j \in [0, d-1]$,* $\Pr\left[B_k^{j+1} \mid A_k, B_k^1, \ldots, B_k^j\right] \geq 1 - \exp\left(-\frac{|\Gamma_k^{j+1}|}{1152N^2V^2}\right)$.

Proof. Suppose $A_k, B_k^1, \ldots, B_k^j$ happen. We denote $A_k^j = [A_k, B_k^1, \ldots, B_k^j]$. We argue that event A_k^j implies $P_{t-1}^i(0:k) \leq \frac{1}{3NV}$ for every bidder $i \in M^1$ and every round $t \in \Gamma_k^j = [T_k^j + 1, T_k^{j+1}]$. Recall that $P_{t-1}^i(0:k) = \frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[\max_{i' \neq i} b_s^{i'} \leq k]$. Because $|M^1| \geq 2$, the event $\mathbb{I}[\max_{i' \neq i} b_s^{i'} \leq k]$ implies that there exists $i^* \in M^1$, $i^* \neq i$, such that $b_s^{i^*} \leq k$. Thus,

$$\begin{aligned}
P_{t-1}^i(0:k) &\leq \frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq k] \\
&= \frac{1}{t-1}\Bigg(\sum_{s=1}^{T_k} \mathbb{I}[\exists i \in M^1, b_s^i \leq k] \\
&\quad + \sum_{s \in \Gamma_k^1 \cup \cdots \cup \Gamma_k^j} \mathbb{I}[\exists i \in M^1, b_s^i \leq k] \\
&\quad + \sum_{s = T_k^j + 1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq k]\Bigg) \\
\text{(by event } A_k^j) &\leq \frac{1}{t-1}\Bigg(T_k \frac{1}{4NV} + (T_k^j - T_k)\frac{1}{8NV} \\
&\quad + \sum_{s = T_k^j + 1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq k]\Bigg) \\
&\leq \frac{1}{t-1}\Bigg(T_k^j \frac{1}{4NV} + (t - 1 - T_k^j) \cdot 1\Bigg).
\end{aligned}$$

Because $T_k^j \leq t - 1 \leq T_k^{j+1} = cT_k^j$ with $c = 1 + \frac{1}{12NV}$, we have

$$P_{t-1}^i(0:k) \;\leq\; \frac{1}{4NV} + \frac{T_k^{j+1} - T_k^j}{T_k^{j+1}} \;\leq\; \frac{1}{3NV}$$

for any round $t \in \Gamma_k^{j+1}$. Then, Lemma 6.1 implies $\Pr[b_t^i = b \mid H_{t-1}, A_k^j] \leq \gamma_t$ for every $b \leq k+1$. Consider the event $[\exists i \in M^1, b_t^i \leq k+1]$. Using union bounds over $i \in M^1$

and $b \in \{0, 1, \ldots, k+1\}$,

$$\begin{aligned}
&\Pr\left[\exists i \in M^1, b_t^i \leq k+1 \;\middle|\; H_{t-1}, A_k^j\right] \\
&\leq |M^1| \cdot \Pr\left[b_t^i \leq k+1 \;\middle|\; H_{t-1}, A_k^j\right] \\
&\leq |M^1|(k+2)\gamma_t \quad \leq \quad |M^1|V\gamma_t.
\end{aligned}$$

Let $Z_t = \mathbb{I}[\exists i \in M^1, b_t^i \leq k+1] - |M^1|V\gamma_t$ and let $X_t = \sum_{s=T_k^j+1}^{t} Z_s$. We have $\mathbb{E}[Z_t \mid A_k^j, H_{t-1}] \leq 0$. Therefore, the sequence $X_{T_k^j+1}, X_{T_k^j+2}, \ldots, X_{T_k^{j+1}}$ is a supermartingale (with respect to the sequence of history $H_{T_k^j}, H_{T_k^j+1}, \ldots, H_{T_k^{j+1}-1}$). By Azuma's inequality, for any $\Delta > 0$, we have

$$\Pr\left[\sum_{t=T_k^j+1}^{T_k^{j+1}} Z_t \geq \Delta \;\middle|\; A_k^j\right] \leq \exp\left(-\frac{\Delta^2}{2|\Gamma_k^{j+1}|}\right).$$

Let $\Delta = \frac{|\Gamma_k^{j+1}|}{24NV}$. Then with probability at least $1 - \exp\left(-\frac{|\Gamma_k^{j+1}|}{1152N^2V^2}\right)$ we have $\sum_{t \in \Gamma_k^{j+1}} Z_t < \frac{|\Gamma_k^{j+1}|}{24NV}$, implying $\sum_{t \in \Gamma_k^{j+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq k+1] < \frac{|\Gamma_k^{j+1}|}{24NV} + \sum_{t \in \Gamma_k^{j+1}} |M^1|V\gamma_t \;\leq\; \frac{|\Gamma_k^{j+1}|}{24NV} + |M^1|V\frac{|\Gamma_k^{j+1}|}{12N^2V^2} \;\leq\; \frac{|\Gamma_k^{j+1}|}{8NV}$, which proves the claim. □

Proof of Lemma 6.4. Suppose A_k holds. We want to show that A_{k+1} holds with high probability. Using Claim 6.1 with $j = 0, 1, \ldots, d-1$, we have, with probability at least

$1 - \sum_{j=1}^{d} \exp\left(- \frac{|\Gamma_k^j|}{1152N^2V^2}\right)$, all the events $B_k^1, \ldots, B_k^d$ hold, which implies

$$\begin{aligned}
&\frac{1}{T_{k+1}} \sum_{t=1}^{T_{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq k+1] \\
&\leq \frac{1}{T_{k+1}} \Big(T_k \cdot 1 + \sum_{t \in \Gamma_k^1 \cup \cdots \cup \Gamma_k^d} \mathbb{I}[\exists i \in M^1, b_t^i \leq k+1] \Big) \\
&\leq \frac{1}{T_{k+1}} \Big(T_k \cdot 1 + (T_{k+1} - T_k) \cdot \frac{1}{8NV} \Big) \\
&\leq \frac{1}{8NV} + \Big(1 - \frac{T_k}{T_{k+1}} \Big) \frac{1}{8NV} \quad \leq \quad \frac{1}{4NV},
\end{aligned}$$

where in the third inequality we used $T_{k+1} \geq (8NV)T_k$. Thus A_{k+1} holds. □

Using induction from $k = 0, 1, \ldots$ to $v^1 - 4$, we have all events $A_0, A_1, \ldots, A_{v^1-3}$ happen with probability at least $1 - \exp\left(- \frac{T_b}{24NV}\right) - \sum_{k=0}^{v^1-4} \sum_{j=1}^{d} \exp\left(- \frac{|\Gamma_k^j|}{1152N^2V^2}\right)$. We then lower bound the probability, obtaining the following corollary:

Corollary 6.1. *Suppose* $|M^1| \geq 2$. $\Pr[A_{v^1-3}] \geq 1 - \exp\left(- \frac{T_b}{24NV}\right) - 2\exp\left(- \frac{T_b}{1152N^2V^2}\right)$.

Proof. Using Lemma 6.3 and Lemma 6.4 from $k = 0$ to $v_1 - 4$, we get

$$\begin{aligned}
&\Pr\left[A_{v^1-3}\right] \geq \Pr\left[A_0, A_1, \ldots, A_{v^1-3}\right] \\
&\geq 1 - \exp\left(-\tfrac{T_b}{24NV}\right) - \sum_{k=0}^{v^1-4} \sum_{j=1}^{d} \exp\left(- \tfrac{|\Gamma_k^j|}{1152N^2V^2}\right).
\end{aligned}$$

Note that $|\Gamma_k^j| = T_k^j - T_k^{j-1} = cT_k^{j-1} - cT_k^{j-2} = c|\Gamma_k^{j-1}|$ for any $k \in \{0, 1, \ldots, v^1 - 4\}$ and $j \in \{2, \ldots, d\}$, and that $|\Gamma_k^1| = c|\Gamma_{k-1}^d|$ for any $k \in \{1, 2, \ldots, v^1 - 4\}$. We also note that

$|\Gamma_0^1| = (c-1)T_0 = T_b$. Thus,

$$\begin{aligned}
&\sum_{k=0}^{v^1-4}\sum_{j=1}^{d}\exp\left(-\tfrac{|\Gamma_k^j|}{1152N^2V^2}\right)\\
&= \sum_{s=0}^{(v^1-3)d-1}\exp\left(-\tfrac{c^sT_b}{1152N^2V^2}\right)\\
&\leq \sum_{s=0}^{\infty}\exp\left(-\tfrac{c^sT_b}{1152N^2V^2}\right)\\
&= \exp\left(-\tfrac{T_b}{1152N^2V^2}\right)\left(1+\sum_{s=1}^{\infty}\exp\left(-\tfrac{(c^s-1)T_b}{1152N^2V^2}\right)\right).
\end{aligned}$$

It remains to prove $\sum_{s=1}^{\infty}\exp\left(-\frac{(c^s-1)T_b}{1152N^2V^2}\right) \leq 1$. Since $c^s-1 \geq c-1+(s-1)(c^2-c), \forall s \geq 1$, we have

$$\begin{aligned}
&\sum_{s=1}^{\infty}\exp\left(-\tfrac{(c^s-1)T_b}{1152N^2V^2}\right)\\
&\leq \sum_{s=1}^{\infty}\exp\left(-\tfrac{(c-1)T_b}{1152N^2V^2}\right)\left(\exp\left(-\tfrac{(c^2-c)T_b}{1152N^2V^2}\right)\right)^{s-1}\\
&\leq \sum_{s=1}^{\infty}\left(\tfrac{1}{2}\right)^s = 1,
\end{aligned}$$

where the second "$\leq$" is because $\exp\left(-\frac{(c^2-c)T_b}{1152N^2V^2}\right) \leq \exp\left(-\frac{(c-1)T_b}{1152N^2V^2}\right) \leq \frac{1}{2}$ by the choice of T_b. □

6.4.3 Eliminating $v^1 - 2$

Assume that A_{v^1-3} has happened. We continue partitioning the time horizon after T_{v^1-3}, all the way to infinity, to show two points: (1) the frequency of bids in $\{0, 1, \ldots, v^1-3\}$ from bidders in M^1 approaches 0; (2) the frequency of v^1-2 also approaches 0. Again let $c = 1 + \frac{1}{12NV}$. Let $T_a^0 = T_{v^1-3}, T_a^{k+1} = cT_a^k, \Gamma_a^{k+1} = [T_a^k+1, T_a^{k+1}], k \geq 0$. Let

$\delta_t = (\frac{1}{t})^{\frac{1}{3}}, t \geq 0$. For each $k \geq 0$, define

$$F_{T_a^k} = \tfrac{1}{4NVc^k} + \sum_{s=0}^{k-1} \tfrac{c-1}{c^{k-s}} \delta_{T_a^s} + \sum_{s=0}^{k-1} |M^1| V \tfrac{c-1}{c^{k-s}} \gamma_{T_a^s},$$

and

$$\widetilde{F}_{T_a^k} = \tfrac{1}{c^k} + \sum_{s=0}^{k-1} \tfrac{c-1}{c^{k-s}} \delta_{T_a^s} + \sum_{s=0}^{k-1} |M^1| V \tfrac{c-1}{c^{k-s}} \gamma_{T_a^s}.$$

Claim 6.2. *If T_b is sufficiently large such that $\delta_{T_a^k} + |M^1| V \gamma_{T_a^k} \leq \frac{1}{4NV}$, then $F_{T_a^{k+1}} \leq F_{T_a^k} \leq \frac{1}{4NV}$ for every $k \geq 0$ and $\lim_{k \to \infty} F_{T_a^k} = \lim_{k \to \infty} \widetilde{F}_{T_a^k} = 0$.*

Lemma 6.5. *Suppose $|M^1| \geq 2$. Let T_b be any sufficiently large constant. Let A_a^k be the event*

$$\left[\forall s \leq k, \; \frac{1}{T_a^s} \sum_{t=1}^{T_a^s} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \leq F_{T_a^s} \right]$$

Then, $\Pr[A_a^k] \geq 1 - \exp\left(-\frac{T_b}{24NV}\right) - 2\exp\left(-\frac{T_b}{1152N^2V^2}\right) - 2\exp\left(-(\frac{T_b}{1152N^2V^2})^{\frac{1}{3}}\right)$*. Moreover, if $|M^1| \geq 3$, we can include the following in event A_a^k: $\forall s \leq k$, $\frac{1}{T_a^s} \sum_{t=1}^{T_a^s} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 2] \leq \widetilde{F}_{T_a^s}$.*

The proof of Lemma 6.5 is similar to Lemma 6.4 except that we use Lemma 6.2 to argue that bidders bid $v^1 - 2$ with low frequency. See details in the appendix.

Proof of Theorem 6.1. Suppose $|M^1| \geq 3$. We note that the event A_a^k implies that for any time $t \in \Gamma_a^k = [T_a^{k-1}+1, T_a^k]$,

$$\begin{aligned}
&\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 2] \\
&\le \frac{1}{t}\sum_{s=1}^{T_a^k}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 2] \\
&\le \frac{1}{t}T_a^k \widetilde{F}_{T_a^k} \\
&\le c\widetilde{F}_{T_a^k} \qquad \text{(because } t \ge \tfrac{1}{c}T_a^k\text{)}. \qquad (6.4)
\end{aligned}$$

We note that $A_a^{k-1} \supseteq A_a^k$, so by Lemma 6.5 with probability at least $\Pr[\cap_{k=0}^{\infty} A_a^k] = \lim_{k\to\infty}\Pr[A_a^k] \ge 1 - \exp\left(-\frac{T_b}{24NV}\right) - 2\exp\left(-\frac{T_b}{1152N^2V^2}\right) - 2\exp\left(-(\frac{T_b}{1152N^2V^2})^{\frac{1}{3}}\right)$ all events $A_a^0, A_a^1, \ldots, A_a^k, \ldots$ happen. Then, according to (6.4) and Claim 6.2, we have

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 2] \le \lim_{k\to\infty} c\widetilde{F}_{T_a^k} = 0.$$

Letting $T_b \to \infty$ proves the first result of the theorem. The second result follows from the observation that, when $\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 2] \le \frac{1}{3NV}$, all bidders in M^1 will choose bids in $\{0, 1, \ldots, v^1 - 2\}$ with probability at most $(v^1 - 1)\gamma_{t+1}$ in round $t+1$ according to Lemmas 6.1 and 6.2, and that $(v^1 - 1)\gamma_{t+1} \to 0$ as $t \to \infty$. □

6.5 Experimental Results

6.5.1 $|M^1| = 2$: Convergence to Two Equilibria

For the case of $|M^1| = 2$, we showed in Theorem 6.2 that any mean-based algorithm must converge to one of the two equilibria where the two players in M^1 bid $v^1 - 1$ or $v^1 - 2$. One may wonder whether there is a theoretical guarantee of which equilibrium will be obtained. We give experimental results to show that, in fact, *both* equilibria can be obtained under a *same* randomized mean-based algorithm in different runs. We demonstrate this by the ε_t-Greedy algorithm (defined in Example 6.1). Interestingly, under the same setting, the MWU algorithm always converges to the equilibrium of $v^1 - 1$. In the experiment, we let $n = |M^1| = 2$, $v^1 = v^2 = V = 4$.

ε_t-Greedy converges to two equilibria We run ε_t-Greedy with $\varepsilon_t = \sqrt{1/t}$ for 1000 simulations. In each simulations, we run it for $T = 1000$ rounds. After it finishes, we use the frequency of bids from bidder 1 to determine which equilibrium the algorithm will converge to: if the frequency of bid 2 is above 0.8, we consider it converging to the equilibrium of $v^1 - 2$; if the frequency of bid 3 is above 0.8, we consider it converging to the equilibrium of $v^1 - 1$; if neither happens, we consider it as "not converged yet". Among the 1000 simulations, we found 774 times of converging to $v^1 - 1$, 226 times of converging to $v^1 - 2$, and 0 times of "not converged yet".

We give two figures of the changes of bid frequencies and mixed strategies of player 1 and 2: Figure 6.1 is for $v^1 - 1$; Figure 6.2 is for the case of converging to $v^1 - 2$. The x-axis is round number t and the y-axis is the frequency $\frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[b_s^i = b]$ of each bid $b \in \{0, 1, 2, 3\}$ or the mixed strategy $\boldsymbol{x}_t^i = (x_t^i(0), x_t^i(1), x_t^i(2), x_t^i(3))$. For clarity, we only show the first 500 rounds.

MWU always converges to $v^1 - 1$ We run MWU with $\varepsilon_t = \sqrt{1/t}$. Same as the previous experiment, we run the algorithm for 1000 simulations and count how many times the algorithm converges to the equilibrium of $v^1 - 2$ and $v^1 - 1$. We found that, in all 1000 simulations, MWU converged to $v^1 - 1$. Figure 6.3 shows the changes of bid frequencies and mixed strategies of both players.

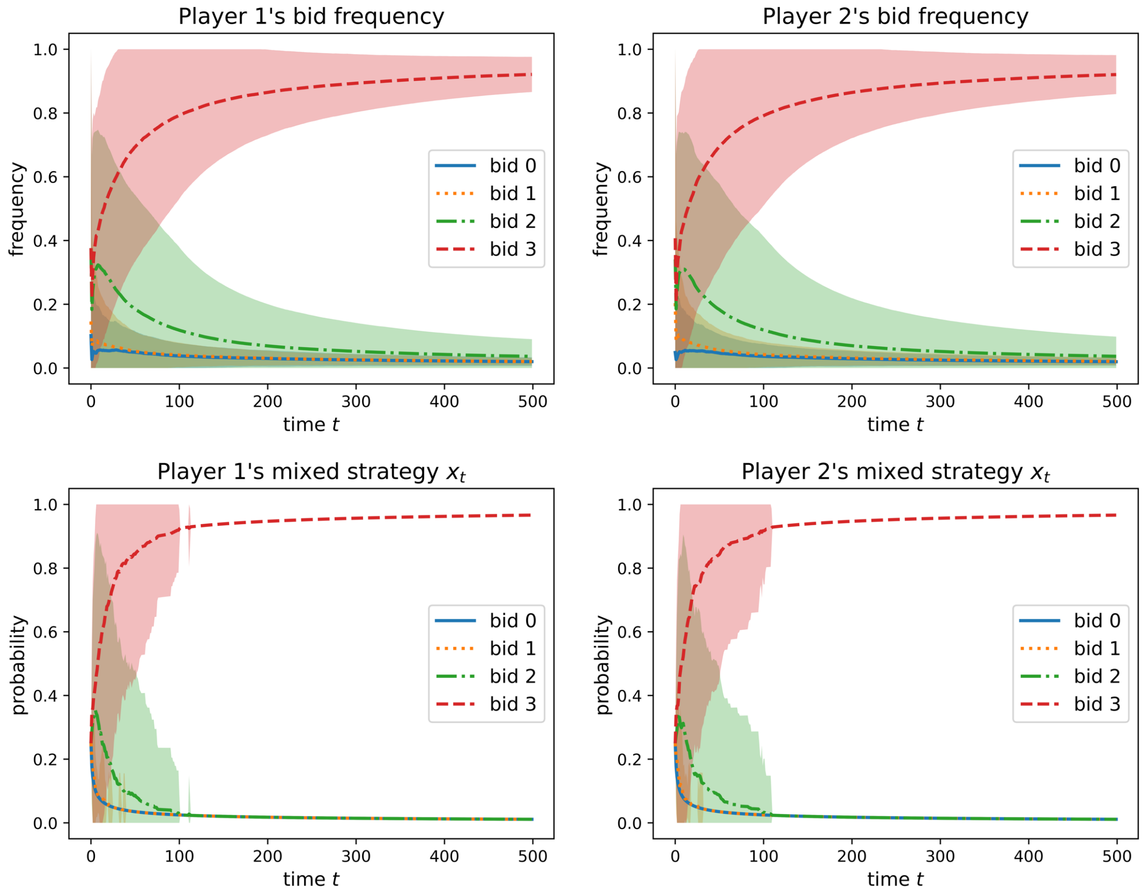


Figure 6.1: Player 1 and 2's bid frequencies and mixed strategies in the case of $|M^1| = 2$, $v^1 = v^2 = 4$, using ε_t-Greedy algorithm, and converging to the $v^1 - 1 = 3$ equilibrium. Converging to $v^1 - 1$ happens in 774 out of 1000 simulations. The curves and the shaded regions are the means and 2-standard deviation intervals among the 774 simulations.

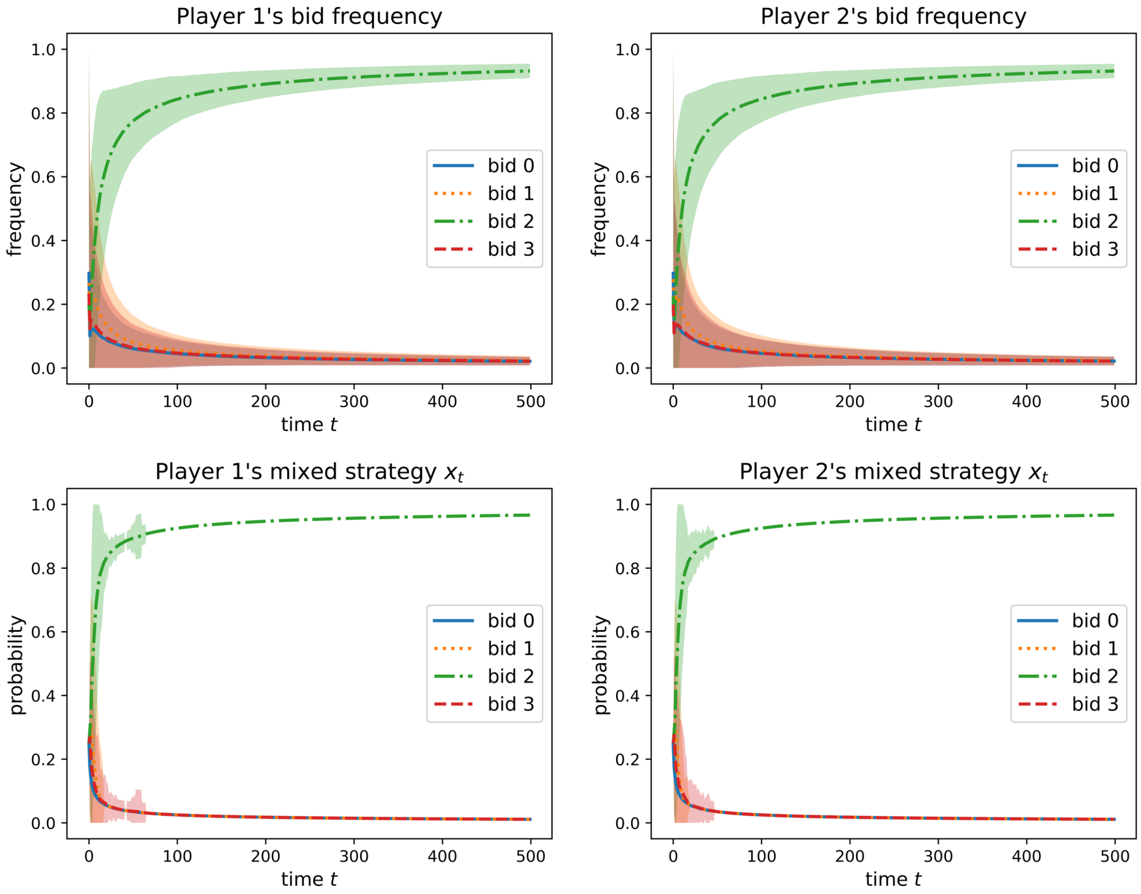


Figure 6.2: Player 1 and 2's bid frequencies and mixed strategies in the case of $|M^1| = 2$, $v^1 = v^2 = 4$, using ε_t-Greedy algorithm, and converging to the $v^1 - 2 = 2$ equilibrium. Converging to $v^1 - 2$ happens in 226 out of 1000 simulations. The curves and the shaded regions are the means and 2-standard deviation intervals among the 226 simulations.

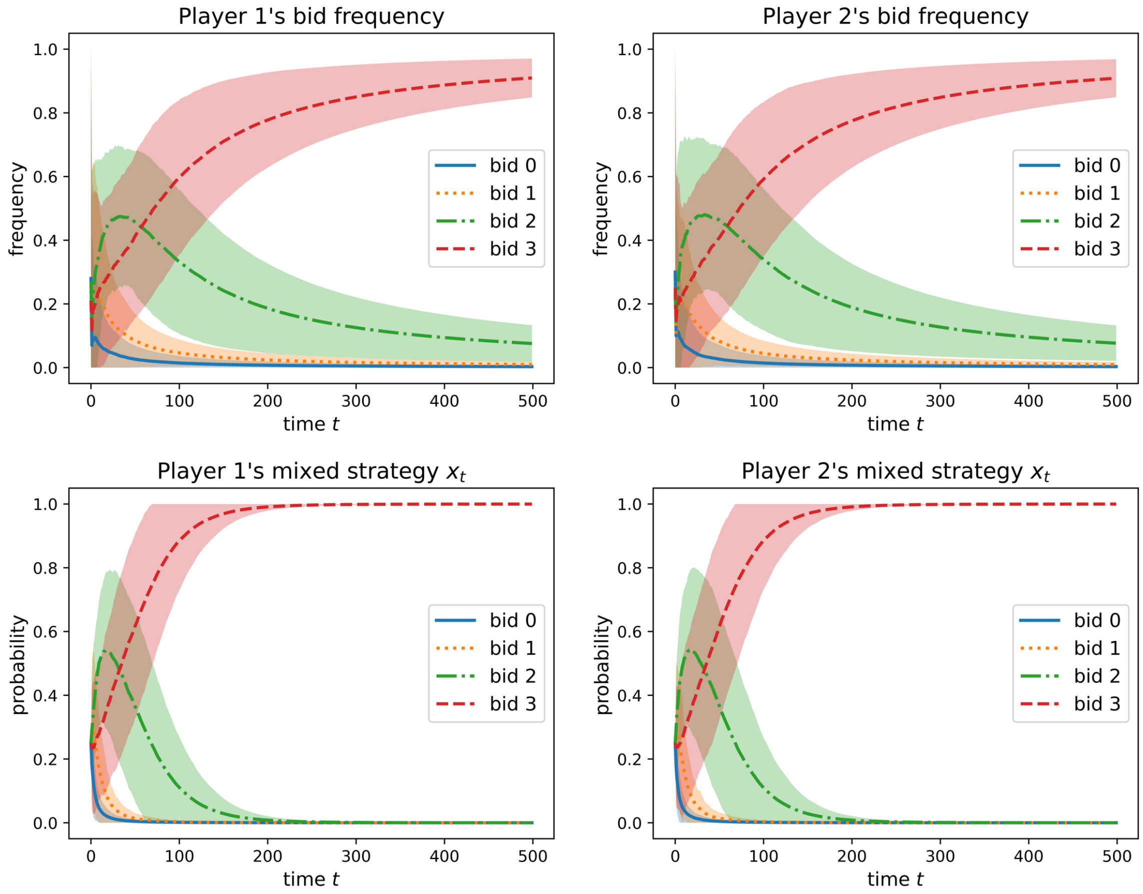


Figure 6.3: Player 1 and 2's bid frequencies and mixed strategies in the case of $|M^1| = 2$, $v^1 = v^2 = 4$, using MWU algorithm. The curves and the shaded regions are the means and 2-standard deviation intervals of 1000 simulations.

6.5.2 $|M^1| = 1$: Non-Convergence

For the case of $|M^1| = 1$, we showed that not all mean-based algorithms can converge to equilibrium, using the example of Follow the Leader (Example 6.2). Here we experimentally demonstrate that such non-convergence phenomena can also happen with more natural (and even no-regret) mean-based algorithms like ε-Greedy and MWU.

In the experiment we let $n = 2$, $v^1 = 8, v^2 = 6$. We run ε_t-Greedy and MWU both with $\varepsilon_t = 1/\sqrt{t}$ for $T = 20000$ rounds.

For ε_t-Greedy, Figure 6.4 shows that the two bidders do not converge to a pure-strategy equilibrium, either in time-average or last-iterate. According to Proposition 6.1, a pure-strategy equilibrium must have bidder 1 bidding $v^2 = 6$ and bidder 2 bidding $v^2 - 1 = 5$. But figure (b) shows that bidder 2's frequency of bidding 5 does not converge to 1. The frequency oscillates and we do not know whether it will stabilize at some limit less than 1. Looking closer, we see that bidder 2 constantly switches between bids 5 and 3, and bidder 1 switches between 5 and 6. Intuitively, this is because: in the ε_t-Greedy algorithm, when bidder 1 bids $v^2 = 6$ with high probability, she also sometimes (with probability ε_t) chooses bids uniformly at random, in which case the best response for bidder 2 is to bid $v^2/2 = 3$; but after bidder 2 switches to 3, bidder 1 will find it beneficial to lower her bid from 6 to 5; then, bidder 2 will switch to 5 to compete with bidder 1, winning the item with probability 1/2; but then bidder 1 will increase to 6 to outbid bidder 2; ... In this way, they enter a cycle.

For MWU, Figure 6.5 shows that bidder 1's bid frequency and mixed strategy seem to converge to bidding $v^2 = 6$ (left two figures); while it seems unclear whether bidder 2's bid frequency and mixed strategy converge (right two figures).

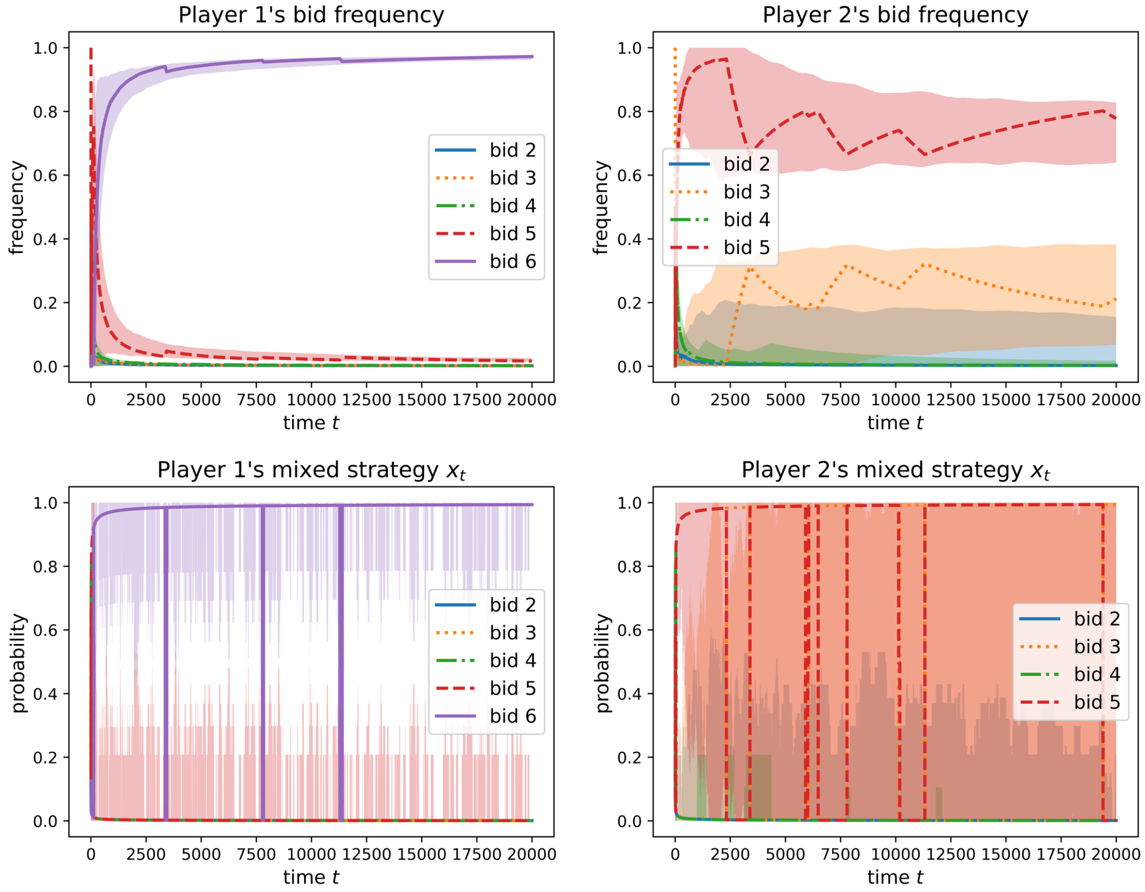


Figure 6.4: Player 1 and 2's bid frequencies and mixed strategies in the case of $|M^1| = 1$, $v^1 = 8, v^2 = 6$, using ε_t-Greedy algorithm. The curves are the results from one simulation. The shaped regions are 2-standard deviation intervals from 100 simulations. Player 1's frequency of bid 6 seems to converge to 1, but the mixed strategy does not last-iterate converge; it switches between bids 5 and 6. Player 2's bid frequency oscillates; the mixed strategy switches between bids 3 and 5.

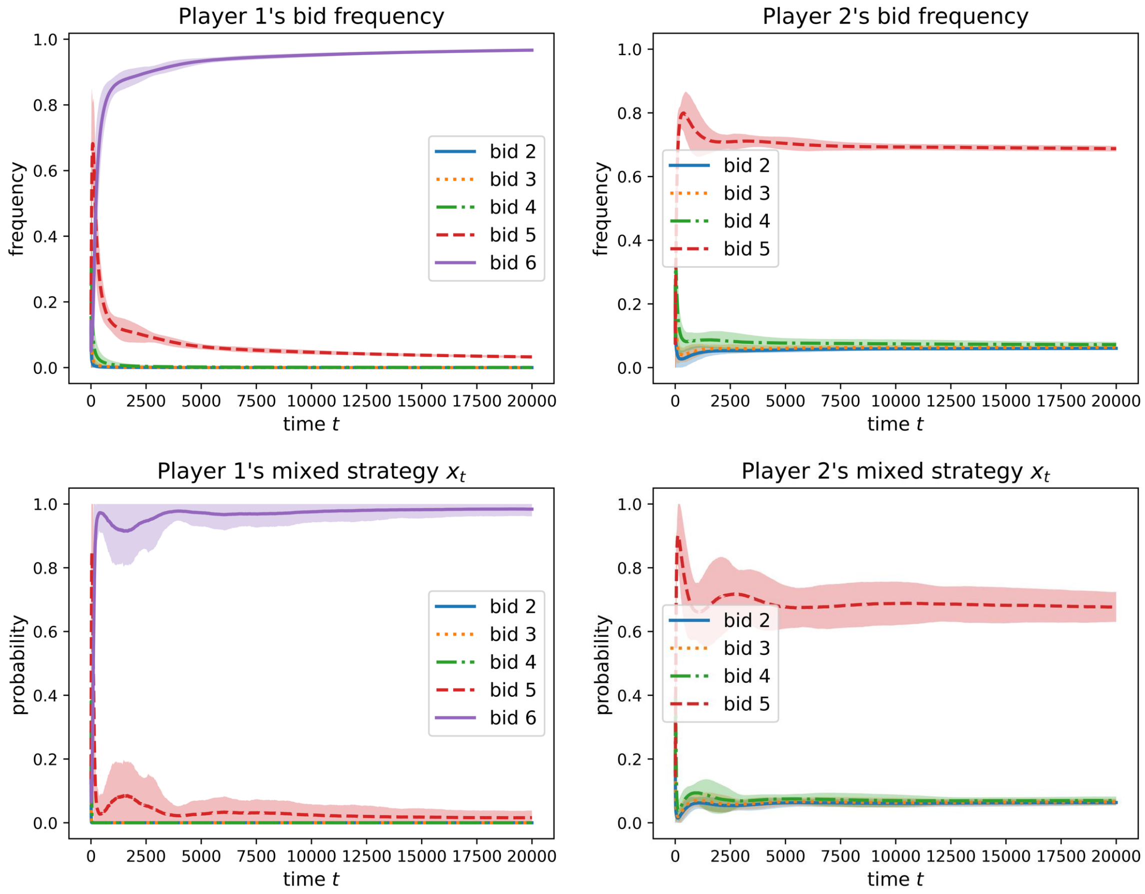


Figure 6.5: Player 1 and 2's bid frequencies and mixed strategies in the case of $|M^1| = 1$, $v^1 = 8, v^2 = 6$, using MWU algorithm. The curves and the shaded regions are the means and 2-standard deviation intervals of 100 simulations.

6.6 Disucssion

This work showed that, in repeated first price auctions with fixed values, mean-based learning bidders converge to a Nash equilibrium in the presence of *competition*, in the sense that at least two bidders share the highest value. Without competition, we gave non-convergence examples using mean-based algorithms that are not necessarily no-regret. Understanding the convergence property of no-regret algorithms in the absence of competition is a natural and interesting future direction.

The convergence result we give is in the limit sense. As observed by [WXY22], many no-regret algorithms actually need an exponential time to converge to Nash equilibrium in some iterative-dominance-solvable game. Our theoretical analysis for the first price auction demonstrates a $T = O(c^{O(v^1)})$ upper bound on the convergence time for the case of $|M^1| = 3$. But the convergence time in our experiments is significantly shorter. The exact convergence rate remains open.

Analyzing repeated first price auctions where bidders have time-varying values is also a natural, yet possibly challenging, future direction.

6.7 Omitted Proofs in Section 6.3

6.7.1 Proof of Theorem 6.2

Suppose $|M^1| = 2$. We will prove that, for any sufficiently large integer T_b, with probability at least $1 - \exp\left(-\frac{T_b}{24NV}\right) - 2\exp\left(-\frac{T_b}{1152N^2V^2}\right) - \frac{6}{e-2}\left(\frac{48NV}{T_b}\right)^{3e/4}$, one of following two events must happen:

- $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1, b_s^i = v^1 - 2] = 1$;
- $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1, b_s^i = v^1 - 1] = 1$ and $\lim_{t\to\infty} \Pr[b_t^i = v^1 - 1] = 1$.

And if $n \geq 3$ and $v^3 = v^1 - 1$, only the second event happens. Letting $T_b \to \infty$ proves Theorem 6.2.

We reuse the argument in Section 6.4.2. Assume $v^1 \geq 3$.[2] Recall that we defined $c = 1 + \frac{1}{12NV}$, $d = \lceil \log_c(8NV) \rceil$; T_b is any integer such that $\gamma_{T_b} < \frac{1}{12N^2V^2}$ and $\exp\left(-\frac{(c-1)T_b}{1152N^2V^2}\right) \leq \frac{1}{2}$; $T_0 = 12NVT_b$; $T_{v^1-3} = c^{(v^1-3)d}T_0$. We defined A_{v^1-3} to be the event $\frac{1}{T_{v^1-3}} \sum_{t=1}^{T_{v^1-3}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \leq \frac{1}{4NV}$. According to Corollary 6.1, A_{v^1-3} holds with probability at least $1 - \exp\left(-\frac{T_b}{24NV}\right) - 2\exp\left(-\frac{T_b}{1152N^2V^2}\right)$. Suppose A_{v^1-3} holds.

Now we partition the time horizon after T_{v^1-3} as follows: let $T_a^0 = T_{v^1-3}, T_a^k = C(k + 24NV)^2$, $\forall k \geq 0$, where $C = \frac{T_{v^1-3}}{(24NV)^2}$, so that $T_a^0 = C(0 + 24NV)^2$. Denote $\Gamma_a^{k+1} = [T_a^k + 1, T_a^{k+1}]$, with $|\Gamma_a^{k+1}| = T_a^{k+1} - T_a^k$. (We note that the notations here have different meanings than those in Section 6.4.3.) We define $\delta_t = (\frac{1}{t})^{1/8}, t \geq 0$. For each $k \geq 0$, we define

$$F_{T_a^k} = \frac{T_a^0}{T_a^k}\frac{1}{4NV} + \sum_{s=0}^{k-1} \frac{T_a^{s+1} - T_a^s}{T_a^k}\delta_{T_a^s} + \sum_{s=0}^{k-1} \frac{T_a^{s+1} - T_a^s}{T_a^k}|M^1|V\gamma_{T_a^s}.$$

Let A_a^k be event

$$A_a^k = \left[\frac{1}{T_a^k}\sum_{t=1}^{T_a^k} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \leq F_{T_a^k}\right].$$

We note that $A_a^0 = A_{v^1-3}$ because $F_{T_a^0} = \frac{1}{4NV}$.

In the proof we will always let T_b to be sufficiently large. This implies that all the times $T_0, T_{v^1-3}, T_a^0, T_a^k$, etc., are sufficiently large.

Additional Notations, Claims, and Lemmas

Claim 6.3. *When T_b is sufficiently large,*

- $F_{T_a^{k+1}} \leq F_{T_a^k} \leq \frac{1}{4NV}$ *for every* $k \geq 0$.
- $\lim_{k\to\infty} F_{T_a^k} = 0$.

[2] If $v^1 = 1$, Theorem 6.2 trivially holds. If $v^1 = 2$, we let $T_{v^1-3} = T_0 = T_b$; A_{v^1-3} holds with probability 1 since $\frac{1}{T_{v^1-3}} \sum_{t=1}^{T_{v^1-3}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] = 0$; the argument for $v^1 \geq 3$ will still apply.

Proof. Since $\delta_{T_a^0} \to 0$ and $\gamma_{T_a^0} \to 0$ as $T_b \to \infty$, when T_b is sufficiently large we have

$$F_{T_a^1} = \frac{T_a^0}{T_a^1}\frac{1}{4NV} + \frac{T_a^1 - T_a^0}{T_a^1}\left(\delta_{T_a^0} + |M^1|V\gamma_{T_a^0}\right) \le \frac{T_a^0}{T_a^1}\frac{1}{4NV} + \frac{T_a^1 - T_a^0}{T_a^1}\frac{1}{4NV} = \frac{1}{4NV} = F_{T_a^0}.$$

Since $\delta_{T_a^s}$ and $\gamma_{T_a^s}$ are both decreasing, we have

$$\begin{aligned} F_{T_a^k} &> \sum_{s=0}^{k-1}\frac{T_a^{s+1} - T_a^s}{T_a^k}\delta_{T_a^s} + \sum_{s=0}^{k-1}\frac{T_a^{s+1} - T_a^s}{T_a^k}|M^1|V\gamma_{T_a^s} \\ &\ge \sum_{s=0}^{k-1}\frac{T_a^{s+1} - T_a^s}{T_a^k}\delta_{T_a^k} + \sum_{s=0}^{k-1}\frac{T_a^{s+1} - T_a^s}{T_a^k}|M^1|V\gamma_{T_a^k} \ = \ \delta_{T_a^k} + |M^1|V\gamma_{T_a^k}. \end{aligned}$$

Thus,

$$F_{T_a^{k+1}} \overset{\text{by definition}}{=} \frac{T_a^k}{T_a^{k+1}}F_{T_a^k} + \frac{T_a^{k+1} - T_a^k}{T_a^{k+1}}\left(\delta_{T_a^k} + |M^1|V\gamma_{T_a^k}\right) < \frac{T_a^k}{T_a^{k+1}}F_{T_a^k} + \frac{T_a^{k+1} - T_a^k}{T_a^{k+1}}F_{T_a^k} = F_{T_a^k}.$$

Then we prove $\lim_{k\to\infty} F_{T_a^k} = 0$. For every $0 < \varepsilon < \frac{1}{4NV}$, we can find k sufficiently large such that $\delta_{T_a^k} \le \frac{\varepsilon}{6}$, and $\gamma_{T_a^k} \le \frac{\varepsilon}{6|M^1|V}$. For any $l \ge \lceil k/\varepsilon \rceil$, we have $\frac{T_a^0}{T_a^l} \le \frac{T_a^k}{T_a^l} \le \frac{\varepsilon}{6}$. Then

$$\begin{aligned} F_{T_a^l} &= \frac{T_a^0}{T_a^l}\frac{1}{4NV} + \sum_{s=0}^{l-1}\frac{T_a^{s+1} - T_a^s}{T_a^l}(\delta_{T_a^s} + |M^1|V\gamma_{T_a^s}) \\ &\le \frac{\varepsilon}{3} + 2\sum_{s=0}^{k-1}\frac{T_a^{s+1} - T_a^s}{T_a^l} + \sum_{s=k}^{l-1}\frac{T_a^{s+1} - T_a^s}{T_a^l}(\delta_{T_a^k} + |M^1|V\gamma_{T_a^k}) \\ &\le \frac{\varepsilon}{3} + 2\frac{T_a^k}{T_a^l} + \delta_{T_a^k} + |M^1|V\gamma_{T_a^k} \\ &\le \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon. \end{aligned}$$

Since $F_{T_a^k}$ is non-negative, we have $\lim_{k\to\infty} F_{T_a^k} = 0$. □

Claim 6.4. $\sum_{s=0}^{\infty}\exp\left(-\frac{1}{2}|\Gamma_a^{s+1}|\delta_{T_a^s}^2\right) \le \frac{2}{e-2}\frac{1}{C^{3e/4}} \le \frac{2}{e-2}\left(\frac{48NV}{T_b}\right)^{3e/4}$.

Proof. Recall that $|\Gamma_a^{s+1}| = T_a^{s+1} - T_a^s$, $\delta^2_{T_a^s} = (\frac{1}{T_a^s})^{1/8}$, and $T_a^s = C(s + 24NV)^2$. Hence,

$$\begin{aligned}
\sum_{s=0}^{\infty} \exp\left(-\frac{1}{2}|\Gamma_a^{s+1}|\delta^2_{T_a^s}\right) &= \sum_{s=0}^{\infty} \exp\left(-\frac{1}{2}\big(T_a^{s+1} - T_a^s\big)\big(\frac{1}{T_a^s}\big)^{1/4}\right) \\
&= \sum_{s=0}^{\infty} \exp\left(-\frac{1}{2}C\big(2(s+24NV)+1\big)\big(\frac{1}{C(s+24NV)^2}\big)^{1/4}\right) \\
&\leq \sum_{s=0}^{\infty} \exp\left(-C^{3/4}\big(s+24NV\big)\big(\frac{1}{s+24NV}\big)^{1/2}\right) \\
&= \sum_{s=0}^{\infty} \exp\left(-C^{3/4}\sqrt{s+24NV}\right) \\
&\leq \sum_{x=2}^{\infty} \exp\left(-C^{3/4}\sqrt{x}\right) \\
&\leq \int_{x=1}^{\infty} \exp\left(-C^{3/4}\sqrt{x}\right) \mathrm{d}x \\
\text{(using } e^x \geq x^e \text{ for } x \geq 0\text{)} \quad &\leq \int_{x=1}^{\infty} \frac{1}{(C^{3/4}\sqrt{x})^e} \mathrm{d}x = \frac{1}{C^{3e/4}} \cdot \frac{2}{e-2}.
\end{aligned}$$

Substituting $C = \frac{T_{v^1-3}}{(24NV)^2} = \frac{c^{(v^1-3)d}12NVT_b}{(24NV)^2} \geq \frac{12NVT_b}{(24NV)^2} = \frac{T_b}{48NV}$ proves the claim. □

Claim 6.5. $\frac{T_a^k}{T_a^{k+1}} \geq 1 - \frac{2}{k+24NV}$.

Proof. By definition, $\frac{T_a^k}{T_a^{k+1}} = \frac{(k+24NV)^2}{(k+24NV+1)^2} = 1 - \frac{2(k+24NV)+1}{(k+24NV+1)^2} \geq 1 - \frac{2}{k+24NV+1} \geq 1 - \frac{2}{k+24NV}$. □

Claim 6.6. *When A_a^k holds, we have, for every $t \in \Gamma_a^{k+1} = [T_a^k + 1, T_a^{k+1}]$, $\frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \leq F_{T_a^k} + \frac{2}{k+24NV} \leq \frac{1}{2NV} - 2\gamma_t$.*

Proof. When A_a^k holds, for every $t \in \Gamma_a^{k+1}$,

$$\begin{aligned}\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 3] &\le \frac{1}{t-1}\left(T_a^k F_{T_a^k} + (t-1-T_a^k)\right) \\ (\text{since } T_a^k \le t-1 \le T_a^{k+1}) &\le F_{T_a^k} + \frac{T_a^{k+1} - T_a^k}{T_a^{k+1}} \\ (\text{by Claim 6.5}) &\le F_{T_a^k} + \frac{2}{k+24NV}.\end{aligned}$$

Since $F_{T_a^k} \le \frac{1}{4NV}$ by Claim 6.3 and $\gamma_t \le \frac{1}{12N^2V^2}$ by assumption, the above expression is further bounded by $\frac{1}{4NV} + \frac{2}{k+24NV} \le \frac{1}{4NV} + \frac{2}{24NV} = \frac{1}{3NV} \le \frac{1}{2NV} - 2\gamma_t$. □

Lemma 6.6. *For every $k \ge 0$,* $\Pr[A_a^{k+1} \mid A_a^k] \ge 1 - \exp\left(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2\right)$.

Proof. Given A_a^k, according to Claim 6.6, it holds that for every $t \in \Gamma_a^{k+1}$, $\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1 - 3] \le \frac{1}{2NV} - 2\gamma_t$. Then according to Lemma 6.1, for any history H_{t-1},

$$\Pr[\exists i \in M^1, b_t^i \le v^1 - 3 \mid H_{t-1}, A_a^k] \le |M^1|V\gamma_t.$$

Let $Z_t = \mathbb{I}[\exists i \in M^1, b_t^i \le v^1 - 3] - |M^1|V\gamma_t$ and let $X_t = \sum_{s=T_a^k+1}^{t} Z_s$. We have $\mathbb{E}[Z_t \mid H_{t-1}, A_a^k] \le 0$. Therefore, the sequence $X_{T_a^k+1}, X_{T_a^k+2}, \ldots, X_{T_a^{k+1}}$ is a supermartingale (with respect to the sequence of history $H_{T_a^k}, H_{T_a^k+1}, \ldots, H_{T_a^{k+1}-1}$). By Azuma's inequality, for any $\Delta > 0$, we have

$$\Pr\left[\sum_{t\in\Gamma_a^{k+1}} Z_t \ge \Delta \,\middle|\, A_a^k\right] \le \exp\left(-\frac{\Delta^2}{2|\Gamma_a^{k+1}|}\right).$$

Let $\Delta = |\Gamma_a^{k+1}|\delta_{T_a^k}$. Then with probability at least $1 - \exp\big(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2\big)$, we get $\sum_{t\in\Gamma_a^{k+1}}\mathbb{I}[\exists i \in M^1, b_t^i \le v^1 - 3] < \Delta + |M^1|V\sum_{t\in\Gamma_a^{k+1}}\gamma_t \le |\Gamma_a^{k+1}|\delta_{T_a^k} + |M^1|V|\Gamma_a^{k+1}|\gamma_{T_a^k}$,

which implies

$$\begin{aligned}
&\frac{1}{T_a^{k+1}}\sum_{t=1}^{T_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \le v^1-3] \\
&= \frac{1}{T_a^{k+1}}\Bigg(\sum_{t=1}^{T_a^{k}} \mathbb{I}[\exists i \in M^1, b_t^i \le v^1-3] + \sum_{t\in\Gamma_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \le v^1-3]\Bigg) \\
&\le \frac{1}{T_a^{k+1}}\left(T_a^k F_{T_a^k} + |\Gamma_a^{k+1}|\delta_{T_a^k} + |M^1||V||\Gamma_a^{k+1}|\gamma_{T_a^k}\right) \\
&= F_{T_a^{k+1}} \qquad \text{(by definition)}
\end{aligned}$$

and thus A_a^{k+1} holds. □

Denote by $f_t^i(b)$ the frequency of bid b in the first t rounds for bidder i: $f_t^i(b) = \frac{1}{t}\sum_{s=1}^t \mathbb{I}[b_s^i = b]$. Let $f_t^i(0 : v^1-3) = \frac{1}{t}\sum_{s=1}^t \mathbb{I}[b_s^i \le v^1-3]$.

Claim 6.7. *If the history H_{t-1} satisfies $f_{t-1}^i(v^1-1) > 2(X+V\gamma_t)$ and $\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b_s^i \le v^1-3] \le X$ for some $X \in [0,1]$, then we have $\Pr[b_t^{i'} = v^1-2 \mid H_{t-1}] \le \gamma_t$ for the other $i' \ne i \in M^1$.*

Proof. Consider $\alpha_{t-1}^{i'}(v^1-1)$ and $\alpha_{t-1}^{i'}(v^1-2)$. On the one hand,

$$\alpha_{t-1}^{i'}(v^1-1) = 1 \times (1 - f_{t-1}^i(v^1-1)) + \frac{1}{2}\times f_{t-1}^i(v^1-1) = 1 - \frac{1}{2} f_{t-1}^i(v^1-1). \tag{6.5}$$

On the other hand, since having more bidders with bids no larger than v^1-2 only decreases the utility of a bidder who bids v^1-2, we can upper bound $\alpha_{t-1}^{i'}(v^1-2)$ by

$$\begin{aligned}
\alpha_{t-1}^{i'}(v^1-2) &\le 2 \times f_{t-1}^i(0 : v^1-3) + 1 \times (1 - f_{t-1}^i(v^1-1) - f_{t-1}^i(0:v^1-3)) \\
&= 1 - f_{t-1}^i(v^1-1) + f_{t-1}^i(0:v^1-3) \\
&\le 1 - f_{t-1}^i(v^1-1) + X, \qquad (6.6)
\end{aligned}$$

where the last inequality holds because $f^i_{t-1}(0:v^1-3) \le \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b^i_s \le v^1-3] \le X$. Combining (6.5) and (6.6), we get

$$\alpha^{i'}_{t-1}(v^1-1) - \alpha^{i'}_{t-1}(v^1-2) \ge (1-\frac{1}{2}f^i_{t-1}) - (1-f^i_{t-1}+X) = \frac{1}{2}f^i_{t-1}(v^1-1) - X > V\gamma_t.$$

This implies $\Pr[b^{i'}_t = v^1-2 \mid H_{t-1}] \le \gamma_t$ according to the mean-based property. □

Proof of the General Case

We consider $k = 0, 1, \dots$ to ∞. For each k, we suppose $A^0_a, A^1_a, \dots, A^k_a$ hold, which happens with probability at least $1-\sum_{s=0}^{k-1}\exp\big(-\frac{1}{2}|\Gamma^{s+1}_a|\delta^2_{T^s_a}\big)$ according to Lemma 6.6, given that $A^0_a = A_{v^1-3}$ already held. The proof is divided into two cases based on $f^i_{T^k_a}(v^1-1)$.

Case 1: *For all $k \ge 0$, $f^i_{T^k_a}(v^1-1) \le 16(F_{T^k_a} + \frac{2}{k+24NV} + V\gamma_{T^k_a})$ for both $i \in M^1$.*

We argue that the two bidders in M^1 converge to playing v^1-2 in this case.

According to Lemma 6.6, all events $A^0_a, A^1_a, \dots, A^k_a, \dots$ happen with probability at least $1-\sum_{k=0}^{\infty}\exp\left(-\frac{1}{2}|\Gamma^{k+1}_a|\delta^2_{T^k_a}\right)$. Claim 6.6 and Claim 6.3 then imply that, for both $i \in M^1$,

$$\lim_{t\to\infty} f^i_t(0:v^1-3) \le \lim_{k\to\infty}\left(F_{T^k_a} + \frac{2}{k+24NV}\right) = 0.$$

Because for every $t \in \Gamma^{k+1}_a = [T^k_a+1, T^{k+1}_a]$ we have $f^i_t(v^1-1) \le \frac{T^{k+1}_a}{t} f^i_{T^k_a}(v^1-1) \le \frac{T^{k+1}_a}{T^k_a} f^i_{T^k_a}(v^1-1) \le 2f^i_{T^k_a}(v^1-1)$ and by condition $f^i_{T^k_a}(v^1-1) \to 0$ as $k \to \infty$, we have $\lim_{t\to\infty} f^i_t(v^1-1) = 0$. Therefore, $\lim_{t\to\infty} f^i_t(v^1-2) = \lim_{t\to\infty} 1 - f^i_t(0:v^1-3) - f^i_t(v^1-1) = 1$, which implies

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\mathbb{I}[\forall i \in M^1, b^i_s = v^1-2] = 1.$$

Case 2: *There exists $k \ge 0$ such that $f^i_{T^k_a}(v^1-1) > 16(F_{T^k_a} + \frac{2}{k+24NV} + V\gamma_{T^k_a})$ for some $i \in M^1$.*

If this case happens, we argue that the two bidders in M^1 converge to playing v^1-1.

We first prove that, after $\ell = k + 24NV$ periods (i.e., at time $T_a^{k+\ell}$), the frequency of $v^1 - 1$ for *both* bidders in M^1 is greater than $4(F_{T_a^{k+\ell}} + \frac{2}{(k+\ell)+24NV} + V\gamma_{T_a^{k+\ell}})$, with high probability.

Lemma 6.7. *Suppose that, at time T_a^k, A_a^k holds and for some $i \in M^1$, $f^i_{T_a^k}(v^1 - 1) > 16(F_{T_a^k} + \frac{2}{k+24NV} + V\gamma_{T_a^k})$ holds. Then, with probability at least*

$$1 - 2\sum_{j=k}^{k+\ell-1} \exp\left(-\frac{1}{2}|\Gamma_a^{j+1}|\delta^2_{T_a^j}\right),$$

the following events happen at time $T_a^{k+\ell}$, where $\ell = k + 24NV$:

- *$A_a^{k+\ell}$;*
- *For both $i \in M^1$, $f^i_{T_a^{k+\ell}}(v^1 - 1) > 4(F_{T_a^{k+\ell}} + \frac{2}{(k+\ell)+24NV} + V\gamma_{T_a^{k+\ell}})$.*

Proof. We prove by an induction from $j = k$ to $k + \ell - 1$. Given A_a^j, A_a^{j+1} happens with probability at least $1 - \exp\left(-\frac{1}{2}|\Gamma_a^{j+1}|\delta^2_{T_a^j}\right)$ according to Lemma 6.6. Hence, with probability at least $1 - \sum_{j=k}^{k+\ell-1} \exp\left(-\frac{1}{2}|\Gamma_a^{j+1}|\delta^2_{T_a^j}\right)$, all events $A_a^k, A_a^{k+1}, \ldots, A_a^{k+\ell}$ happen.

Now we consider the second event. For all $t \in \Gamma_a^{j+1}$, noticing that $\frac{T_a^k}{t-1} \geq \frac{T_a^k}{T_a^{j+1}} \geq \frac{T_a^k}{T_a^{k+\ell}} = \frac{(k+24NV)^2}{(2(k+24NV))^2} = \frac{1}{4}$, we have

$$\begin{aligned} f^i_{t-1}(v^1 - 1) &\geq \frac{T_a^k}{t-1} f^i_{T_a^k}(v^1 - 1) \geq \frac{1}{4} f^i_{T_a^k}(v^1 - 1) \\ \text{(by condition)} &> 4(F_{T_a^k} + \frac{2}{k+24NV} + V\gamma_{T_a^k}) \qquad (6.7) \\ (F_{T_a^k} \text{ and } \gamma_{T_a^k} \text{ are decreasing in } k) &\geq 4(F_{T_a^j} + \frac{2}{j+24NV} + V\gamma_{T_a^j}). \end{aligned}$$

According to Claim 6.6, given A_a^j we have $\frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \leq F_{T_a^j} + \frac{2}{j+24NV} \leq \frac{1}{2NV} - 2\gamma_t$. Using Claim 6.7 with $X = F_{T_a^j} + \frac{2}{j+24NV}$, we have, for bidder $i' \neq i, i' \in M^1$, $\Pr[b_t^{i'} = v^1 - 2 \mid H_{t-1}] \leq \gamma_t$. By Lemma 6.1, $\Pr[b_t^{i'} \leq v^1 - 3 \mid$

$H_{t-1}] \leq (V-1)\gamma_t$. Combining the two, we get $\Pr[b_t^{i'} = v^1 - 1 \mid H_{t-1}] \geq 1 - V\gamma_t$. Let $\Delta = |\Gamma_a^{k+1}|\delta_{T_a^k}$. Similar to the proof of Lemma 6.6, we can use Azuma's inequality to argue that, with probability at least $1 - \exp(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2)$, it holds that

$$\sum_{t \in \Gamma_a^{j+1}} \mathbb{I}[b_t^{i'} = v^1 - 1] \geq \sum_{t \in \Gamma_a^{j+1}} (1 - V\gamma_t - \delta_{T_a^j}) \geq |\Gamma_a^{j+1}|(1 - V\gamma_{T_a^j} - \delta_{T_a^j}).$$

An induction shows that, with probability at least $1 - \sum_{j=k}^{k+\ell-1} \exp\left(-\frac{1}{2}|\Gamma_a^{j+1}|\delta_{T_a^j}^2\right)$, $\sum_{t \in \Gamma_a^{j+1}} \mathbb{I}[b_t^{i'} = v^1 - 1] \geq |\Gamma_a^{j+1}|(1 - V\gamma_{T_a^j} - \delta_{T_a^j})$ holds for all $j \in \{k, \ldots, k+\ell-1\}$. Therefore,

$$\begin{aligned}
f_{T_a^{k+\ell}}^{i'}(v^1 - 1) &\geq \frac{1}{T_a^{k+\ell}}\left(0 + \sum_{t \in \Gamma_a^{k+1} \cup \cdots \cup \Gamma_a^{k+\ell}} \mathbb{I}[b_t^{i'} = v^1 - 1]\right) \\
&\geq \frac{1}{T_a^{k+\ell}}\left(|\Gamma_a^{k+1}|(1 - V\gamma_{T_a^k} - \delta_{T_a^k}) + \cdots + |\Gamma_a^{k+\ell}|(1 - V\gamma_{T_a^{k+\ell-1}} - \delta_{T_a^{k+\ell-1}})\right) \\
&\geq \frac{1}{T_a^{k+\ell}}\left((|\Gamma_a^{k+1}| + \cdots + |\Gamma_a^{k+\ell}|) \cdot (1 - V\gamma_{T_a^k} - \delta_{T_a^k})\right) \\
&= \frac{T_a^{k+\ell} - T_a^k}{T_a^{k+\ell}}(1 - V\gamma_{T_a^k} - \delta_{T_a^k}) \\
&= \frac{4(k+24NV)^2 - (k+24NV)^2}{4(k+24NV)^2}(1 - V\gamma_{T_a^k} - \delta_{T_a^k}) \\
&= \frac{3}{4}(1 - V\gamma_{T_a^k} - \delta_{T_a^k}) \\
&\overset{\text{(assuming } T_b \text{ is large enough)}}{>} 4\left(F_{T_a^{k+\ell}} + \frac{2}{(k+\ell) + 24NV} + V\gamma_{T_a^{k+\ell}}\right).
\end{aligned}$$

This proves the claim for $i' \in M^1$. The claim for $i \in M^1$ follows from (6.7) and the fact that $F_{T_a^k}$ and $\gamma_{T_a^k}$ are decreasing in k. □

We denote by $k_0 = k + \ell$ the time period at which $f_{T_a^{k_0}}^i(v^1 - 1) > 4(F_{T_a^{k_0}} + \frac{2}{k_0 + 24NV} + V\gamma_{T_a^{k_0}})$ for both $i \in M^1$. We continuing the analysis for each period $k \geq k_0$. Define sequence $(G_{T_a^k})$:

$$G_{T_a^k} = \frac{T_a^{k_0}}{T_a^k} \cdot 4 \left(F_{T_a^{k_0}} + \frac{2}{k_0 + 24NV} + V\gamma_{T_a^{k_0}} \right) + \sum_{s=k_0}^{k-1} \frac{T_a^{s+1} - T_a^s}{T_a^k} (1 - V\gamma_{T_a^s} - \delta_{T_a^s}), \quad \text{for } k \geq k_0,$$

where we recall that $\delta_t = (\frac{1}{t})^{1/8}$. We note that $f^i_{T_a^{k_0}}(v^1 - 1) > G_{T_a^{k_0}} = 4\big(F_{T_a^{k_0}} + \frac{2}{k_0+24NV} + V\gamma_{T_a^{k_0}}\big)$.

Claim 6.8. *When T_b is sufficiently large,*

- $G_{T_a^k} \geq 4\big(F_{T_a^{k_0}} + \frac{2}{k_0+24NV} + V\gamma_{T_a^{k_0}}\big)$ *for every $k \geq k_0$.*
- $\lim_{k\to\infty} G_{T_a^k} = 1$.

Proof. Since $1 - V\gamma_{T_a^s} - \delta_{T_a^s} \to 1$ as $T_b \to \infty$, for sufficiently large T_b we have $1 - V\gamma_{T_a^s} - \delta_{T_a^s} \geq 4\big(F_{T_a^{k_0}} + \frac{2}{k_0+24NV} + V\gamma_{T_a^{k_0}}\big)$ and hence $G_{T_a^k} \geq 4\big(F_{T_a^{k_0}} + \frac{2}{k_0+24NV} + V\gamma_{T_a^{k_0}}\big)$.

Now we prove $\lim_{k\to\infty} G_{T_a^k} = 1$. Consider the second term in $G_{T_a^k}$, $\sum_{s=k_0}^{k-1} \frac{T_a^{s+1}-T_a^s}{T_a^k}(1 - V\gamma_{T_a^s} - \delta_{T_a^s})$. Since $\sum_{s=\sqrt{k}}^{k-1} \frac{T_a^{s+1}-T_a^s}{T_a^k} = \sum_{s=\sqrt{k}}^{k-1} \frac{2(s+24NV)+1}{(k+24NV)^2} = \frac{(k+\sqrt{k}+48NV)(k-\sqrt{k})}{(k+24NV)^2} \to 1$ and $1 - V\gamma_{T_a^k} - \delta_{T_a^k} \to 1$ as $k \to \infty$, for any $\varepsilon > 0$ we can always find $K \geq k_0$ such that $\sum_{s=\sqrt{k}}^{k-1} \frac{T_a^{s+1}-T_a^s}{T_a^k} \geq 1 - \varepsilon/2$ for every $k \geq K$ and $1 - V\gamma_{T_a^s} - \delta_{T_a^s} \geq 1 - \varepsilon/2$ for every $s \geq \sqrt{k}$. Hence, $G_{T_a^k} \geq \sum_{s=\sqrt{k}}^{k-1} \frac{T_a^{s+1}-T_a^s}{T_a^k}(1 - V\gamma_{T_a^s} - \delta_{T_a^s}) \geq (1-\varepsilon/2)(1-\varepsilon/2) \geq 1 - \varepsilon$. In addition, $G_{T_a^k} \leq 1$ when T_b is sufficiently large. Therefore $\lim_{k\to\infty} G_{T_a^k} = 1$. □

Lemma 6.8. *Fix any k. Suppose A_a^k holds and $f_{T_a^k}(v^1 - 1) > G_{T_a^k}$ holds for both $i \in M^1$. Then, the following four events happen with probability $\geq 1 - 3\exp\left(-\frac{1}{2}|\Gamma_a^{k+1}|\delta^2_{T_a^k}\right)$:*

- A_a^{k+1}*;*
- $f^i_{T_a^{k+1}}(v^1 - 1) > G_{T_a^{k+1}}$ *holds for both $i \in M^1$;*
- $f^i_t(v^1 - 1) > (1 - \frac{2}{k+24NV})G_{T_a^k}$ *holds for both $i \in M^1$, for any $t \in \Gamma_a^{t+1}$.*

- $\boldsymbol{x}_t^i(v^1-1) = \Pr[b_t^i = v^1 - 1 \mid H_{t-1}] \geq 1 - V\gamma_t$ *for both* $i \in M^1$*, for any* $t \in \Gamma_a^{k+1}$.

Proof. By Lemma 6.6, A_a^{k+1} holds with probability at least $1-\exp\left(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2\right)$. Now we consider the second event. For every $t \in \Gamma_a^{k+1}$, we have

$$\begin{aligned} f_{t-1}^i(v^i-1) &\geq \frac{T_a^k}{T_a^{k+1}} f_{T_a^k}(v^i-1) \\ \text{(by condition)} &> \frac{T_a^k}{T_a^{k+1}} G_{T_a^k} \\ \text{(by Claim 6.5)} &\geq \left(1 - \frac{2}{k+24NV}\right) G_{T_a^k} \\ &\geq \frac{1}{2} G_{T_a^k} \\ \text{(by Claim 6.8)} &\geq 2\left(F_{T_a^k} + \frac{2}{k+24NV} + V\gamma_{T_a^k}\right). \end{aligned} \tag{6.8}$$

In addition, according to Claim 6.6 A_a^k implies $\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b_s^i \leq v^1-3] \leq F_{T_a^k} + \frac{2}{k+24NV} \leq \frac{1}{2NV} - 2\gamma_t$. Using Claim 6.7 with $X = F_{T_a^k} + \frac{2}{k+24NV}$, we get $\Pr[b_t^i = v^1 - 2 \mid H_{t-1}] \leq \gamma_t$. Additionally, by Lemma 6.1 we have $\Pr[b_t^i \leq v^1 - 3 \mid H_{t-1}] \leq (V-1)\gamma_t$. Therefore,

$$\Pr[b_t^i = v^1 - 1 \mid H_{t-1}] \geq 1 - V\gamma_t. \tag{6.9}$$

Using Azuma's inequality with $\Delta = |\Gamma_a^{k+1}|\delta_{T_a^k}$, we have with probability at least $1 - \exp(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2)$, $v\sum_{t\in\Gamma_a^{k+1}}\mathbb{I}[b_t^i = v^1-1] > \sum_{t\in\Gamma_a^{k+1}}(1-V\gamma_t-\delta_{T_a^k}) \geq |\Gamma_a^{k+1}|(1-V\gamma_{T_a^k}-\delta_{T_a^k})$. It follows that $f_{T_a^{k+1}}^i(v^1-1) > \frac{1}{T_a^{k+1}}\left(T_a^k G_{T_a^k} + |\Gamma_a^{k+1}|(1-V\gamma_{T_a^k}-\delta_{T_a^k})\right) = G_{T_a^{k+1}}$ by definition.

Using a union bound, the first event A_a^{k+1} and the second event that $f_{T_a^{k+1}}^i(v^1-1) > G_{T_a^{k+1}}$ holds for both $i \in M^1$ happen with probability at least $1 - 3\exp(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2)$. The third event is given by (6.8) and the forth event is given by (6.9). □

We use Lemma 6.8 from k to ∞; from its third and fourth events, combined with Claim 6.8, we get $\lim_{t\to\infty} f_t^i(v^1-1) \geq \lim_{k\to\infty}\left(1-\frac{2}{k+24NV}\right)G_{T_a^k} = 1$ and $\lim_{t\to\infty}\boldsymbol{x}_t^i =$

$\mathbf{1}_{v^1-1}$, which happens with probability at least $1 - 3\sum_{k=0}^{\infty} \exp(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2)$. This concludes the analysis for Case 2.

Combining Case 1 and Case 2, we have that either $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1, b_s^i = v^1 - 2] = 1$ happens or $\lim_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\forall i \in M^1, b_s^i = v^1 - 1] = 1$ happens (in which case we also have $\lim_{t\to\infty} \boldsymbol{x}_t^i = \mathbf{1}_{v^1-1}$) with overall probability at least $1 - \exp\big(-\frac{T_b}{24NV}\big) - 2\exp\big(-\frac{T_b}{1152N^2V^2}\big) - 3\sum_{k=0}^{\infty} \exp(-\frac{1}{2}|\Gamma_a^{k+1}|\delta_{T_a^k}^2)$. Using Claim 6.4 concludes the proof.

The special case of $v^3 = v^1 - 1$

Claim 6.9. *Given $f_t^i(v^1-2) \geq 1 - \frac{1}{4+2NV}$ for all $i \in M^1$, we have $\Pr[b_t^3 = v^1 - 2 \mid H_{t-1}] \geq 1 - V\gamma_t$.*

Proof. If $f_t^i(v^1 - 2) \geq 1 - \varepsilon$, $\varepsilon = \frac{1}{4+2NV}$, for all $i \in M^1$ then the frequency of the maximum bid to be $v^1 - 2$ is at least $1 - 2\varepsilon$, which implies $\alpha_{t-1}^3(v^1 - 2) \geq 2\frac{1}{N}(1 - 2\varepsilon)$. For any $b \leq v^1 - 3$, $\alpha_{t-1}^3(b) \leq V2\varepsilon$. Since $\gamma_t < \frac{1}{12N^2V^2} < \frac{1}{NV}$, we have $\alpha_{t-1}^3(v^1 - 2) - \alpha_{t-1}^3(b) \geq 2\frac{1}{N}(1 - 2\varepsilon) - 2V\varepsilon > V\gamma_t$, which implies, according to mean-based property, $\Pr[b_t^3 = v^1 - 2] \geq 1 - V\gamma_t$. □

Claim 6.10. *If history H_{t-1} satisfies $f_{t-1}^i(v^1-2) \geq \frac{9}{10}$ for $i \in M^1$ and $f_{t-1}^3(v^1-2) \geq \frac{9}{10}$, then $\Pr[b_{t-1}^{i'} = v^1 - 2 \mid H_{t-1}] \leq \gamma_t$.*

Proof. If $f_{t-1}^i(v^1-2) \geq \frac{9}{10}$ for $i \in M^1$ and $f_{t-1}^3(v^1-2) \geq \frac{9}{10}$, then we have $\frac{1}{t-1}\sum_{s=1}^{t-1} \mathbb{I}[|\{i \notin M^1 : b_s^i = v^1 - 2\}| \geq 2] \geq 1 - 2 \times \frac{1}{10} = \frac{4}{5}$ and $P_{t-1}^{i'}(0 : v^1 - 3) \leq 1 - f_{t-1}^3(v^1 - 2) \leq \frac{1}{10}$.

Recall that $P_t^i(k) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}[\max_{j\neq i} b_s^j = k]$. By $P_t^i(0 : k)$ we mean $\sum_{\ell=0}^{k} P_t^i(\ell)$. And we can calculate

$$\begin{aligned}
&\alpha_{t-1}^{i'}(v^1-1)-\alpha_{t-1}^{i'}(v^1-2)\\
&\geq P_{t-1}^{i'}(v^1-1)\times(\frac{1}{2}-0)+\frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[|\{i\notin M^1 : b_s^i=v^1-2\}|\geq 2]\times(1-\frac{2}{3})\\
&\quad+P_{t-1}^{i'}(0:v^1-3)\times(1-2)\\
&\geq 0+\frac{1}{3}\times\frac{4}{5}-\frac{1}{10}=\frac{1}{6}\\
&>V\gamma_t,
\end{aligned}$$

which implies $\Pr[b_{t-1}^{i'}=v^1-2 \mid H_{t-1}]\leq\gamma_t$ according to mean-based property. □

We only provide a proof sketch here; the formal proof is complicated but similar to the above proof for Case 2 and hence omitted. We prove by contradiction. Suppose Case 1 happens, that is, at each time step T_a^k the frequency of v^1-1 for both bidders $i\in M^1$, $f_{T_a^k}^i(v^1-1)$, is upper bounded by the threshold $16(F_{T_a^k}+\frac{2}{k+24NV}+V\gamma_{T_a^k})$, which approaches 0 as $k\to\infty$. Assuming $A_a^0,\dots,A_a^k$ happen (which happens with high probability), the frequency of $0:v^1-3$ is also low. Thus, $f_t^i(v^1-2)$ must be close to 1. Then, according to Claim 6.9, bidder 3 will bid v^1-2 with high probability. Using Azuma's inequality, with high probability, the frequency of bidder 3 bidding v^1-2 in all future periods will be approximately 1, which increases $f_t^3(v^1-2)$ to be close to 1 after several periods. Then, according Claim 6.10, bidder $i\in M^1$ will switch to bid v^1-1. After several periods, the frequency $f_{T_a^k}^i(v^1-1)$ will exceed $16(F_{T_a^k}+\frac{2}{k+24NV}+V\gamma_{T_a^k})$ and thus satisfy Case 2. This leads to a contradiction.

6.7.2 Proof of Proposition 6.2

We consider a simple case where there are only two bidders with the same type $v^1=v^2=3$. Let $V=3$. The set of possible bids is $\mathcal{B}^1=\mathcal{B}^2=\{0,1,2\}$. Denote $f_t^i(b)=\frac{1}{t}\sum_{s=1}^t\mathbb{I}[b_s^i=b]$ the frequency of bidder i's bid in the first t rounds.

Claim 6.11. *For $i \in \{1,2\}$, $\alpha_t^i(1) - \alpha_t^i(2) = f_t^{3-i}(0) - \frac{f_t^{3-i}(2)}{2}$ and $\alpha_t^i(1) - \alpha_t^i(0) = f_t^{3-i}(1) + \frac{f_t^{3-i}(0)}{2}$.*

Proof. We can express $\alpha_t^i(b)$ using the frequencies as the following: $\alpha_t^i(0) = \frac{3f_t^{3-i}(0)}{2}$; $\alpha_t^i(1) = f_t^{3-i}(1) + 2f_t^{3-i}(0) = 1 + f_t^{3-i}(0) - f_t^{3-i}(2)$; $\alpha_t^i(2) = \frac{f_t^{3-i}(2)}{2} + 1 - f_t^{3-i}(2)$. Then the claim follows from direct calculation. □

We construct a γ_t-mean-based algorithm $\mathcal{A}$ (Algorithm 6.1) with $\gamma_t = O(\frac{1}{t^{1/4}})$ such that, with constant probability, $\lim_{t\to\infty} f_t^i(1) = 1$ but in infinitely many rounds the mixed strategy $\boldsymbol{x}_t^i = \mathbf{1}_2$. The key idea is that, when $\alpha_t^i(1) - \alpha_t^i(2)$ is positive but lower than $V\gamma_t$ in some round t (which happens infinitely often), we let the algorithm bid 2 with certainty in round $t+1$. This does not violate the mean-based property.

We note that this algorithm has no randomness in the first T_0 rounds. It bids 1 in the first $T_0 - T_0^{2/3}$ rounds and bid 0 in the remaining $T_0^{2/3}$ rounds. Define round $T_k = 32^k T_0$ for $k \geq 0$. Let $\gamma_t = 1$ for $1 \leq t \leq T_0$ and $\gamma_t = T_k^{-1/4} = O(t^{-1/4})$ for $t \in [T_k + 1, T_{k+1}]$ and all $k \geq 0$.

Claim 6.12. *Algorithm 6.1 is a γ_t-mean-based algorithm with $\gamma_t = O(t^{-1/4})$.*

Proof. We only need to verify the mean-based property in round $t \geq T_0 + 1$ since $\gamma_t = 1$ for $t \leq T_0$. The proof follows by the definition and is straightforward: If the condition in Line 5 holds, where $\arg\max_b \alpha_{t-1}(b) = 1$ and $\alpha_{t-1}^i(1) - \alpha_{t-1}^i(2) \leq V\gamma_t$, then the mean-based property does not apply to bids 1 and 2 and the algorithm bids 0 with probability $0 \leq \gamma_t$. Otherwise, according to Line 8, the algorithm bids $b' \notin \arg\max_b \alpha_{t-1}(b)$ with probability at most $T_{k+1}^{-1/3} \leq \gamma_t$. □

For $k \geq 0$, let A_k be the event that for both $i \in \{1,2\}$, it holds that $T_k^{-\frac{1}{3}} \leq f_{T_k}^i(0) \leq 2T_k^{-\frac{1}{3}}$ and $f_{T_k}^i(2) = \frac{k}{T_k}$. Since both bidders submit deterministic bids in the first T_0 rounds, it is easy to check that A_0 holds probability 1.

The following two claims show that if $A_0, A_1, \dots$ all happen, then the dynamics time-average converges to 1 while in the meantime, both of the bidders bid 2 at round $T_k + 1$ for all $k \geq 0$.

Claim 6.13. *For any $k \geq 0$ and $i \in \{1, 2\}$, if A_{k+1} holds, then $f_t^i(1) \geq 1 - 64T_{k+1}^{-\frac{1}{3}} - \frac{32k}{T_{k+1}}$ holds for any $t \in [T_k, T_{k+1}]$. In particular, if A_k holds for all $k \geq 0$, then $\lim_{t\to\infty} f_t^i(1) = 1$ for $i \in \{1, 2\}$.*

Proof. Let A_{k+1} holds. Then $2T_{k+1}^{-\frac{1}{3}} \geq f_{T_{k+1}}^i(0) \geq \frac{tf_t^i(0)}{T_{k+1}} \geq \frac{f_t^i(0)}{32}$, which implies that $f_{T_k}^i(0) \leq 64T_{k+1}^{-\frac{1}{3}}$. Similarly, we have $f_t^i(2) \leq \frac{32k}{T_{k+1}}$. The claim follows by $f_t^i(1) = 1 - f_t^i(0) - f_t^i(2)$. □

Claim 6.14. *If A_k happens, then both of the bidders bid 2 at round $T_k + 1$.*

Proof. According to Claim 6.11, we know that for any $i \in \{1, 2\}$ and any $t > T_0$, $\alpha_{t-1}^i(1) - \alpha_{t-1}^i(0) = f_{t-1}^{3-i}(1) + \frac{f_{t-1}^{3-i}(0)}{2} > 0$. Thus $\arg\max_b\{\alpha_{t-1}^i(b)\} \neq 0$ for any history H_{t-1}. Again by Claim 6.11, we have for any $i \in \{1, 2\}$, $0 < T_k^{-\frac{1}{3}} - \frac{k}{T_k} \leq \alpha_{T_k}^i(1) - \alpha_{T_k}^i(2) = f_{T_k}^{3-i}(0) - \frac{f_{T_k}^{3-i}(2)}{2} \leq f_{T_k}^{3-i}(0) \leq 2T_{k+1}^{-\frac{1}{3}} < 3T_{k+1}^{-\frac{1}{4}} = V\gamma_{T_k+1}$. It follows from Lines 5-6 of Algorithm 6.1 that both bidders bid 2 at round $T_k + 1$. □

We now bound the probability of A_{k+1} given A_k happens. This will be used later to derive a constant lower bound on the probability that A_k happens for all $k \geq 0$.

Lemma 6.9. *For any $k \geq 0$, $\Pr[A_{k+1} \mid A_k] \geq 1 - 4\exp\left(\frac{T_{k+1}^{\frac{1}{3}}}{900}\right)$.*

Proof. Suppose A_k happens. We know from Claim 6.14 that both bidders bid 2 in round T_k+1. The following claim shows the behaviour of the algorithm in rounds $[T_k+2, T_{k+1}]$.

Claim 6.15. *For any $i \in \{1, 2\}$ and any $t \in [T_k + 2, T_{k+1}]$, $\Pr[b_t^i = 1 \mid A_k] = 1 - T_{k+1}^{-\frac{1}{3}}$, and $\Pr[b_t^i = 0 \mid A_k] = T_{k+1}^{-\frac{1}{3}}$.*

Proof. According to the definition of Algorithm 6.1, it suffices to prove that for any $t \in [T_k + 2, T_{k+1}]$ and $i \in \{1, 2\}$, $\arg\max_b\{\alpha^i_{t-1}(b)\} = 1$ holds. We prove it by induction. For the base case, it is easy to verify that $\alpha^i_{T_k+1}(1) - \alpha^i_{T_k+1}(2) = f^{3-i}_{T_k+1}(0) - \frac{f^{3-i}_{T_k+1}(2)}{2} > 0, \forall i \in \{1, 2\}$. Suppose the claim holds for all of the rounds $[T_k + 2, t]$. Then none of the bidders bids 2 in rounds $[T_k + 2, t]$. It follows that for any $i \in \{1, 2\}$, $\alpha^i_t(1) - \alpha^i_t(2) = f^{3-i}_t(0) - \frac{f^{3-i}_t(2)}{2} \geq \frac{f^{3-i}_{T_k}(0)}{32} - \frac{k+1}{2T_k} \geq \frac{1}{32T_k^{\frac{1}{3}}} - \frac{k+1}{T_k} > 0$ (since $T_0 > 64^{\frac{3}{2}}$). Therefore $\arg\max_b\{\alpha^i_{t-1}(b)\} = 1$. This completes the induction step. □

From the above proof we can also conclude that for $i \in \{1, 2\}$, $f^i_{T_{k+1}}(2) = \frac{k+1}{T_{k+1}}$.

Note that the bidding strategies of both bidders at different rounds in $[T_k+2, T_{k+1}]$ are independent. Specifically, we have $\Pr[b^i_s = 0] = T_{k+1}^{-\frac{1}{3}}$ for $i \in \{1, 2\}$ and $s \in [T_k+2, T_{k+1}]$. Therefore, by Chernoff bound, we have for $i \in \{1, 2\}$, the total number of bids 0 between rounds $[T_k + 2, T_{k+1}]$ lies between $[\frac{29}{30}(T_{k+1} - T_k - 1)T_{k+1}^{-\frac{1}{3}}, \frac{31}{30}(T_{k+1} - T_k - 1)T_{k+1}^{-\frac{1}{3}}]$ with probability at least $1 - 2\exp\left(-\frac{T_{k+1}-T_k-1}{450T_{k+1}^{\frac{2}{3}}}\right) \geq 1 - 2\exp\left(-\frac{T_{k+1}^{\frac{1}{3}}}{900}\right)$. Therefore, with probability at least $1 - 4\exp\left(-\frac{T_{k+1}^{\frac{1}{3}}}{900}\right)$, both of the above events happens. It implies that for $i \in \{1, 2\}$, the frequency of bids 0 is at least

$$
\begin{aligned}
f^i_{T_{k+1}}(0) &\geq \frac{1}{T_{k+1}}\left(T_k f^i_{T_k}(0) + \frac{29}{30}\frac{T_{k+1} - T_k - 1}{T_{k+1}^{\frac{1}{3}}}\right) \\
&\geq \frac{1}{T_{k+1}}\left(\frac{T_k}{T_k^{\frac{1}{3}}} + \frac{29}{30}\frac{\frac{30}{32}T_{k+1}}{T_{k+1}^{\frac{1}{3}}}\right) \\
&= \frac{32^{\frac{1}{3}}}{32T_{k+1}^{\frac{1}{3}}} + \frac{29}{32T_{k+1}^{\frac{1}{3}}} \geq \frac{1}{T_{k+1}^{\frac{1}{3}}},
\end{aligned}
$$

and the frequency of bids 0 is at most

$$\begin{aligned} f^i_{T_{k+1}}(0) &\le \frac{1}{T_{k+1}}\left(T_k f^i_{T_k}(0) + \frac{31}{30}\frac{T_{k+1}-T_k-1}{T_{k+1}^{\frac{1}{3}}}\right) \\ &\le \frac{1}{T_{k+1}}\left(\frac{2T_k}{T_k^{\frac{1}{3}}} + \frac{31}{30}\frac{T_{k+1}}{T_{k+1}^{\frac{1}{3}}}\right) \\ &= \frac{2\times 32^{\frac{1}{3}}}{32T_{k+1}^{\frac{1}{3}}} + \frac{31}{30T_{k+1}^{\frac{1}{3}}} \le \frac{2}{T_{k+1}^{\frac{1}{3}}}. \end{aligned}$$

Therefore, A_{k+1} holds. This completes the proof of Lemma 6.9. □

Using a union bound, we have $\Pr[\forall k \ge 0, A_k \text{ holds}] = \Pr[A_0]\prod_{k=0}^{\infty}\Pr[A_{k+1} \mid A_k]$ is at least $1 - 4\sum_{j=1}^{\infty}\exp(-T_j^{\frac{1}{3}}/900)$. Since $T_0 = 10^{12}$, $\exp(-T_0^{1/3}/900) \le \frac{1}{16}$ and $T_j = 32^j T_0$, we can further lower bound the probability by $1 - \frac{1}{4}\sum_{j=1}^{\infty}\exp\left(-32^{j/3}\right) \ge \frac{1}{2}$. Therefore, with probability at least $\frac{1}{2}$, A_k holds for all $k \ge 0$. By Claim 6.13, the dynamics time-average converges to the equilibrium of 1, but by Claim 6.14, both bidders' mixed strategies do not converge in the last-iterate sense. This completes the proof of Proposition 6.2.

6.8 Omitted Proofs in Section 6.4

6.8.1 Proof of Lemma 6.2

Let $\Gamma = \{s \le t-1 | \exists i \in M^1, b^i_s \le v^1 - 3\}$. The premise of the lemma says $\frac{|\Gamma|}{t-1} \le \frac{1}{3NV}$. First, note that

$$\begin{aligned} P^i_{t-1}(0 : v^1 - 3) &= \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\max_{i' \ne i} b^{i'}_s \le v^1 - 3] \\ &\le \frac{1}{t-1}\sum_{s=1}^{t-1}\mathbb{I}[\exists i \in M^1, b^i_s \le v^1 - 3] = \frac{|\Gamma|}{t-1} \le \frac{1}{3NV}. \end{aligned} \tag{6.10}$$

Then, according to (6.2),

$$\begin{aligned}&\alpha^i_{t-1}(v^1-1)-\alpha^i_{t-1}(v^1-2)\\&\quad=Q^i_{t-1}(v^1-1)+P^i_{t-1}(v^1-2)-2Q^i_{t-1}(v^1-2)-P^i_{t-1}(0:v^1-3).\end{aligned}\tag{6.11}$$

Using $Q^i_{t-1}(v^1-1)\geq\frac{1}{N}P^i_{t-1}(v^1-1)$ and $Q^i_{t-1}(v^1-2)\leq\frac{1}{2}P^i_{t-1}(v^1-2)$ from (6.1), we can lower bound (6.11) by $\frac{1}{N}P^i_{t-1}(v^1-1)-P^i_{t-1}(0:v^1-3)$. With (6.10), we get $\alpha^i_{t-1}(v^1-1)-\alpha^i_{t-1}(v^1-2)\geq\frac{1}{N}P^i_{t-1}(v^1-1)-\frac{1}{3NV}$.

If $\frac{1}{N}P^i_{t-1}(v^1-1)-\frac{1}{3NV}>V\gamma_t$, then $\alpha^i_{t-1}(v^1-1)-\alpha^i_{t-1}(v^1-2)>V\gamma_t$. By the mean-based property, $\Pr[b^i_t=v^1-2\mid H_{t-1}]\leq\gamma_t$.

Suppose $\frac{1}{N}P^i_{t-1}(v^1-1)-\frac{1}{3NV}\leq V\gamma_t$, which is equivalent to $P^i_{t-1}(v^1-1)\leq\frac{1}{3V}+NV\gamma_t$. Consider $Q^i_{t-1}(v^1-2)$. By the definition of Γ, in all rounds $s\notin\Gamma$ and $s\leq t-1$, we have that all bidders in M^1 bid v^1-2 or v^1-1. If bidder i wins with bid v^1-2 in round $s\notin\Gamma$, she must be tied with at least two other bidders in M^1 since $|M^1|\geq3$; if bidder i wins with bid v^1-2 (tied with at least one other bidder) in round $s\in\Gamma$, that round contributes at most $\frac{1}{2}$ to the summation in $Q^i_{t-1}(v^1-2)$. Therefore,

$$Q^i_{t-1}(v^1-2)\leq\frac{1}{t-1}\left(\frac{(t-1)-|\Gamma|}{3}+\frac{|\Gamma|}{2}\right)=\frac{1}{3}+\frac{1}{6}\frac{|\Gamma|}{t-1}\leq\frac{1}{3}+\frac{1}{18NV}.\tag{6.12}$$

We then consider $P^i_{t-1}(v^1-2)$. Since $P^i_{t-1}(0:v^1-3)+P^i_{t-1}(v^1-2)+P^i_{t-1}(v^1-1)=1$, and recalling that $P^i_{t-1}(0:v^1-3)\leq\frac{1}{3NV}$ and $P^i_{t-1}(v^1-1)\leq\frac{1}{3V}+NV\gamma_t$, we get

$$P^i_{t-1}(v^1-2)=1-P^i_{t-1}(0:v^1-3)-P^i_{t-1}(v^1-1)\geq1-\frac{1}{3NV}-\frac{1}{3V}-NV\gamma_t.\tag{6.13}$$

Combining (6.11) with (6.10), (6.12), and (6.13), we get

$$\begin{aligned}
&\alpha_{t-1}^{i}(v^1-1)-\alpha_{t-1}^{i}(v^1-2)\\
&\geq 0+\left(1-\frac{1}{3NV}-\frac{1}{3V}-NV\gamma_t\right)-2\left(\frac{1}{3}+\frac{1}{18NV}\right)-\frac{1}{3NV}\\
&=\frac{1}{3}-\frac{3N+7}{9NV}-NV\gamma_t\\
&\geq\frac{1}{3}-\frac{3N+7}{9NV}-\frac{1}{12}\quad(\text{because } \gamma_t\leq\frac{1}{12NV})\\
&\geq\frac{1}{4}-\frac{3N+7}{27N}\quad(\text{because } V\geq 3)\\
&\geq\frac{1}{12N}\quad(\text{because } N\geq 3)\\
&\geq V\gamma_t\quad(\text{because } \gamma_t\leq\frac{1}{12NV}).
\end{aligned}$$

Therefore, by the mean-based property, $\Pr[b_t^i=v^1-2\mid H_{t-1}]\leq\gamma_t$.

6.8.2 Proof of Claim 6.2

Since $\delta_{T_a^0}\to 0$ and $\gamma_{T_a^0}\to 0$ as $T_b\to\infty$, when T_b is sufficiently large we have

$$F_{T_a^1}=\frac{1}{c}\frac{1}{4NV}+\frac{c-1}{c}\left(\delta_{T_a^0}+|M^1|V\gamma_{T_a^0}\right)\leq\frac{1}{c}\frac{1}{4NV}+\frac{c-1}{c}\frac{1}{4NV}\leq\frac{1}{4NV}=F_{T_a^0}.$$

By definition, for every $k\geq 1$

$$F_{T_a^{k+1}}=\frac{1}{c}F_{T_a^k}+\frac{c-1}{c}\left(\delta_{T_a^k}+|M^1|V\gamma_{T_a^k}\right),\quad F_{T_a^k}=\frac{1}{c}F_{T_a^{k-1}}+\frac{c-1}{c}\left(\delta_{T_a^{k-1}}+|M^1|V\gamma_{T_a^{k-1}}\right).$$

Using the fact that $F_{T_a^k}\leq F_{T_a^{k-1}}$ and that $\delta_{T_a^k}+|M^1|V\gamma_{T_a^k}$ is decreasing in k, we have $F_{T_a^{k+1}}\leq F_{T_a^k}\leq\frac{1}{4NV}$. Similarly, we have $\widetilde{F}_{T_a^{k+1}}\leq\widetilde{F}_{T_a^k}$ for any $k\geq 0$.

Note that $\delta_{T_a^k}\to 0$ and $\gamma_{T_a^0}\to 0$ as $k\to+\infty$. Therefore, for any $0<\varepsilon\leq\frac{1}{4NV}$, we can find k sufficiently large such that $\frac{1}{c^{k/2}} \leq \frac{\varepsilon}{6}$, $\delta_{T_a^s} \leq \frac{\varepsilon}{6}$, and $\gamma_{T_a^s} \leq \frac{\varepsilon}{6|M^1|V}$. Then we have

$$\begin{aligned}
F_{T_a^k} \leq \widetilde{F}_{T_a^k} &= \frac{1}{c^k} + \sum_{s=0}^{k-1} \frac{c-1}{c^{k-s}} \delta_{T_a^s} + \sum_{s=0}^{k-1} |M^1| V \frac{c-1}{c^{k-s}} \gamma_{T_a^s} \\
&\leq \frac{\varepsilon}{3} + 2 \sum_{s=0}^{k/2-1} \frac{c-1}{c^{k-s}} + \sum_{s=k/2}^{k-1} \frac{c-1}{c^{k-s}} (\delta_{T_a^{k/2}} + |M^1| V \frac{c-1}{c^{k-s}} \gamma_{T_a^{k/2}}) \\
&\leq \frac{\varepsilon}{3} + 2 \frac{1}{c^{k/2}} + \frac{\varepsilon}{3} \sum_{s=k/2}^{k-1} \frac{c-1}{c^{k-s}} \\
&\leq \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon.
\end{aligned}$$

Thus for any $l \geq k$, we have $F_{T_a^l} \leq \widetilde{F}_{T_a^l} \leq \varepsilon$. Since $F_{T_a^k}$ and $\widetilde{F}_{T_a^k}$ are both positive, we have $\lim_{k \to \infty} F_{T_a^k} = \lim_{k \to \infty} \widetilde{F}_{T_a^k} = 0$. □

6.8.3 Proof of Lemma 6.5

We will use an induction to prove the following:

$$\Pr[A_a^{k+1}] \geq 1 - \exp\left(-\frac{T_b}{24NV}\right) - 2\exp\left(-\frac{T_b}{1152N^2V^2}\right) - \sum_{s=0}^{k} \exp\left(-\frac{1}{2}|\Gamma_a^{s+1}|\delta_{T_a^s}^2\right).$$

We do not assume $|M^1| \geq 3$ for now. The base case follows from Corollary 6.4 because A_a^0 is the same as A_{v^1-3}. Suppose A_a^k happens. Consider A_a^{k+1}. For any round $t \in \Gamma_a^{k+1}$,

$$\begin{aligned}
P_{t-1}^i(0 : v^1 - 3) &\leq \frac{1}{t-1} \sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \\
&= \frac{1}{t-1}\Bigg(\sum_{s=1}^{T_a^k} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] + \sum_{s=T_a^k+1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \Bigg) \\
\text{(by } F_{T_a^k} \leq \frac{1}{4NV}\text{)} \ &\leq \frac{1}{t-1} \left(\frac{T_a^k}{4NV} + (t - 1 - T_a^k) \right) \\
\text{(by } T_a^k \leq t-1 \leq T_a^{k+1}\text{)} \ &\leq \frac{1}{T_a^k} \left(\frac{T_a^k}{4NV} + T_a^{k+1} - T_a^k \right) \\
\text{(by } T_a^{k+1} = cT_a^k\text{)} \ &= \frac{1}{3NV}.
\end{aligned}$$

By Lemma 6.1, for any history H_{t-1} that satisfies A_a^k, we have

$$\Pr[\exists i \in M^1, b_t^i \leq v^1 - 3 \ \mid \ H_{t-1}, A_a^k] \leq |M^1| V \gamma_t. \tag{6.14}$$

Let $Z_t = \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] - |M^1| V \gamma_t$ and let $X_t = \sum_{s=T_a^k+1}^{t} Z_s$. We have $\mathbb{E}[Z_t \mid A_a^k, H_{t-1}] \leq 0$. Therefore, the sequence $X_{T_a^k+1}, X_{T_a^k+2}, \ldots, X_{T_a^{k+1}}$ is a supermartingale (with respect to the sequence of history $H_{T_a^k}, H_{T_a^k+1}, \ldots, H_{T_a^{k+1}-1}$). By Azuma's inequality, for any $\Delta > 0$, we have

$$\Pr\left[\sum_{t \in \Gamma_a^{k+1}} Z_t \geq \Delta \ \middle| \ A_a^k \right] \leq \exp\left(-\frac{\Delta^2}{2|\Gamma_a^{k+1}|} \right).$$

Let $\Delta = |\Gamma_a^{k+1}| \delta_{T_a^k}$. Then with probability at least $1 - \exp\big(-\frac{1}{2} |\Gamma_a^{k+1}| \delta_{T_a^k}^2 \big)$, we have

$$\begin{aligned}
\sum_{t \in \Gamma_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \ &< \ \Delta + |M^1| V \sum_{t \in \Gamma_a^{k+1}} \gamma_t \\
&\leq \ |\Gamma_a^{k+1}| \delta_{T_a^k} + |M^1| V |\Gamma_a^{k+1}| \gamma_{T_a^k},
\end{aligned} \tag{6.15}$$

which implies

$$
\begin{aligned}
&\frac{1}{T_a^{k+1}} \sum_{t=1}^{T_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \\
&= \frac{1}{T_a^{k+1}} \left(\sum_{t=1}^{T_a^k} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] + \sum_{t \in \Gamma_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 3] \right) \\
&\leq \frac{1}{T_a^{k+1}} \left(T_a^k F_{T_a^k} + |\Gamma_a^{k+1}| \delta_{T_a^k} + |M^1| V |\Gamma_a^{k+1}| \gamma_{T_a^k} \right) \\
\text{(since } T_a^{k+1} = cT_a^k\text{)} \quad &= \frac{1}{c} F_{T_a^k} + \frac{c-1}{c} \delta_{T_a^k} + |M^1| V \frac{c-1}{c} \gamma_{T_a^k} \\
\text{(by definition)} \quad &= F_{T_a^{k+1}}
\end{aligned}
$$

and thus A_a^{k+1} holds.

Now we suppose $|M^1| \geq 3$. We can change (6.14) to $\Pr[\exists i \in M^1, b_t^i \leq v^1 - 2 \mid H_{t-1}, A_a^k] \leq |M^1| V \gamma_t$ due to Lemma 6.2 and the fact that $\frac{1}{t-1} \sum_{s=1}^{t-1} \mathbb{I}[\exists i \in M^1, b_s^i \leq v^1 - 3] \leq \frac{1}{3NV}$. The definition of Z_t is changed accordingly, and (6.15) becomes

$$
\sum_{t \in \Gamma_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 2] < |\Gamma_a^{k+1}| \delta_{T_a^k} + |M^1| V |\Gamma_a^{k+1}| \gamma_{T_a^k},
$$

which implies

$$
\frac{1}{T_a^{k+1}} \sum_{t=1}^{T_a^{k+1}} \mathbb{I}[\exists i \in M^1, b_t^i \leq v^1 - 2] \leq \frac{1}{T_a^{k+1}} \left(T_a^k \widetilde{F}_k + |\Gamma_a^{k+1}| \delta_{T_a^k} + |M^1| V |\Gamma_a^{k+1}| \gamma_{T_a^k} \right) = \widetilde{F}_{T_a^{k+1}}.
$$

To conclude, by induction,

$$
\begin{aligned}
\Pr[A_a^{k+1}] &= \Pr[A_a^k] \Pr[A_a^{k+1} | A_a^k] \\
&\geq \Pr[A_a^k] - \exp\left(-\frac{1}{2} |\Gamma_a^{k+1}| \delta_{T_a^k}^2 \right) \\
&\geq 1 - \exp\left(-\frac{T_b}{24NV} \right) - 2 \exp\left(-\frac{T_b}{1152 N^2 V^2} \right) - \sum_{s=0}^{k} \exp\left(-\frac{1}{2} |\Gamma_a^{s+1}| \delta_{T_a^s}^2 \right).
\end{aligned}
$$

As $\delta_t = (\frac{1}{t})^{\frac{1}{3}}$ and $|\Gamma_a^s| = c^{s+d(v^1-3)-1}(c-1)T_0$, $T_a^s = c^{s+d(v^1-3)}T_0$ (let $v^1-3=0$ if $v^1<3$), we have

$$
\begin{aligned}
&\sum_{s=0}^{k} \exp\left(-\frac{1}{2}|\Gamma_a^{s+1}|\delta_{T_a^s}^2\right) \\
&= \sum_{s=0}^{k} \exp\left(-\frac{1}{2}c^{\frac{1}{3}(s+d(v^1-3))}(c-1)(T_0)^{\frac{1}{3}}\right) \\
&= \exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}\right)\left(1+\sum_{s=1}^{k}\exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}(c^{\frac{s}{3}}-1)\right)\right) \\
&\le \exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}\right)\left(1+\sum_{s=1}^{k}\exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}s(c^{\frac{1}{3}}-1)\right)\right) \\
&\le \exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}\right)\left(1+\sum_{s=1}^{k}(\frac{1}{2})^s\right) \\
&\le 2\exp\left(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}\right),
\end{aligned}
$$

where in the last but one inequality we suppose that T_0 is large enough so that $\exp\big(-\frac{1}{2}c^{\frac{1}{3}d(v^1-3)}(c-1)(T_0)^{\frac{1}{3}}s(c^{\frac{1}{3}}-1)\big) \le \frac{1}{2}$. Substituting $T_0 = 12NVT_b = \frac{1}{c-1}T_b$, $c = 1+\frac{1}{12NV}$, and $c^d = 8NV$ gives

$$
\begin{aligned}
\sum_{s=0}^{k} \exp\left(-\frac{1}{2}|\Gamma_a^{s+1}|\delta_{T_a^s}^2\right) &\le 2\exp\left(-\left(\frac{(8NV)^{(v^1-3)}T_b}{1152N^2V^2}\right)^{\frac{1}{3}}\right) \\
&\le 2\exp\left(-\left(\frac{T_b}{1152N^2V^2}\right)^{\frac{1}{3}}\right),
\end{aligned}
$$

concluding the proof.

Part III

Incentive Issues in Machine Learning Systems

Chapter 7

Incentives and Polarization in Recommender Systems

joint work with
Kun Jin, Andrew Estornell, Xiaoying Zhang,
Yiling Chen, Yang Liu [LJE+24]

This part of my dissertation focuses on the incentive issues in machine learning systems. As the strategic behaviors of humans or algorithms are ubiquitous, understanding the impact of such behaviors is essential to the design of socially responsible AI systems. The following chapter investigates the incentive issues in one of the most successful commercial applications of machine learning algorithms: recommender systems.

7.1 Introduction

From restaurant selection, video watching, to apartment renting, recommender systems play a pivotal role across a plethora of real-world domains. These systems match users with content they like, and help creators (those producing the content) identify their target audiences. Nevertheless, behind such success, concerns have emerged regarding possible harmful outcomes of recommender systems, in particular, *filter bubbles* [MWY+20, AGS20] and *polarization* [SLL21] – outcomes with insufficient *recommendation diversity* and *creation diversity*. Recommendation diversity, meaning the diversity of the contents recommended to a user, is key to users' engagement and retention on the platform. Meanwhile, creation diversity, meaning the variety of content created on the platform, is a determinant of the platform's long-term health. In extreme cases, insufficient creation diversity can lead to consensus or polarization, where the latter can cause conflict and hatred, diminish people's mutual understanding, and cause societal crises. Therefore, from both business and social responsibility perspectives, championing and improving diversity in recommender systems is equally important as optimizing recommendation relevancy.

There is increasing emphasis in academia and industry on investigating and improving the diversity of recommender systems, combating filter bubbles and polarization. Popular diversity-boosting approaches include applying post-processing procedures such as reranking [CG98, ZMKL05] and setting diversity-aware objectives in addition to relevance maximization [SYCY13, ZH08, Hur13, WRB+18, CWM+17]. These methods aim to increase the recommendation diversity for users. Assuming that the contents on the platform are static, these methods have been shown to bring diversity gain to the system.

However, an important aspect is overlooked in the aforementioned approaches: users and contents on a recommendation platform are not static entities – they can be *influenced* by the recommendation made by the system. In content creation platforms like YouTube, TikTok, and Twitter, recommendations naturally affect both content users and content creators. It is well known that the exposure to recommended items can shift a user's preference [JCL+19, KBKW21, DM22]. On the other hand, the creators have the *incentive* to change their creation styles constantly to attract their audience better (and to make more profits from the platform) [ER23, HKJ+23, JGS24]. While the effects of recommendation on either users or creators have been investigated separately, to our knowledge no previous work considers both effects. The dual influence of recommendation on users and creators causes complicated dynamics where users and creators interact and their preferences evolve together. Such evolution might exacerbate filter bubble and polarization effects. Whether the aforementioned diversity-boosting approaches still work in a dynamic environment with dual influence is questionable.

Our Contributions The first contribution of our work is to define a novel, natural dynamics model to capture the dual influence of a recommender system on adaptive users and strategic creators, which we call user-creator feature dynamics (Section 7.2). We leverage the users' and items'/creators' embedding vectors to represent their preferences and creation styles, and use cosine similarity to characterize the relevance of creations and users' interests (which is common in the recommender system literature and practice). This model allows us to formally reason about the impact of various design choices on the long-term diversity of a recommender system with dual influence.

Our second contribution is to demonstrate that, under realistic conditions, the user-creator feature dynamics of any recommender system with dual influence must unavoidably converge to polarization (Section 7.3), i.e., the preferences of users and the contents of creators will become tightly clustered into two opposite groups, significantly reducing the diversity of the system. We demonstrate that this phenomenon still occurs even after applying diversity-boosting interventions to the system.

Then, (in Section 7.4) we investigate some real-world designs of recommendation algorithms in order to look for techniques that mitigate polarization. Interestingly, we find that some common efficiency-improving methods, such as top-k truncation, can both prevent the system from polarization and improve the creation diversity. We also provide empirical results (Section 7.5) on both synthetic and real-world (MovieLens) data. As predicted by our theory, we find that systems with dual influence more easily converge to polarization under diversity-boosting designs, while efficiency-oriented and relevance-optimizing designs can in fact improve the long-term diversity of the system. This could explain why polarization does not always happen in reality. Section 7.6 concludes and offers additional discussions.

7.1.1 Related Work

Diversity in Recommendations Diversity, filter bubbles, and polarization are important topics in recommender system research. They are closely related but with different focuses. On the one hand, filter bubbles are frequently defined as decreasing recommendation diversity over time [AGS20], which describes both the process and the outcome of insufficiently diverse recommendations. On the other hand, polarization describes the negative outcome of insufficient mutual understanding between people [SLL21]. In content platforms, an example of polarization is people creating content with strong agreement or disagreement with other content under the same topic, e.g., political opinions. To combat these negative outcomes, previous works propose diversity-boosting approaches including re-ranking [CG98, ZMKL05] and diversity-aware objective optimization [SYCY13, ZH08, Hur13, WRB$^+$18, CWM$^+$17, ZWL23, CLGB24]. Despite having positive effects in situations where user preferences and creation styles are fixed, these approaches overlooked the dynamic nature of recommender systems. We show that certain approaches are not effective under the dual influence of recommendation.

Opinion Dynamics Opinion dynamics study the effect of people exchanging opinions with others on social networks [SSL07, GJ10, LS14, AL15]. Our model of a recommender system with dual influence on users and creators resembles a bipartite social network, and our conclusion that the system converges to polarization is conceptually similar to people reaching consensus on social networks [ACFO13, CCLP15, MTG18, ZAAZ22]. However, the technique we use to prove our conclusion (absorbing Markov chain) significantly differs from the main technique (stability of ODE) in the mentioned works.

Performative effects of recommender systems The phenomenon that predictive systems like recommender systems can impact the individuals interacting with those systems (e.g., users and creators) is related to the literature of performative prediction [PZMDH20, HJMD22]. These impacts can be direct, such as individuals ostensibly modifying their features in order to obtain more desirable outcomes [LR22]. Prior works on the performative effects of recommender systems (e.g., [BPT18, JCL$^+$19, DM22, YLN$^+$22, ER23, YLN$^+$23, PMB23, HKJ$^+$23, AB23, YLW$^+$24, AVWZ24, JGS24]) only consider one-sided impact, either on users or on creators. Differing from them, our work studies two-sided impacts, i.e., on both users and creators. We provide Table 7.1 to compare our work with some previous works.

Table 7.1: Comparison between our work and some previous works on performative effects of recommender systems

Works	**Adaptive Users?**	**Adaptive Creators?**	**Creator Reward**	**Dynamics or Equilibrium?**	**Content Adjustment Cost**
Ours	Yes	Yes	User engagement	Dynamics	Implicit
[ER23]	No	Yes	Exposure	Dynamics	Explicit
[YLN$^+$23]	No	Yes	User engagement	Dynamics	No cost
[PMB23]	No	Yes	User engagement	Dynamics	No cost
[JGS24]	No	Yes	Exposure	Equilibrium	Explicit
[HKJ$^+$23]	No	Yes	Exposure	Equilibrium	No cost
[BPT18]	No	Yes	Exposure	Equilibrium	No cost
[AVWZ24]	No	Yes	User engagement	Equilibrium	No cost
[YLW$^+$24]	No	Yes	Designed by a welfare-maximizing platform	Dynamics	No cost
[DM22]	Yes	No[1]	N/A	Dynamics	N/A
[YLN$^+$22]	Yes	No[1]	N/A	Dynamics	N/A
[AB23]	Adversarial	No[1]	N/A	Dynamics	N/A

[1]: These works study the design of recommendation algorithms for the platform with a fixed set of content, without explicitly modeling the content creators.

7.2 Model: User-Creator Feature Dynamics

We define a *dynamics* model for user preferences and content/creator features in a recommender system. Let $\boldsymbol{U}^t = [\boldsymbol{u}_j^t]_{j=1}^m = [\boldsymbol{u}_1^t, \ldots, \boldsymbol{u}_m^t] \in \mathbb{R}^{d\times m}$ be a population of m users and $\boldsymbol{V}^t = [\boldsymbol{v}_i^t]_{i=1}^n = [\boldsymbol{v}_1^t, \ldots, \boldsymbol{v}_n^t] \in \mathbb{R}^{d\times n}$ be a population of n creators at time t, where each vector $\boldsymbol{u}_j^t, \boldsymbol{v}_i^t \in \mathbb{S}^{d-1}$ represent the preference/feature vector of each user and creator respectively, assumed to be on the unit sphere $\mathbb{S}^{d-1}$ with ℓ_2-norm. These are the true features of users and creators, which may or may not equal the features learned by the recommender system. Then $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ denotes the state of the dynamics at time t. The dynamics evolve as follows at each time step t:

1) Recommendation: Each user $j \in [m]$ is recommended a creator, where creator $i \in [n]$ is chosen with a probability

$$p_{ij}^t = p_{ij}^t(\boldsymbol{U}^t, \boldsymbol{V}^t). \tag{7.1}$$

While we allow a wide array of different functions $p_{ij}^t(\cdot)$, a common example of such functions is the so-called *softmax function*:

$$p_{ij}^t = \mathrm{softmax}(\boldsymbol{u}_j^t, \boldsymbol{V}^t; \beta) = \frac{\exp(\beta\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle)}{\sum_{i=1}^n \exp(\beta\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle)}. \tag{7.2}$$

A larger β means that the recommendation is more sensitive to the *relevance* of a creator to a user, measured by $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle$.

2) User update: After recommendation, each user $j \in [m]$ updates their feature vector $\boldsymbol{u}_j^t$, based on which creator, say i_j^t, was recommended to them:

$$\boldsymbol{u}_j^{t+1} = \mathcal{P}\big(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_{i_j^t}^t, \boldsymbol{u}_j^t)\boldsymbol{v}_{i_j^t}^t\big). \tag{7.3}$$

Here, $\eta_u \in [0,1]$ is a parameter controlling the rate of update, $f(\boldsymbol{v}_i, \boldsymbol{u}_j)$ is a function that quantifies the impact of creator i's content on user j (discussed in detail later), and $\mathcal{P}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}$ is the projection back onto the unit sphere. Our user update model generalizes [DM22], which considers $\boldsymbol{u}_j^{t+1} = \mathcal{P}(\boldsymbol{u}_j^t + \eta_u \langle \boldsymbol{v}_{i_j^t}^t, \boldsymbol{u}_j^t \rangle \boldsymbol{v}_{i_j^t}^t)$, by replacing the inner product with a general function f.

3) Creator update: Creators also update their feature vectors based on which users are recommended their content. For each creator $i \in [n]$, let $J_i^t = \{j : i_j^t = i\}$ be the set of users being recommended creator i, then $\boldsymbol{v}_i^t$ is updated by:

$$\boldsymbol{v}_i^{t+1} = \mathcal{P}\Big(\boldsymbol{v}_i^t + \frac{\eta_c}{|J_i^t|} \sum_{j \in J_i^t} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) \boldsymbol{u}_j^t\Big), \tag{7.4}$$

where $\eta_c \in [0,1]$ is a parameter controlling the rate of update, and $g(\boldsymbol{u}_j, \boldsymbol{v}_i)$ is a function that quantifies the impact of user j on creator i.

Impact functions f and g Our results apply to any impact functions f and g that satisfy the following natural assumptions. First, $f(\boldsymbol{v}_i, \boldsymbol{u}_j)$ and the inner product $\langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle$ have the same sign:

$$f(\boldsymbol{v}_i, \boldsymbol{u}_j) \text{ is } \begin{cases} > 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle > 0 \\ < 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle < 0 \\ = 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle = 0. \end{cases}$$

This means that if a user *likes* the content $(\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t \rangle > 0)$, then the user vector $\boldsymbol{u}_j^t$ will be updated *towards* the direction of the creator vector $\boldsymbol{v}_j^t$. If the user *dislikes* the content $(\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t \rangle < 0)$, then the user vector $\boldsymbol{u}_j^t$ will move *away from* $\boldsymbol{v}_j^t$. Such "biased assimilation" user behavior is well documented in the literature [DM22]. Further, we assume upper and lower bounds on $|f|$:

$$|f(\boldsymbol{v}_i, \boldsymbol{u}_j)| \leq 1, \qquad |f(\boldsymbol{v}_i, \boldsymbol{u}_j)| \geq L_f > 0 \text{ whenever } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle \neq 0.$$

"$|f(\boldsymbol{v}_i, \boldsymbol{u}_j)| \geq L_f > 0$" means that the exposure to an item that a user likes or dislikes always has a non-negligible impact on the user's preference. For example, $f(\boldsymbol{v}_i, \boldsymbol{u}_j) = \operatorname{sign}(\langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle) \cdot a + \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle \cdot b$ satisfies both assumptions with $L_f = a > 0$ and $b \geq 0$.

For g, likewise assume that its sign is the same as $\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle$:

$$g(\boldsymbol{u}_j, \boldsymbol{v}_i) \text{ is } \begin{cases} > 0 & \text{if } \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle > 0 \\ < 0 & \text{if } \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle < 0 \\ = 0 & \text{if } \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle = 0. \end{cases}$$

Intuitively, this captures the incentive of a creator who aims to maximize the average ratings from users who are recommended their items. On video platforms for example, if the creators are rewarded based on the average rating of their videos, they will try to reinforce their creation styles based on the users who give positive feedback ($\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle > 0$) so that their creations are more likely to be recommended to those users. Meanwhile, the creators will also change their creation styles based on negative feedback ($\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle < 0$), but in the opposite direction of the negative-feedback users' interests, so that their creations are less likely to be recommended to those users. Taking both scenarios into account, the creator moves towards the weighted average of user preferences $\sum_{j \in J_i^t} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) \boldsymbol{u}_j^t$, which is captured by our update rule (7.4). A particular example of g is the sign function $g(\boldsymbol{u}_j, \boldsymbol{v}_i) = \operatorname{sign}(\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle) \in \{-1, 0, 1\}$. We will only consider the sign function g in order to simplify the theoretical presentation. We believe that all our results can be generalized to other g functions satisfying similar conditions as f; the details are left as future work.

7.3 Theoretical Result: Unavoidable Polarization

Having defined the user-creator feature dynamics in a recommender system with dual influence, we now theoretically study how such dynamics evolve. Our main result is: if every creator can be recommended to every user with some non-zero probability, then the dynamics must eventually *polarize*.

Definition 7.1 (consensus and bi-polarization)**.** *Let $R > 0$. The dynamics $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ is said to reach:*

- *R-*consensus *if there exists a vector $\boldsymbol{c} \in \mathbb{R}^d$ such that every feature vector is R-close to $\boldsymbol{c}$: $\forall \boldsymbol{u}_j^t, \|\boldsymbol{u}_j^t - \boldsymbol{c}\|_2 \le R$ and $\forall \boldsymbol{v}_i^t, \|\boldsymbol{v}_i^t - \boldsymbol{c}\|_2 \le R$.*
- *R-*bi-polarization *if there exists a vector $\boldsymbol{c} \in \mathbb{R}^d$ such that every feature vector is R-close to $+\boldsymbol{c}$ or $-\boldsymbol{c}$: $\forall \boldsymbol{u}_j^t$, $\|\boldsymbol{u}_j^t - \boldsymbol{c}\|_2 \le R$ or $\|\boldsymbol{u}_j^t + \boldsymbol{c}\|_2 \le R$, and $\forall \boldsymbol{v}_i^t$, $\|\boldsymbol{v}_i^t - \boldsymbol{c}\|_2 \le R$ or $\|\boldsymbol{v}_i^t + \boldsymbol{c}\|_2 \le R$.*

The dynamics is said to reach $(R, \boldsymbol{c})$-consensus (or $(R, \boldsymbol{c})$-bi-polarization) if the dynamics reaches R-consensus (or R-bi-polarization) with the vector $\boldsymbol{c}$.

Consensus is any state where all users and creators have similar feature vectors (with maximum difference R), implying that they have similar interests or preferences. Bi-polarization is any state where all users and creators are clustered into two groups with exactly opposite features (e.g., Republicans vs Democrats). Mathematically, consensus is a special case of bi-polarization.

Proposition 7.1. *Bi-polarization states are **absorbing**: once the dynamics reaches $(R, \boldsymbol{c})$-bi-polarization with some $R \in [0, 1]$ and $\boldsymbol{c} \in \mathbb{S}^{d-1}$, it will satisfy $(R, \boldsymbol{c})$-bi-polarization forever. The same holds for consensus.*

Proof. See Section 7.9. □

A natural property of a recommender system is that every creator can be recommended to every user with some non-zero probability: $p_{ij}^t \geq p_0 > 0$ with some constant p_0. This is satisfied by the softmax function, which is a rough model of real-world recommendation algorithms [CAS16, KBKW21]: $p_{ij}^t = \frac{\exp(\beta\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle)}{\sum_{i=1}^n \exp(\beta\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle)} \geq \frac{\exp(-\beta)}{n\exp(\beta)} = p_0 > 0$. Moreover, many large-scale real-world recomendation systems (e.g., Yahoo! [LCLS10] and Kuaishou [GLZ+22]) intentionally insert small random traffic attempting to improve recommendation diversity or explore users' interests [JMvE18, YCX+18], which will cause all recommendation probabilities to be non-zero. We show in Theorem 7.1 that, however, a recommender system satisfying $p_{ij}^t \geq p_0 > 0$ must converge to polarization, under some additional conditions on the users' and creators' update rates:

Theorem 7.1. *Suppose* $g(\boldsymbol{u}_j, \boldsymbol{v}_i) = \operatorname{sign}(\langle \boldsymbol{u}_j, \boldsymbol{v}_i\rangle)$*, the update rates* $\eta_c \leq \frac{\eta_u L_f}{2}$ *and* $\eta_u < \frac{1}{2}$*, and the recommendation probability* $p_{ij}^t \geq p_0 > 0, \forall i, j, t$*. Then, from almost all initial states, the dynamics* $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ *will eventually reach R-consensus or R-bi-polarization for any* $R > 0$*.*

In other words, if the users' and creators' updates are not too fast and all recommendation probabilities are non-zero, then all users and creators will eventually converge to at most two clusters (regardless of the feature dimension d). Since creators in one cluster produce similar contents, users in such a polarized system can never receive diverse recommendations. This means that the naïve attempt of imposing $p_{ij}^t \geq p_0 > 0$ cannot improve the diversity of a recommender system with dual influence. The conditions on the update rates η_u, η_c are only assumed to simplify the proof of Theorem 7.1. Our experiments (in Section 7.5) will show that polarization still occurs even without those conditions.

Theorem 7.1 does not characterize the rate of convergence of the user-creator feature dynamics to polarization, which we leave as an open question.

The proof of Theorem 7.1 is an absorbing Markov chain argument. It uses the following lemma:

Lemma 7.1. *Suppose $\eta_c \leq \frac{\eta_u L_f}{2}$ and $\eta_u < \frac{1}{2}$. For any $R > 0$, for almost every state $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ in the state space, there exists a path $(\boldsymbol{U}^t, \boldsymbol{V}^t) \rightarrow (\boldsymbol{U}^{t+1}, \boldsymbol{V}^{t+1}) \rightarrow \cdots \rightarrow (\boldsymbol{U}^{t+T}, \boldsymbol{V}^{t+T})$ of finite length that leads to an R-bi-polarization state $(\boldsymbol{U}^{t+T}, \boldsymbol{V}^{t+T})$.*

The proof of this lemma (in Appendix 7.10) is involved. It uses induction on the number of creators n. The base case of $n = 1$ is proved by a potential function argument. For $n \geq 2$, we first construct a path that leads the *subsystem* of $n-1$ creators and all users to R-bi-polarization. Then, depending on where the remaining creator is, we construct a sequence of recommendations that leads the remaining creator to one of the two clusters formed by the $n-1$ creators and all users. Such recommendations will move some users out of the formed clusters, which requires extra care in the proof.

Proof of Theorem 7.1. For any state $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ in the state space, by Lemma 7.1 there exists a path $(\boldsymbol{U}^t, \boldsymbol{V}^t) \rightarrow \cdots \rightarrow (\boldsymbol{U}^{t+T}, \boldsymbol{V}^{t+T})$ of length T that leads to R-bi-polarization. Because every creator can be recommended to a user with probability at least p_0, each transition $(\boldsymbol{U}^{t'}, \boldsymbol{V}^{t'}) \rightarrow (\boldsymbol{U}^{t'+1}, \boldsymbol{V}^{t'+1})$ happens with probability at least p_0^m. So, the path of length T has probability at least $p_0^{mT} > 0$, and the probability that the dynamics *does not* reach R-bi-polarization after KT steps is at most $(1 - p_0^{mT})^K$, which $\rightarrow 0$ as $K \rightarrow \infty$. Therefore, with probability 1 the dynamics will reach R-bi-polarization eventually. □

7.4 Discussions on Real-World Designs

Next, we discuss how 4 types of real-world recommender system designs affect the user-creator feature dynamics: top-k truncation, threshold truncation, diversity-boosting, and uniform traffic.

7.4.1 Top-k Truncation

A prevalent practice in modern two-stage recommendation algorithms on large-scale platforms, such as YouTube [CAS16], is to first filter out items that are unlikely to be relevant to a user, then make recommendations from the remaining items. In particular, we consider the top-k truncation policy: for every user j, find the k most relevant creators, namely, the k creators whose inner products with the user $\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t \rangle$ are largest (equivalently, the k creators whose probabilities p_{ij}^t of being recommended to user j are highest), then recommend one of those k creators to user j with probability proportional to p_{ij}^t. The other creators will not be recommended. This practice significantly reduces the computation cost and improves the relevancy of recommendations. Interestingly, we show that such a practice also has the potential to improve the long-term diversity of a recommender system with dual influence.

Definition 7.2 (clusters)**.** *We say a state* $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ forms q clusters *if there exist* $\boldsymbol{c}_1, \dots, \boldsymbol{c}_q \in \mathbb{R}^d$ *and a small number* $R > 0$ *such that every feature vector is in the* ℓ_2 *ball of some* $\boldsymbol{c}_i$ *with radius* R *(denoted by* $B(\boldsymbol{c}_\ell, R) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{c}_\ell\|_2 \le R\}$*), and* $B(\boldsymbol{c}_\ell, 2R) \cap B(\boldsymbol{c}_{\ell'}, 2R) = \emptyset$ *for* $\ell \neq \ell'$*.*

It is clear that consensus has a single cluster, and bi-polarization has two.

Proposition 7.2. *With top-k truncation, there exist states* $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ *that form* $\lfloor n/k \rfloor$ *clusters and are absorbing (i.e., once the system forms* $\lfloor n/k \rfloor$ *clusters, it forms* $\lfloor n/k \rfloor$ *clusters forever).*

Proof. Let $R > 0$ be any small number. Let $\boldsymbol{c}_1, \dots, \boldsymbol{c}_{\lfloor n/k \rfloor} \in \mathbb{R}^d$ be $\lfloor n/k \rfloor$ vectors that satisfy $B(\boldsymbol{c}_\ell, 2R) \cap B(\boldsymbol{c}_{\ell'}, 2R) = \emptyset$ for $\ell \neq \ell'$, where $B(\boldsymbol{c}, R)$ is the ball centered at $\boldsymbol{c}$ with radius R: $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{c}\|_2 \le R\}$. Consider user and creator features $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ that satisfy: every ball $B(\boldsymbol{c}_\ell, R)$ ($\ell = 1, \dots, \lfloor n/k \rfloor$) contains k creator vectors, and every user vector $\boldsymbol{u}_j^t$ is in one of the balls $B(\boldsymbol{c}_\ell, R)$. By definition, $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ form $\lfloor n/k \rfloor$ clusters. We show that, after one step of update, the new state $(\boldsymbol{U}^{t+1}, \boldsymbol{V}^{t+1})$ must still form $\lfloor n/k \rfloor$ clusters. Consider any user j. Suppose $\boldsymbol{u}_j^t \in B(\boldsymbol{c}_\ell, R)$, then the distance from $\boldsymbol{u}_j^t$ to any creator $\boldsymbol{v}_i^t \in B(\boldsymbol{c}_\ell, R)$ is at most $2R$:

$$\|\boldsymbol{u}_j^t - \boldsymbol{v}_i^t\| \leq 2R.$$

The distance from $\boldsymbol{u}_j^t$ to any creator $\boldsymbol{v}_{i'}^t$ not in $B(\boldsymbol{c}_\ell, R)$ is greater than $2R$:

$$\|\boldsymbol{u}_j^t - \boldsymbol{v}_{i'}^t\| > 2R$$

because $\boldsymbol{v}_{i'}^t$ is in some other ball $B(\boldsymbol{c}_{\ell'}, R)$ that satisfies $B(\boldsymbol{c}_{\ell'}, 2R) \cap B(\boldsymbol{c}_\ell, 2R) = \emptyset$. This implies that the inner products between user j and the creators in ball $B(\boldsymbol{c}_\ell, R)$ are greater than that with the creators in other ball:

$$\begin{aligned}
\forall \boldsymbol{v}_i^t \in B(\boldsymbol{c}_\ell, R), \quad \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle &= 1 - \frac{1}{2}\|\boldsymbol{u}_j^t - \boldsymbol{v}_i^t\|_2^2 \\
&\geq 1 - \frac{1}{2}(2R)^2 \\
&> 1 - \frac{1}{2}\|\boldsymbol{u}_j^t - \boldsymbol{v}_{i'}^t\| = \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle, \quad \forall \boldsymbol{v}_{i'}^t \in B(\boldsymbol{c}_{\ell'}, R).
\end{aligned}$$

Since $B(\boldsymbol{c}_\ell, R)$ contains k creators, these k creators are the k-most relevant ones to user j, so user j will only be recommended these creators. Then, by applying Proposition 7.1 to each of the $\lfloor n/k \rfloor$ balls separately, we see that each ball is a R-consensus and hence absorbing. So, the new state $(\boldsymbol{U}^{t+1}, \boldsymbol{V}^{t+1})$ still forms $\lfloor n/k \rfloor$ clusters with these $\lfloor n/k \rfloor$ balls. □

This result is in contrast with Theorem 7.1 which shows that a recommender system where every creator can be recommended to every user ($p_{ij}^t > 0$) is doomed to polarize. With top-k truncation where some $p_{ij}^t = 0$, polarization can be avoided. Experiments in

Section 7.5.5 support our prediction that top-k truncation can reduce polarization and improve diversity.

7.4.2 Threshold Truncation

Besides top-k truncation, threshold truncation is another way to filter out irrelevant creators: set a threshold $\tau \in [-1, 1]$ such that any user-creator pair with inner product $\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle < \tau$ is not recommended. A natural choice is $\tau = 0$, meaning that users will not receive recommendations predicted to be "disliked" by them. Increasing τ is similar to increasing the β in the softmax function, which improves recommendation relevance.

Proposition 7.3. *In d-dimensional feature space, if user-creator pairs with $\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle < 0$ are not recommended, then there exist stable states with $d + 1$ clusters.*

Proof. The d-dimensional simplex centered at the original has $d+1$ vectors with negative inner products with each other. They form $d + 1$ clusters. Since user-creator pairs with negative inner product $\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle < 0$ are not recommended, recommendations only happen within each cluster. By Proposition 7.1, each cluster is absorbing, so the whole system is stable, keep forming $d + 1$ clusters forever. □

Although truncation at $\tau = 0$ allows stable states with $d + 1$ clusters to exist, the dynamics does not necessarily converge to such states; it can still end up with stable states with fewer clusters. In fact, experiments (in Section 7.5.6) show that truncation at $\tau = 0$ is *not good* for diversity and causes severe polarization, while truncation at a large threshold like $\tau = 0.707$ is better at reducing polarization.

7.4.3 Diversity Boosting

Diversity boosting aims to explore users' interests and improve users' experience by diversifying recommendation. For example, when making recommendations, the model optimizes the objective:

$$h_{rel}(\langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle) + \rho h_{div}(list_i, \boldsymbol{v}_j), \tag{7.5}$$

where h_{rel}, h_{div} rewards the recommendation relevance and diversity respectively and $list_i$ records the recent list of recommended items to user i. h_{div} can take a simple form of $\sum_{j' \in list_i} 1 - \langle \boldsymbol{v}_{j'}, \boldsymbol{v}_j \rangle$, and $\rho > 0$ controls the strength of diversity-boosting. Despite being successful when users' preferences and items are fixed, this design alone cannot prevent bi-polarization in our dual-influence dynamics, since the conditions in Theorem 7.1 are still satisfied and the users' and creators' update rules remain the same. Experiments in Section 7.7 support our claim.

7.4.4 Uniform Traffic

Adding a small fraction of uniform traffic to the personalized recommendations is another method proposed in previous works to improve recommendation diversity or to explore user preferences [JMvE18, GLZ+22, BCIV23, BV18, LCL+23]. This method gives a non-zero lower bound on the probability of every creator being recommended to every user. So, as a corollary of our Theorem 7.1, it causes a recommender system with dual influence to polarize. Such an observation is striking as it demonstrates that optimizing for recommendation diversity in a static setting can ultimately lead to a huge loss of the system diversity in the long run.

7.5 Experiments: Synthetic Data

We perform experiments to investigate the behavior of user-creator feature dynamics and the effect of top-k truncation and threshold truncation on the dynamics. This section presents the experiment results on synthetic data. Section 7.7 presents the results on a real-world dataset (MovieLens 20M).

7.5.1 Experiment Setup

The dynamics is initialized by randomly generating user and creator features on the unit sphere in $\mathbb{R}^d$. We pick $d = 10$, number of creators $n = 50$, number of users $m = 100$. We use the softmax recommendation probability function (7.2). We simulate the dynamics for $T = 1000$ steps, repeated 100 times each with a new initialization. We choose the sign impact function $g(\boldsymbol{u}_j, \boldsymbol{v}_i) = \operatorname{sign}(\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle)$ for creator updates. For user updates, we choose inner product $f(\boldsymbol{v}_i, \boldsymbol{u}_j) = \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle$. The inner product function is studied in previous works on users' preference dynamics (but not creators') [DM22]. Note that the inner product does not satisfy the condition $|f(\boldsymbol{v}_i, \boldsymbol{u}_j)| \geq L_f$ needed in Theorem 7.1. However, we still observe convergence to polarization in nearly all experiments. Thus, even when this condition does not hold, users and creators still tend towards polarization in practice.

Three key parameters in our model are β (sensitivity of the softmax function), η_c (creator update rate), and η_u (user update rate). We set them to $\beta = 1, \eta_c = \eta_u = 0.1$ by default, and change one parameter at a time to see its effect on the dynamics. We also test what happens when some dimensions of the user features are *fixed* features that are not updated.

Measures To quantify the behavior of the dynamics, given user and creator feature vectors $(\boldsymbol{U}, \boldsymbol{V})$ we compute the following measures, which cover diversity, relevancy, and polarization of the system:

- *Creator Diversity* (CD): diversity of the creator features, measured by their average pairwise distance [ZMKL05, NHH+14]:

$$\mathrm{CD}(\boldsymbol{V}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|.$$

- *Recommendation Diversity* (RD): diversity of the contents recommended to a user, measured by the weighted variance of the contents:

$$\mathrm{RD}(\boldsymbol{U}, \boldsymbol{V}; \beta) = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij} \|\boldsymbol{v}_i - \overline{\boldsymbol{v}}_j\|^2,$$

  where $\overline{\boldsymbol{v}}_j = \sum_{i=1}^n p_{ij} \boldsymbol{v}_i$ and $p_{ij} = \frac{\exp(\beta \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle)}{\sum_{i=1}^n \exp(\beta \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle)}$.

- *Recommendation Relevance* (RR): relevance of the contents recommended to a user, measured by the weighted average of inner products:

$$\mathrm{RR}(\boldsymbol{U}, \boldsymbol{V}; \beta) = \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij} \langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle.$$

- *Tendency to Polarization* (TP): This is a novel measure we propose to quantify how close the system is to consensus or bi-polarization, measured by the average absolute inner products between the creators:

$$\mathrm{TP}(\boldsymbol{V}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{n} |\langle \boldsymbol{v}_i, \boldsymbol{v}_k \rangle|.$$

  $\mathrm{TP}(\boldsymbol{V})$ being closer to 1 means that the system is more polarized, because the term $|\langle \boldsymbol{v}_i, \boldsymbol{v}_k \rangle|$ is 1 iff the two vectors $\boldsymbol{v}_i, \boldsymbol{v}_k$ are equal or opposite to each other.

It is worth noting that a high creator diversity is necessary for simultaneously achieving high recommendation relevance and high recommendation diversity. For example, they cannot be simultaneously achieved in a polarized state.

7.5.2 Sensitivity Parameter β

A larger β means that a user will be recommended more relevant content/creator with a higher probability. $\beta = 0$, on the other hand, means that the user receives uniform recommendations across all creators. **Our main observation** from the experiments is:

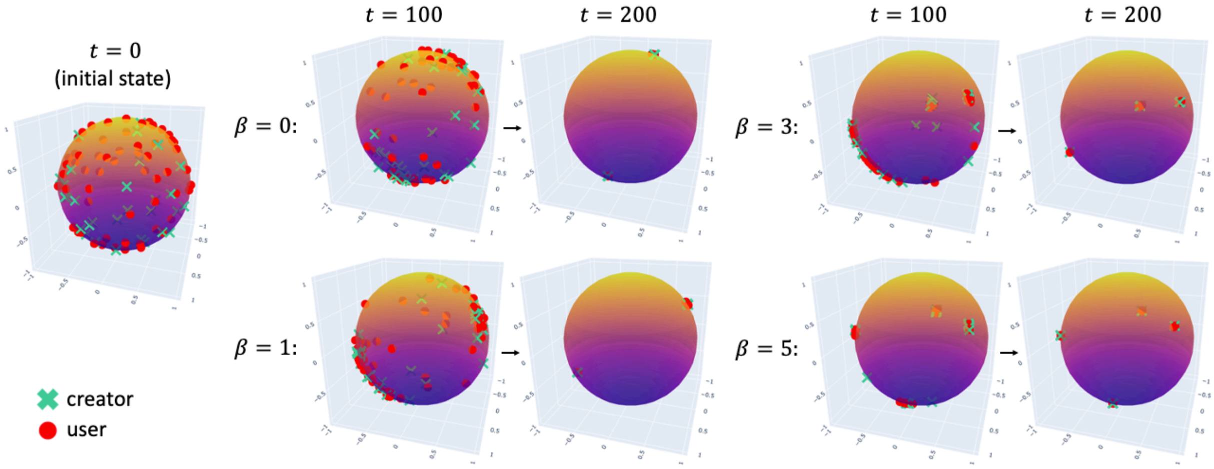


Figure 7.1: Snapshots of the dynamics simulated with the same initialization but different recommendation sensitivity β. A larger β resulted in more clusters at time step $t = 200$.

a larger β leads to higher creator diversity and alleviated polarization in the long run.

Figure 7.1 shows snapshots of the dynamics at different time steps under different β values. Here, we choose dimension $d = 3$ instead of 10 so the feature vectors can be visualized on a 3d sphere. We see that the system tends to form more clusters at time $t = 200$ as β increases.

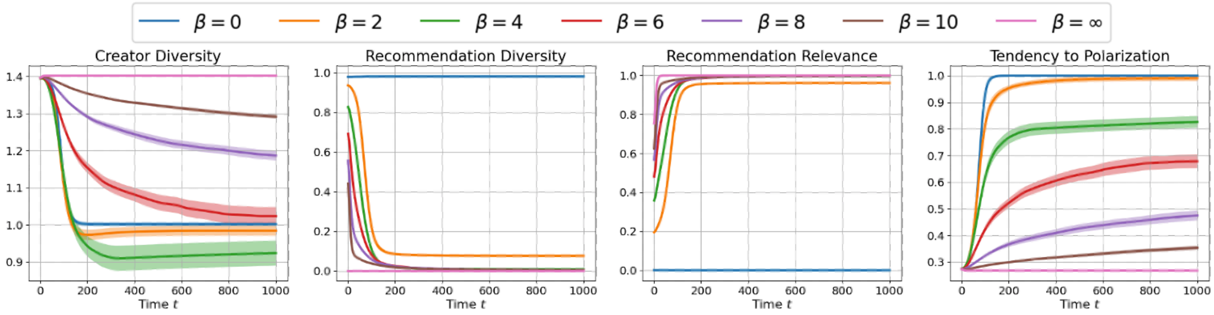


Figure 7.2: Changes of measures over time under different sensitivity parameter β, on synthetic data. Larger β reduces the tendency to polarization.

Figure 7.2 shows the changes of the 4 measures CD, RD, RR, TP over time under different β values. $\beta = 0$ means uniform (non-personalized) recommendation. $\beta = \infty$ means hard-max recommendation: only recommend the single most relevant creator to a user. We see that a more diverse recommendation policy (a smaller β) leads to lower creator diversity and a higher level of polarization in the long run. In particular, while Creator Diversity reaches a similar level under different β in the end, it *drops at a slower rate* with a *larger* β (see $\beta = 5, 6$). Moreover, from the plot of Tendency to Polarization, we see that a larger β *alleviates* polarization, which means improvement in the diversity of the whole system.

An explanation for our observation is the following: When β is smaller, each user receives more uniform recommendations across all creators. So, for different creators, the sets of users recommended to those creators have larger intersections. Since the creator updates are based on the sets of recommended users, different creators will be moving towards more similar directions. This leads to faster polarization. One can also predict this observation from Theorem 7.1: when β is large, the minimum recommendation probability p_0 of the softmax function tends to 0, so it might take a long time for the system to converge to polarization, while with a small β the system polarizes quickly.

7.5.3 Update Rates η_c and η_u

A larger η_c means that creator features are updated faster, and intuitively should lead to faster polarization. This is validated in experiments: Figure 7.3 shows that a larger η_c indeed causes more extreme polarization and lower diversity (both CD and RD). A larger η_u means that user features are updated faster. It has a similar effect of exacerbating polarization as η_c does, as shown in Figure 7.4.

7.5.4 Number of Fixed Dimensions

We also consider the scenario where some dimensions of the user feature vectors are fixed features and thus not updated from round to round (e.g., age, gender), which is a realistic scenario. Formally, we fix the first $k \le d$ dimensions. The remaining $d - k$ dimensions $\boldsymbol{u}_j^t[k+1 : d] = (u_j^t[k+1], \dots, u_j^t[d])$ are updated according to the following

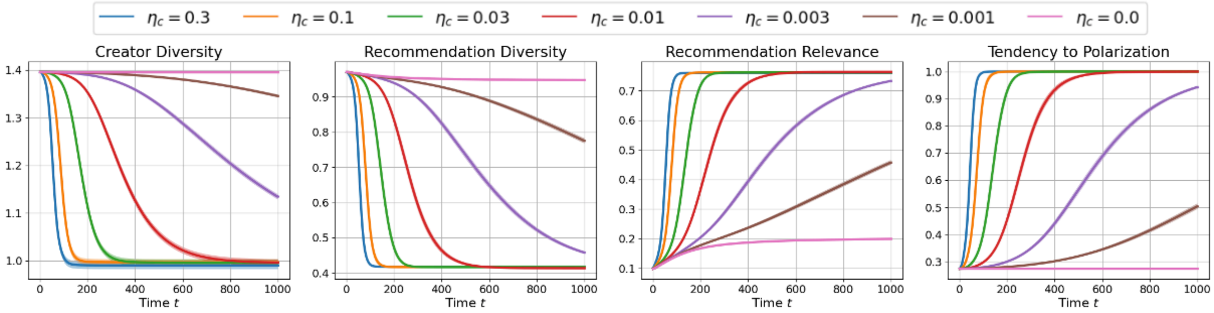


Figure 7.3: Changes of measures over time under different creator update rate η_c, on synthetic data

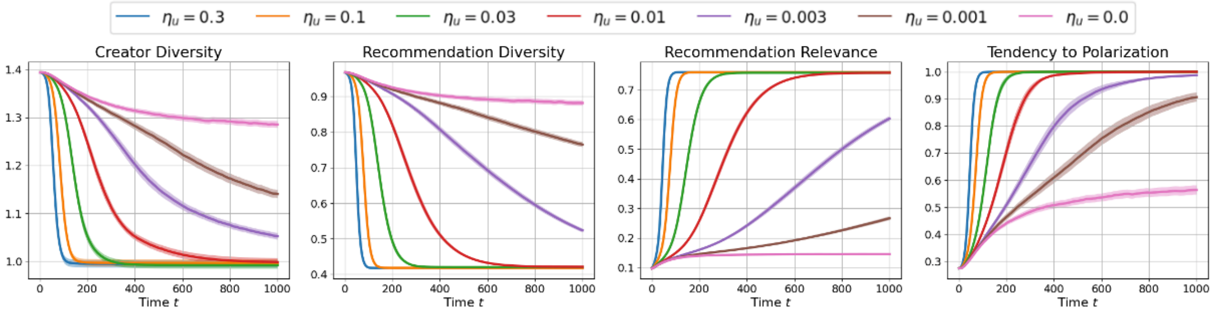


Figure 7.4: Changes of measures over time under different user update rate η_u, on synthetic data

rule: $\boldsymbol{u}_j^{t+1}[k+1:d] = \|\boldsymbol{u}_j^t[k+1:d]\| \cdot \mathcal{P}\big(\boldsymbol{u}_j^t[k+1:d] + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t[k+1:d]\big)$. The multiplication by $\|\boldsymbol{u}_j^t[k+1:d]\|$ ensures unit norm $\|\boldsymbol{u}_j^{t+1}\| = 1$.

The effect of the number of fixed dimensions on the dynamics is shown in Figure 7.5. The **main observation** is: *as the number of fixed dimensions increases, the diversity of the system improves and the degree of polarization is reduced.* This is similar to the effect of decreasing user update rate η_u in Figure 7.4. The observation that fixed dimensions of user features help to improve diversity might be a reason why the recommender systems in practice are not as polarized as our theoretical prediction.

7.5.5 Top-k Truncation

We experimented with top-k truncation. Our **main observation** is: *a small k improves the diversity of the recommender system and reduces polarization.* This is consistent with

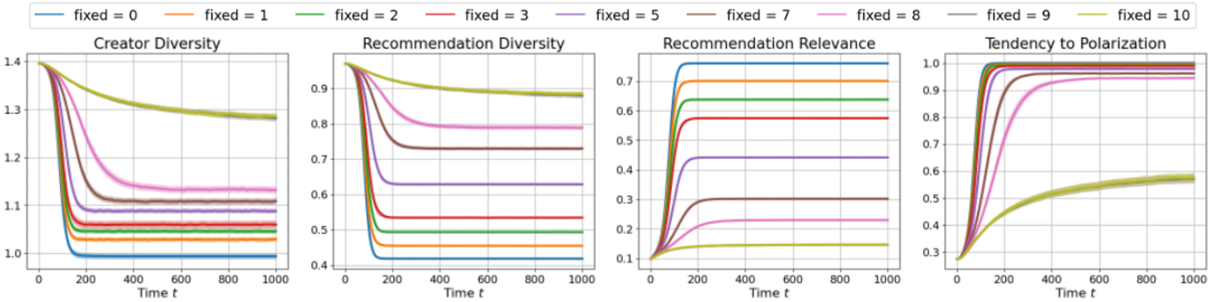


Figure 7.5: Changes of measures over time under different numbers of fixed dimensions, on synthetic data

our theoretical prediction (Proposition 7.2). However, there is a tradeoff between the diversity of recommendations to users (RD) and the diversity of creations in the system (CD and TP). A top-k truncation policy with small k is “not diverse” for users because it exposes a user only to a small set of contents. However, such a policy can lead to a more diverse outcome in the whole system. This tradeoff is worth further studying.

Table 7.2: Diversity improvement by top-k truncation on synthetic data

β	k	Creator Diversity	Recommendation Diversity	Recommendation Relevance	Tendency to Polarization
1	50	$1.00_{\pm.03}$	$\mathbf{0.42}_{\pm\mathbf{0.01}}$	$0.76_{\pm0.01}$	$1.00_{\pm10^{-3}}$
	25	$0.52_{\pm.32}$	$0.03_{\pm0.03}$	$0.97_{\pm0.02}$	$0.91_{\pm0.13}$
	20	$0.91_{\pm.15}$	$0.00_{\pm0.01}$	$1.00_{\pm0.01}$	$0.68_{\pm0.12}$
	10	$1.17_{\pm.06}$	$0.00_{\pm10^{-3}}$	$1.00_{\pm10^{-3}}$	$0.50_{\pm0.07}$
	5	$1.31_{\pm.02}$	$0.00_{\pm10^{-3}}$	$1.00_{\pm10^{-3}}$	$0.35_{\pm0.03}$
	1	$\mathbf{1.40}_{\pm\mathbf{10^{-3}}}$	$0.00_{\pm10^{-3}}$	$\mathbf{1.00}_{\pm\mathbf{10^{-3}}}$	$\mathbf{0.27}_{\pm\mathbf{10^{-3}}}$
3	50	$0.95_{\pm.14}$	$\mathbf{0.02}_{\pm\mathbf{0.02}}$	$0.99_{\pm0.01}$	$0.91_{\pm0.10}$
	25	$0.80_{\pm.24}$	$0.00_{\pm0.01}$	$1.00_{\pm10^{-3}}$	$0.77_{\pm0.13}$
	20	$0.89_{\pm.13}$	$0.00_{\pm10^{-3}}$	$1.00_{\pm10^{-3}}$	$0.74_{\pm0.11}$
	10	$1.18_{\pm.05}$	$0.00_{\pm10^{-3}}$	$1.00_{\pm10^{-3}}$	$0.49_{\pm0.07}$
	5	$1.31_{\pm.02}$	$0.00_{\pm10^{-3}}$	$1.00_{\pm10^{-3}}$	$0.34_{\pm0.03}$
	1	$\mathbf{1.40}_{\pm\mathbf{10^{-3}}}$	$0.00_{\pm10^{-3}}$	$\mathbf{1.00}_{\pm\mathbf{10^{-3}}}$	$\mathbf{0.27}_{\pm\mathbf{10^{-3}}}$

7.5.6 Threshold Truncation

We also experimented with threshold truncation. The effect of a large truncation threshold τ is similar to the effect of a small k in top-k truncation.

Table 7.3 shows the effect of different thresholds in threshold truncation on the long-term diversity of the system. We see that truncating at $\tau = 0$, which corresponds to 90° angle between $\boldsymbol{u}_j$ and $\boldsymbol{v}_i$, is *not good* for diversity, resulting in the lowest creator diversity measure (CD) and highest tendency to polarization (TP). Truncating at a large threshold like 0.707 is good for diversity, instead. Figure 7.6 shows how the diversity measures change over time, under different truncation thresholds.

Table 7.3: Diversity improvement by threshold truncation on synthetic data

β	threshold τ	CD	RD	RR	TP
0	$-\cos(60°) = -0.5$	1.00 ± 0.03	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.99 ± 10^{-3}
	$-\cos(72°) = -0.309$	0.96 ± 0.06	$\mathbf{0.01 \pm 0.02}$	1.00 ± 0.02	0.92 ± 0.10
	$\cos(90°) = 0$	0.03 ± 0.16	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.99 ± 0.04
	$\cos(72°) = 0.309$	0.72 ± 0.30	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.81 ± 0.12
	$\cos(60°) = 0.5$	1.16 ± 0.11	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.47 ± 0.10
	$\cos(45°) = 0.707$	$\mathbf{1.37 \pm 0.02}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.33 \pm 0.02}$
	$\cos(30°) = 0.866$	1.30 ± 0.03	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.55 ± 0.05
1	$-\cos(60°) = -0.5$	0.98 ± 0.04	0.00 ± 0.02	1.00 ± 0.01	0.96 ± 0.04
	$-\cos(72°) = -0.309$	0.92 ± 0.08	0.00 ± 0.02	0.99 ± 0.02	0.87 ± 0.10
	$\cos(90°) = 0$	0.13 ± 0.31	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.97 ± 0.08
	$\cos(72°) = 0.309$	0.85 ± 0.16	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.76 ± 0.11
	$\cos(60°) = 0.5$	1.21 ± 0.07	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.43 ± 0.08
	$\cos(45°) = 0.707$	$\mathbf{1.38 \pm 0.01}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.30 \pm 0.01}$
	$\cos(30°) = 0.866$	1.33 ± 0.02	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.47 ± 0.04
3	$-\cos(60°) = -0.5$	0.91 ± 0.18	$\mathbf{0.01 \pm 0.02}$	1.00 ± 0.01	0.83 ± 0.10
	$-\cos(72°) = -0.309$	0.85 ± 0.23	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.78 ± 0.11
	$\cos(90°) = 0$	0.64 ± 0.33	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.81 ± 0.12
	$\cos(72°) = 0.309$	1.01 ± 0.14	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.64 ± 0.14
	$\cos(60°) = 0.5$	1.26 ± 0.05	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.38 ± 0.06
	$\cos(45°) = 0.707$	$\mathbf{1.39 \pm 0.01}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.28 \pm 0.01}$
	$\cos(30°) = 0.866$	1.37 ± 0.01	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.34 ± 0.01

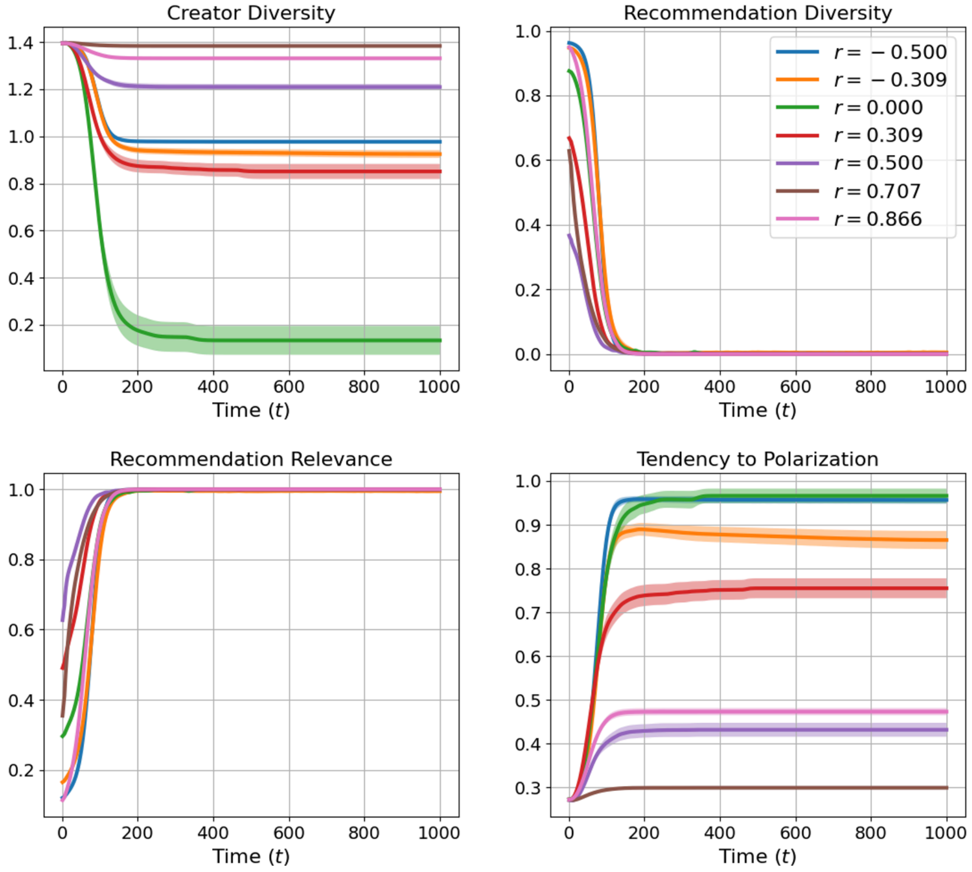


Figure 7.6: Changes of measures over time under different truncation threshold τ, with $\beta = 1$, on synthetic data

7.6 Discussion

Summary of Our Contributions. Our work defines a dynamics model to capture the dual influence of recommender systems on adaptive users and strategic creators. Although our model is a theoretical abstraction, we believe that it captures the essence of a real-world recommender system, and our effort is an important initial endeavor to study diversity in recommender systems with dual influence. We specifically point out different concepts of diversity in recommender systems (creation diversity, recommendation diversity, and tendency to polarization) and provide theoretical and empirical evidences to show that, due to dual influence, myopically optimizing recommendation diversity might hurt the long-term creation diversity and result in polarization of the system. We also explore popular design choices in recommender systems and show an interesting and somewhat counter-intuitive result that designs purely targeting efficiency improvement (e.g., top-k truncation) can alleviate polarization. We believe that the insights from our work are valuable to building healthy and sustainable recommender systems, and our results can inspire more sophisticated solutions for improving the long-term diversity of recommender systems to be developed.

Below, we discuss some real-world recommender system properties and designs that are not covered in our work.

User and Creator Retention and Activeness. In our current model, the users and creators stay in the system from the start to the end. However, in real-world recommender systems, users and creators may leave the platform either permanently or for a certain period. Meanwhile, new users and creators will join the platform. Such join and leave dynamics are also influenced by the recommendations' relevance and diversity, which further complicate the problem. Moreover, users and creators have different activeness levels on the platform, e.g., some users may watch a lot more videos than others, and some creators may post a lot more creations, these effects will also be strongly correlated with the dual influence of the recommender system.

Creation Quality. Creation quality is a major factor influencing users' feedback in addition to the creation style, e.g., well-made cuisine videos could also be fun and liked by gamers and pet lovers, which we need more than a collaborative filtering type of modeling like our current model to capture such features. A potential solution to boost both long-term system diversity and single-shot recommendation diversity is to design mechanisms that can incentivize creators to create higher-quality videos instead of changing their creation styles.

Cold Start. Cold Start is widely used in real-world recommender systems for newly published items. Due to the lack of user-item interactions on new items, the systems randomly recommend these new items to users and collect data for collaborative filtering. In our current model, if we consider the creators creating new items in each time step under their current time creation style, then cold start satisfies the conditions in Theorem 7.1. But if we consider the system to have good enough content understanding ability and can accurately predict the new creations' embeddings, the cold start is not necessary and our model and results in the top-k truncation and threshold truncation parts are valid. We also highlight a subtle difference between cold start and random traffic, if cold start is used on creators instead of items, then after the creator is exposed to users a certain number of times, the system will not guarantee to provide a non-zero probability of recommending this creator, and thus the conditions in Theorem 7.1 may not hold.

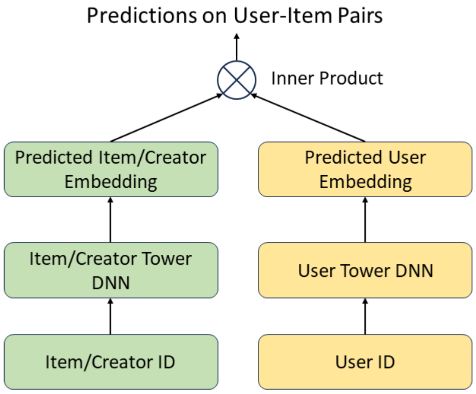


Figure 7.7: Two tower model for the MovieLens experiment, where the two towers both have size 16 × 16 with linear layers and ReLu activations.

7.7 Experiments: Real-World Data

7.7.1 Experiment Setup

Algorithm 7.1: Real-world Recommendation with Dual Influence

```
Input: $t = 0$, actual embedding $U^{(0)}, V^{(0)}$, true labels $Y_{ij}^{(0)} := y(u_i^{(0)}, v_j^{(0)})$, initial
       parameter $\boldsymbol{\theta}^{(0)}$ (which includes the predicted embedding $\hat{U}^{(0)}, \hat{V}^{(0)}$)
repeat
    Let temporary parameter $\boldsymbol{w}^{(0)} \leftarrow \boldsymbol{\theta}^{(t)}$ ;
    Compute loss $\mathcal{L}(\boldsymbol{\theta}^{(t)}, Y^{(t)})$ ;
    for $s = 1$ to $m - 1$ do
        $\boldsymbol{w}^{(s+1)} \leftarrow \boldsymbol{\theta}^{(s)} - \eta \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^{(s)}, Y^{(t)})$ ;
    $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{w}^{(m)}$ ;
    Deliver recommendations based on $\hat{U}^{(t+1)}, \hat{V}^{(t+1)}$ ;
    Update $U^{(t+1)}, V^{(t+1)}$, and $Y^{(t+1)}$ ;
    $t \leftarrow t + 1$ ;
until $\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|_2 \leq \delta$;
```

We conduct experiments on the MovieLens 20M dataset [HK15]. We use a real-world two-tower recommendation model with 16-dimensional tower tops as the user and creator embeddings (Figure 7.7). The model is initialized by fitting a two-tower model [HHG+13] on the existing MovieLens rating data and using the tower tops as the initial user and creator embeddings. Then we follow Algorithm 7.1 to simulate the dynamics.

7.7.2 Effect of Sensitivity Parameter β

Figure 7.8 shows the effect of the recommendation sensitivity parameter β on the system. Similar to the synthetic data experiments, a smaller β (more diverse recommendation for the users in the short term) results in faster polarization. We note that the joint results on CD and TP are more informative than each one alone: despite $\beta = 0$ has a higher creator diversity than $\beta = 2$ at $T = 500$, the system reaches polarization more quickly under $\beta = 0$. The higher creator diversity under $\beta = 0$ is because the two clusters in the bi-polarized state are more balanced so the average pairwise distance between the creators is higher under $\beta = 0$ than under $\beta = 2$.

7.7.3 Effect of Diversity Boosting Parameter ρ

Figure 7.9 shows the effect of using diversity-aware objective (Equation (7.5)) for diversity boosting. We see that myopically promoting the short-term recommendation diversity (using a larger ρ) results a higher creation diversity but also a higher tendency to polarization. A possible explanation for this phenomenon is, similar to the case with β, the system polarizes into two balanced clusters which actually have a large average pairwise distance. In this case, Tendency to Polarization is a better measure for diversity loss than Creator Diversity (average pairwise distance).

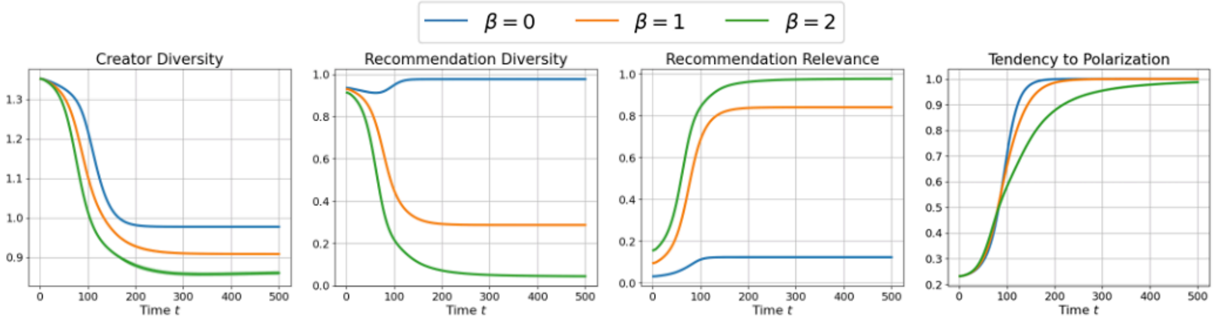


Figure 7.8: Experiment on MovieLens 20M dataset under different recommendation sensitivity β

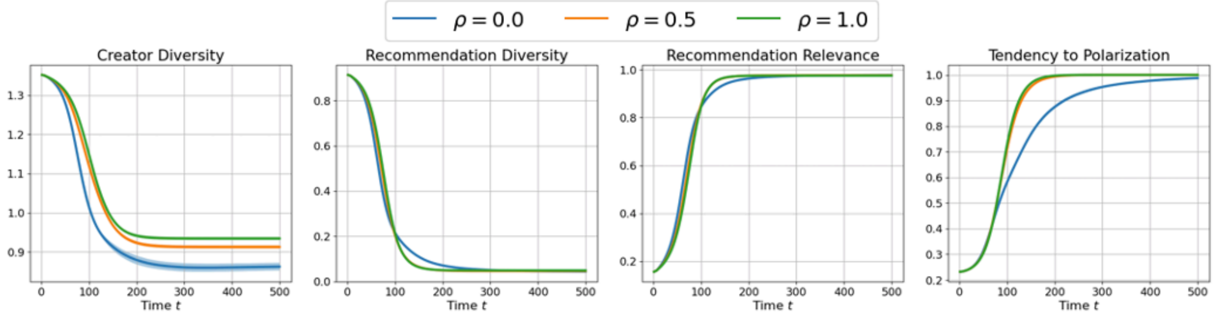


Figure 7.9: Experiment on MovieLens 20M dataset with diversity-aware objective under different ρ

7.7.4 Top-k Truncation

We try top-k truncation (Section 7.4) on the MovieLens 20M dataset. Here, we have $n = 2000$ creators and $m = 2000$ users, with feature dimension $d = 16$. The results for top-k truncation are in Table 7.4 and Figure 7.10. Similar to the experiments with synthetic data, we see that a smaller k improves Creator Diversity (CD) and Recommendation Relevance (RR), reduces Tendency to Polarization (TP), yet worsens Recommendation Diversity (RD).

Table 7.4: Diversity improvement by top-k truncation on MovieLens 20M dataset

β	k	CD	RD	RR	TP
0	2000	1.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	0.00 ± 10^{-3}	1.00 ± 10^{-3}
	1000	0.30 ± 0.04	0.03 ± 0.01	0.88 ± 0.01	1.00 ± 10^{-3}
	500	1.10 ± 0.06	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.43 ± 0.03
	100	1.36 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.28 ± 0.01
	10	1.40 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.20 ± 10^{-3}
	1	$\mathbf{1.40 \pm 10^{-3}}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.20 \pm 10^{-3}}$
1	2000	1.00 ± 10^{-3}	$\mathbf{0.42 \pm 10^{-3}}$	0.92 ± 0.01	1.00 ± 10^{-3}
	1000	0.61 ± 0.16	0.03 ± 0.01	0.97 ± 0.01	0.90 ± 0.06
	500	1.14 ± 0.04	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.41 ± 0.04
	100	1.35 ± 0.01	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.27 ± 10^{-3}
	10	1.40 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.20 ± 10^{-3}
	1	$\mathbf{1.40 \pm 10^{-3}}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.20 \pm 10^{-3}}$
3	2000	0.92 ± 0.07	$\mathbf{0.02 \pm 0.01}$	0.99 ± 10^{-3}	0.91 ± 0.05
	1000	0.65 ± 0.18	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.69 ± 0.14
	500	1.07 ± 0.07	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.48 ± 0.11
	100	1.36 ± 0.01	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.27 ± 0.01
	10	1.40 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.20 ± 10^{-3}
	1	$\mathbf{1.40 \pm 10^{-3}}$	0.00 ± 10^{-3}	$\mathbf{1.00 \pm 10^{-3}}$	$\mathbf{0.20 \pm 10^{-3}}$

7.7.5 Threshold Truncations

Results for threshold truncation (Section 7.4) are in Table 7.5 and Figure 7.11. Similar to synthetic data, we see that a large (but not too large) threshold like 0.707 is good for improving CD and TP.

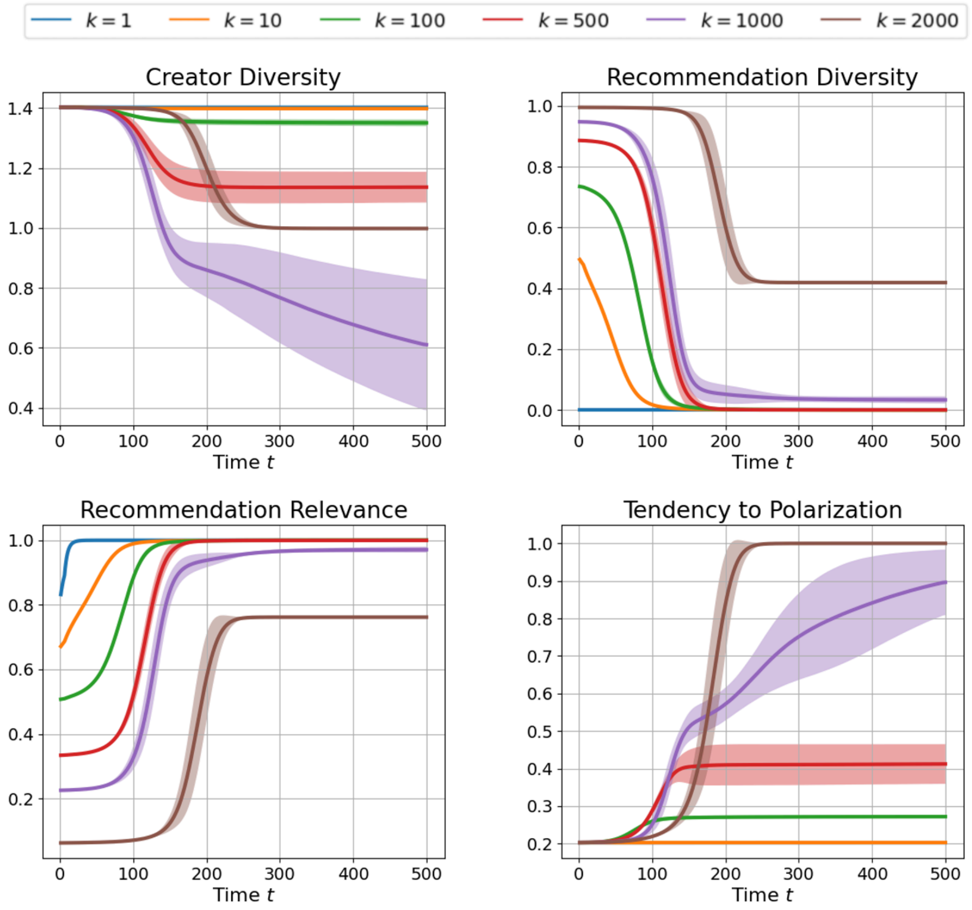


Figure 7.10: Changes of measures over time under different k, with $\beta = 1$, on MovieLens 20M dataset

7.8 Useful Lemmas

This section provides some lemmas that will be used in the proofs. They are mainly about some properties of the dynamics update rule.

Claim 7.1. *For vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ with unit norm $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$, we have:*

- $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = 2(1 - \langle \boldsymbol{x}, \boldsymbol{y} \rangle)$.

- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 1 - \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.

Table 7.5: Threshold truncation with different thresholds on MovieLens 20M dataset

β	threshold τ	CD	RD	RR	TP
0	$-\cos(60°) = -0.5$	1.00 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	1.00 ± 10^{-3}
	$-\cos(72°) = -0.309$	1.00 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	1.00 ± 10^{-3}
	$\cos(90°) = 0$	0.01 ± 0.01	0.00 ± 10^{-3}	1.00 ± 10^{-3}	1.00 ± 10^{-3}
	$\cos(72°) = 0.309$	0.83 ± 0.08	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.72 ± 0.09
	$\cos(60°) = 0.5$	1.20 ± 0.05	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.46 ± 0.07
	$\cos(45°) = 0.707$	$\mathbf{1.39 \pm 10^{-3}}$	0.00 ± 10^{-3}	1.00 ± 10^{-3}	$\mathbf{0.20 \pm 10^{-3}}$
	$\cos(30°) = 0.866$	1.36 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.40 ± 0.01
1	$-\cos(60°) = -0.5$	1.00 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	1.00 ± 10^{-3}
	$-\cos(72°) = -0.309$	0.96 ± 0.03	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.95 ± 0.03
	$\cos(90°) = 0$	0.02 ± 0.02	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.99 ± 10^{-3}
	$\cos(72°) = 0.309$	0.83 ± 0.07	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.66 ± 0.10
	$\cos(60°) = 0.5$	1.18 ± 0.06	0.00 ± 10^{-3}	1.00 ± 0.01	0.44 ± 0.07
	$\cos(45°) = 0.707$	$\mathbf{1.40 \pm 10^{-3}}$	0.00 ± 10^{-3}	1.00 ± 10^{-3}	$\mathbf{0.20 \pm 10^{-3}}$
	$\cos(30°) = 0.866$	1.35 ± 0.01	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.40 ± 0.02
3	$-\cos(60°) = -0.5$	0.77 ± 0.27	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.86 ± 0.09
	$-\cos(72°) = -0.309$	0.80 ± 0.24	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.79 ± 0.13
	$\cos(90°) = 0$	0.04 ± 0.02	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.98 ± 0.01
	$\cos(72°) = 0.309$	0.99 ± 0.11	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.55 ± 0.13
	$\cos(60°) = 0.5$	1.26 ± 0.05	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.36 ± 0.06
	$\cos(45°) = 0.707$	$\mathbf{1.40 \pm 10^{-3}}$	0.00 ± 10^{-3}	1.00 ± 10^{-3}	$\mathbf{0.20 \pm 10^{-3}}$
	$\cos(30°) = 0.866$	1.36 ± 10^{-3}	0.00 ± 10^{-3}	1.00 ± 10^{-3}	0.39 ± 0.01

Lemma 7.2 (Convex Cone Property)**.** *Let* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k \in \mathbb{R}^d$ *be vectors with norm* $\|\boldsymbol{z}_i^t\|_2 = 1$. *Suppose* $\langle \boldsymbol{z}_i, \boldsymbol{y} \rangle > 0$ *for every* $i = 1, \ldots, k$ *for some* $\boldsymbol{y} \in \mathbb{R}^d$. *Let* $\boldsymbol{x} = \mathcal{P}(\sum_{i=1}^k a_i \boldsymbol{z}_i)$ *for some* $a_1, \ldots, a_k \geq 0$ *(namely,* $\boldsymbol{x}$ *is the normalization of some vector in the convex cone formed by* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$*). Then, we have*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \;\geq\; \min_{i=1}^{k} \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle \;>\; 0 \quad \textit{and} \quad \|\boldsymbol{x} - \boldsymbol{y}\|_2 \;\leq\; \max_{i=1}^{k} \|\boldsymbol{z}_i - \boldsymbol{y}\|_2 \;>\; 0.$$

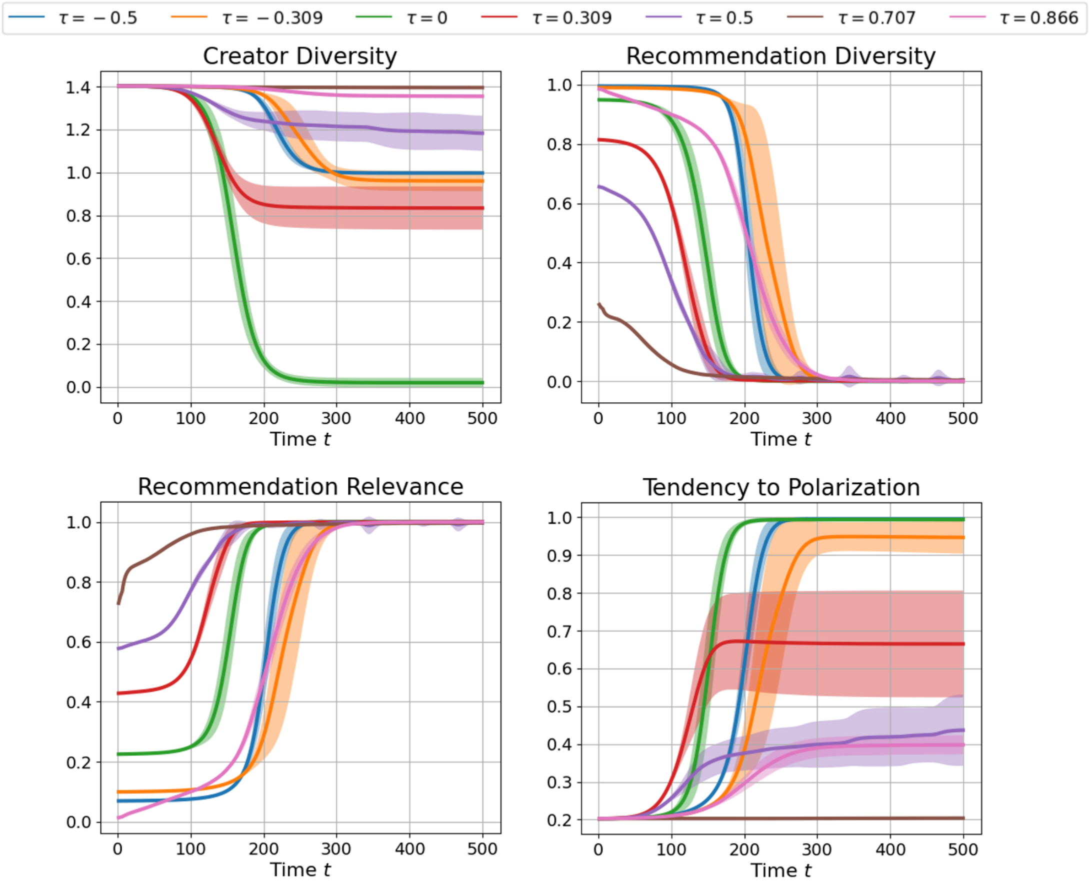


Figure 7.11: Changes of measures over time under truncation with different threshold τ, with $\beta = 1$, on MovieLens 20M dataset

Proof.

$$
\begin{aligned}
\langle \boldsymbol{x},\, \boldsymbol{y} \rangle &= \left\langle \frac{\sum_{i=1}^{k} a_i \boldsymbol{z}_i}{\| \sum_{i=1}^{k} a_i \boldsymbol{z}_i \|_2},\, \boldsymbol{y} \right\rangle = \frac{1}{\| \sum_{i=1}^{k} a_i \boldsymbol{z}_i \|_2} \sum_{i=1}^{k} a_i \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle \\
&\geq \frac{1}{\| \sum_{i=1}^{k} a_i \boldsymbol{z}_i \|_2} \sum_{i=1}^{k} a_i \min_{i=1}^{k} \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle = \min_{i=1}^{k} \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle \frac{\sum_{i=1}^{k} a_i}{\| \sum_{i=1}^{k} a_i \boldsymbol{z}_i \|_2} \\
&\geq \min_{i=1}^{k} \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} a_i} = \min_{i=1}^{k} \langle \boldsymbol{z}_i, \boldsymbol{y} \rangle.
\end{aligned}
$$

This proves the first inequality. To prove the second inequality, we use Claim 7.1 and the first inequality:

$$\|\boldsymbol{x}-\boldsymbol{y}\|_2 \;=\; \sqrt{2(1-\langle \boldsymbol{x},\boldsymbol{y}\rangle)} \;\le\; \sqrt{2(1-\min_i \langle \boldsymbol{z}_i,\boldsymbol{y}\rangle)} \;=\; \max_i \sqrt{2(1-\langle \boldsymbol{z}_i,\boldsymbol{y}\rangle)} \;=\; \max_{i=1}^{k} \|\boldsymbol{z}_i-\boldsymbol{y}\|_2.$$

□

Lemma 7.3. *Let $\boldsymbol{x}^t, \boldsymbol{y}, \boldsymbol{z}^t \in \mathbb{R}^d$ be vectors with norm $\|\boldsymbol{x}^t\|_2 = 1$, $\|\boldsymbol{y}\|_2 \ge 0$, $\|\boldsymbol{z}^t\|_2 \le 1$. Suppose $\langle \boldsymbol{x}^t, \boldsymbol{y}\rangle \ge 0$, $\langle \boldsymbol{z}^t, \boldsymbol{y}\rangle \ge 0$. After the update $\boldsymbol{x}^{t+1} = \mathcal{P}(\boldsymbol{x}^t + \eta \boldsymbol{z}^t)$, we have*

$$\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t,\, \boldsymbol{y}\rangle \;\ge\; \frac{\eta}{1+\eta\|\boldsymbol{z}^t\|_2}\Big(\langle \boldsymbol{z}^t, \boldsymbol{y}\rangle - \|\boldsymbol{z}^t\|_2 \langle \boldsymbol{x}^t, \boldsymbol{y}\rangle\Big).$$

As a corollary, if $\boldsymbol{y} = \boldsymbol{z}^t$ and $\|\boldsymbol{z}^t\|_2 = 1$, then

$$\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t,\, \boldsymbol{z}^t\rangle \;\ge\; \frac{\eta}{1+\eta}\Big(1 - \langle \boldsymbol{x}^t, \boldsymbol{z}^t\rangle\Big).$$

Proof. By definition,

$$\begin{aligned}
\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t,\, \boldsymbol{y}\rangle \;&=\; \Big\langle \frac{\boldsymbol{x}^t + \eta \boldsymbol{z}^t}{\|\boldsymbol{x}^t + \eta \boldsymbol{z}^t\|_2} - \boldsymbol{x}^t,\, \boldsymbol{y}\Big\rangle \\
&=\; \Big(\frac{1}{\|\boldsymbol{x}^t + \eta \boldsymbol{z}^t\|_2} - 1\Big)\cdot \langle \boldsymbol{x}^t, \boldsymbol{y}\rangle \;+\; \frac{\eta}{\|\boldsymbol{x}^t + \eta \boldsymbol{z}^t\|_2}\cdot \langle \boldsymbol{z}^t, \boldsymbol{y}\rangle \\
\text{(because } \|\boldsymbol{x}^t + \eta \boldsymbol{z}^t\|_2 \le 1 + \eta\|\boldsymbol{z}^t\|_2\text{)} \;&\ge\; \Big(\frac{1}{1+\eta\|\boldsymbol{z}^t\|_2} - 1\Big)\cdot \langle \boldsymbol{x}^t, \boldsymbol{y}\rangle \;+\; \frac{\eta}{1+\eta\|\boldsymbol{z}^t\|_2}\cdot \langle \boldsymbol{z}^t, \boldsymbol{y}\rangle \\
&=\; \frac{\eta}{1+\eta\|\boldsymbol{z}^t\|_2}\Big(\langle \boldsymbol{z}^t, \boldsymbol{y}\rangle - \|\boldsymbol{z}^t\|_2 \langle \boldsymbol{x}^t, \boldsymbol{y}\rangle\Big).
\end{aligned}$$

□

Lemma 7.4. *Let $\boldsymbol{x}^t, \boldsymbol{z}^t \in \mathbb{R}^d$ be vectors with norm $\|\boldsymbol{x}^t\|_2 = 1$, $\|\boldsymbol{z}^t\|_2 \le 1$. Suppose $\langle \boldsymbol{x}^t, \boldsymbol{z}^t\rangle \ge 0$ and $\eta > 0$. Then the update $\boldsymbol{x}^{t+1} = \mathcal{P}(\boldsymbol{x}^t + \eta \boldsymbol{z}^t)$ satisfies*

- $\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \boldsymbol{z}^t \rangle \ \geq \ \frac{1}{\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2$.
- $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2 \ \leq \ \eta \|\boldsymbol{z}^t\|_2$.

Proof. Let $\tilde{\boldsymbol{x}}^{t+1} = \boldsymbol{x}^t + \eta \boldsymbol{z}^t$, so $\boldsymbol{x}^t = \mathcal{P}(\tilde{\boldsymbol{x}}^{t+1})$ and $\boldsymbol{z}^t = \frac{1}{\eta}(\tilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t)$. Then we have

$$\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \boldsymbol{z}^t \rangle \ = \ \frac{1}{\eta} \langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \tilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \rangle.$$

Because $\langle \boldsymbol{x}^t, \boldsymbol{z}^t \rangle \geq 0$, the vector $\tilde{\boldsymbol{x}}^{t+1} = \boldsymbol{x}^t + \eta \boldsymbol{z}^t$ has length ≥ 1 and hence is outside (or on the surface) of the d-dimensional unit ball. Since $\boldsymbol{x}^t = \mathcal{P}(\tilde{\boldsymbol{x}}^{t+1})$ is the projection of $\tilde{\boldsymbol{x}}^{t+1}$ onto the unit ball, and $\boldsymbol{z}^t$ is another vector inside the unit ball, by the "Pythagorean property" (Proposition 2.2 in [BG19]), we must have $\langle \boldsymbol{x}^t - \boldsymbol{x}^{t+1}, \tilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^{t+1} \rangle \leq 0$. This implies

$$\begin{aligned} \langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \boldsymbol{z}^t \rangle \ &\geq \ \frac{1}{\eta} \Big(\langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \tilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^t \rangle + \langle \boldsymbol{x}^t - \boldsymbol{x}^{t+1}, \, \tilde{\boldsymbol{x}}^{t+1} - \boldsymbol{x}^{t+1} \rangle \Big) \\ &= \ \frac{1}{\eta} \langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \boldsymbol{x}^{t+1} - \boldsymbol{x}^t \rangle \ = \ \frac{1}{\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2, \end{aligned}$$

which proves the first claim. To prove the second claim, we use Cauchy-Schwarz inequality:

$$\frac{1}{\eta} \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2^2 \ \leq \ \langle \boldsymbol{x}^{t+1} - \boldsymbol{x}^t, \, \boldsymbol{z}^t \rangle \ \leq \ \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2 \|\boldsymbol{z}^t\|_2.$$

This implies $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\|_2 \leq \eta \|\boldsymbol{z}^t\|_2$. □

Lemma 7.5. *Consider a creator $\boldsymbol{v}_i^t$ and a user $\boldsymbol{u}_j^t$. Suppose the user is always recommended creator i (so the user is updated by $\boldsymbol{u}_j^{t+1} = \mathcal{P}(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t) \boldsymbol{v}_i^t)$), and creator i is updated by $\boldsymbol{v}_i^{t+1} = \mathcal{P}(\boldsymbol{v}_i^t + \eta_c \boldsymbol{\alpha}_i^t)$ with $\|\boldsymbol{\alpha}_i^t\|_2 \leq 1$ and $\langle \boldsymbol{v}_i^t, \boldsymbol{\alpha}_i^t \rangle \geq 0$ at each time step. Assume:*

- *The inner product* $\langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0 \rangle > 0$ *initially. (Note that* $\langle \boldsymbol{u}_j^0, \boldsymbol{u}_{j'}^0 \rangle$ *needs not hold.)*

- *There exists some constant* $L_f > 0$ *such that* $f(\boldsymbol{v}_i, \boldsymbol{u}_j) \geq L_f > 0$ *whenever* $\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle > 0$.

- $\eta_c \leq \frac{\eta_u L_f}{2}$ *and* $0 \leq \eta_u < \frac{1}{2}$.

Then, we have $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle > 0$ *in all time steps.*

Proof. We prove by induction. Suppose $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle > 0$ already holds. We prove that $\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle > 0$ will also hold. Take the difference between $\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle$ and $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle$:

$$\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \; = \; \langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t \rangle + \langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle.$$

For $\langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle$, using Lemma 7.3 with $\boldsymbol{x}^t = \boldsymbol{u}_j^t$, $\boldsymbol{z}^t = \boldsymbol{v}_i^t$, and $\eta = \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)$, we get

$$\langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \; \geq \; \tfrac{\eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)}{1 + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big) \; \geq \; \tfrac{\eta_u L_f}{1 + \eta_u L_f} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big).$$

For $\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t \rangle$, by Cauchy-Schwarz inequality and Lemma 7.4,

$$\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t \rangle \; \geq \; -\|\boldsymbol{u}_j^{t+1}\|_2 \cdot \|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2 \; \geq \; -1 \cdot \eta_c \|\boldsymbol{\alpha}_i^t\|_2 \; \geq \; -\eta_c.$$

- If $1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle > \frac{1}{2}(1 + \eta_u L_f)$, then we have

  $$\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \; > \; \eta_u L_f \tfrac{1}{2} - \eta_c \; \geq 0$$

  by the assumption of $\eta_c \leq \frac{\eta_u L_f}{2}$.

- If $1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \leq \frac{1}{2}(1 + \eta_u L_f)$, then we have

$$\begin{aligned} \langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \ &\geq \ 0 - \eta_c \\ \implies \langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle \ &\geq \ \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle - \eta_c \ \geq \ \tfrac{1}{2} - \tfrac{1}{2}\eta_u L_f - \eta_c \ > \ 0 \end{aligned}$$

under the assumption of $\eta_c \leq \frac{\eta_u L_f}{2}$ and $\eta_u < \frac{1}{2}$.

The above two cases together ensure $\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle > 0$. □

Lemma 7.6. *Consider a system of one user and one creator that satisfies $\langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0 \rangle > 0$ and $\langle \boldsymbol{u}_j^0, \boldsymbol{y} \rangle > \langle \boldsymbol{v}_i^0, \boldsymbol{y} \rangle > 0$ for some $\boldsymbol{y} \in \mathbb{R}^d$ with $\|\boldsymbol{y}\| \leq 1$ initially. The creator is always recommended to the user (so the updates are $\boldsymbol{u}_j^{t+1} = \mathcal{P}(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t)$ and $\boldsymbol{v}_i^{t+1} = \mathcal{P}(\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t)$). Suppose $\eta_c \leq \frac{\eta_u L_f}{2}$ and $0 \leq \eta_u < \frac{1}{2}$. Then, we have:*

- *$\langle \boldsymbol{u}_j^t, \boldsymbol{y} \rangle > \langle \boldsymbol{v}_i^t, \boldsymbol{y} \rangle > 0$ for all $t \geq 1$.*
- *Suppose $\langle \boldsymbol{u}_j^0, \boldsymbol{y} \rangle - \langle \boldsymbol{v}_i^0, \boldsymbol{y} \rangle = D > 0$. For any $R < D$, after $T = \frac{8}{3\eta_u L_f} \ln \frac{2}{R^2}$ steps, we have $\langle \boldsymbol{v}_i^T, \boldsymbol{y} \rangle - \langle \boldsymbol{v}_i^0, \boldsymbol{y} \rangle \geq \frac{\eta_c}{\eta_u + \eta_c}(D - R)$.*

Proof. We prove the first item by induction. Suppose $\langle \boldsymbol{u}_j^t, \boldsymbol{y} \rangle > \langle \boldsymbol{v}_i^t, \boldsymbol{y} \rangle > 0$ holds. Consider $t+1$. First, by Lemma 7.2, $\langle \boldsymbol{v}_i^{t+1}, \boldsymbol{y} \rangle > 0$ holds. Then, we prove $\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{y} \rangle >$

$\langle \boldsymbol{v}_i^{t+1}, \boldsymbol{y}\rangle$. Let $f = f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)$.

$$\begin{aligned}
\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{t+1}, \boldsymbol{y}\rangle \;&=\; \Big\langle \frac{\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2}, \boldsymbol{y}\Big\rangle - \Big\langle \frac{\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2}, \boldsymbol{y}\Big\rangle \\
=\; & \Big(\frac{1}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2} - \frac{\eta_c}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2}\Big)\langle \boldsymbol{u}_j^t, \boldsymbol{y}\rangle - \Big(\frac{1}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2} - \frac{\eta_u f}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2}\Big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle \\
>\; & \Big(\frac{1}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2} - \frac{\eta_c}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2}\Big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle - \Big(\frac{1}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2} - \frac{\eta_u f}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2}\Big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle \\
=\; & \Big(\frac{1+\eta_u f}{\|\boldsymbol{u}_j^t + \eta_u f \boldsymbol{v}_i^t\|_2} - \frac{1+\eta_c}{\|\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t\|_2}\Big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle \\
=\; & \Big(\frac{1+\eta_u f}{\sqrt{1 + 2\eta_u f \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle + (\eta_u f)^2}} - \frac{1+\eta_c}{\sqrt{1 + 2\eta_c \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle + (\eta_c)^2}}\Big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle.
\end{aligned}$$

Let $a = \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle \leq 1$. We note that the function

$$h(\eta) = \frac{1+\eta}{\sqrt{1 + 2\eta a + \eta^2}} = \sqrt{\frac{1 + 2\eta + \eta^2}{1 + 2\eta a + \eta^2}} = \sqrt{1 + \frac{(2-2a)\eta}{1 + 2\eta a + \eta^2}} = \sqrt{1 + \frac{2(1-a)}{\frac{1}{\eta} + 2a + \eta}}$$

is increasing in $\eta \in [0, 1]$. Under the assumption of $\eta_c \leq \frac{\eta_u L_f}{2} \leq \frac{\eta_u f}{2} < \eta_u f$, we have $h(\eta_c) \leq h(\eta_u f)$ and hence

$$\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{t+1}, \boldsymbol{y}\rangle \;>\; \big(h(\eta_u f) - h(\eta_c)\big)\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle \;\geq\; 0.$$

We then prove the second item. Using Lemma 7.3 for $\boldsymbol{v}_i^{t+1} = \mathcal{P}(\boldsymbol{v}_i^t + \eta_c \boldsymbol{u}_j^t)$, we get

$$\langle \boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t, \boldsymbol{y}\rangle \geq \frac{\eta_c}{1+\eta_c}\Big(\langle \boldsymbol{u}_j^t, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle\Big).$$

Using Lemma 7.3 for $\boldsymbol{u}_j^{t+1} = \mathcal{P}(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t)$ and using the fact $\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{y}\rangle < 0$ proved in item 1,

$$\langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{y}\rangle \geq \frac{\eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)}{1 + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)}\Big(\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{y}\rangle\Big) \geq \frac{\eta_u}{1+\eta_u}\Big(\langle \boldsymbol{v}_i^t, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^t, \boldsymbol{y}\rangle\Big).$$

Rearranging the above two inequalities:

$$\begin{aligned}\frac{1+\eta_c}{\eta_c}\Big(\langle \boldsymbol{v}_i^{t+1}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{t}, \boldsymbol{y}\rangle\Big) &\geq \langle \boldsymbol{u}_j^{t}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{t}, \boldsymbol{y}\rangle;\\ \frac{1+\eta_u}{\eta_u}\Big(\langle \boldsymbol{u}_j^{t+1}, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^{t}, \boldsymbol{y}\rangle\Big) &\geq \langle \boldsymbol{v}_i^{t}, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^{t}, \boldsymbol{y}\rangle.\end{aligned}$$

Summing the above two inequalities over $t = 0, 1, \ldots, T-1$:

$$\frac{1+\eta_c}{\eta_c}\Big(\langle \boldsymbol{v}_i^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{0}, \boldsymbol{y}\rangle\Big) + \frac{1+\eta_u}{\eta_u}\Big(\langle \boldsymbol{u}_j^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^{0}, \boldsymbol{y}\rangle\Big) \geq 0. \tag{7.6}$$

According to Lemma 7.8, after at most $T = \frac{8}{3\eta_u L_f}\ln\frac{2}{R^2}$ steps, we have $\|\boldsymbol{u}_j^T - \boldsymbol{v}_i^T\|_2 \leq R$. This implies $\langle \boldsymbol{u}_j^T, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^T, \boldsymbol{y}\rangle = \langle \boldsymbol{u}_j^T - \boldsymbol{v}_i^T, \boldsymbol{y}\rangle \leq \|\boldsymbol{u}_j^T - \boldsymbol{v}_i^T\| \leq R$ and hence

$$\Big(\langle \boldsymbol{v}_i^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{0}, \boldsymbol{y}\rangle\Big) - \Big(\langle \boldsymbol{u}_j^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{u}_j^{0}, \boldsymbol{y}\rangle\Big) = \Big(\langle \boldsymbol{u}_j^{0}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{0}, \boldsymbol{y}\rangle\Big) - \Big(\langle \boldsymbol{u}_j^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{T}, \boldsymbol{y}\rangle\Big) \geq D - R. \tag{7.7}$$

Multiplying (7.7) by $\frac{1+\eta_u}{\eta_u}$ and adding to (7.6):

$$\Big(\frac{1+\eta_c}{\eta_c} + \frac{1+\eta_u}{\eta_u}\Big)\Big(\langle \boldsymbol{v}_i^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{0}, \boldsymbol{y}\rangle\Big) \geq \frac{1+\eta_u}{\eta_u}(D - R).$$

This implies

$$\langle \boldsymbol{v}_i^{T}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}_i^{0}, \boldsymbol{y}\rangle \geq \frac{\frac{1+\eta_u}{\eta_u}}{\frac{1+\eta_c}{\eta_c} + \frac{1+\eta_u}{\eta_u}}(D-R) = \frac{\eta_c(1+\eta_u)}{\eta_u(1+\eta_c) + \eta_c(1+\eta_u)}(D-R) \geq \frac{\eta_c}{\eta_u + \eta_c}(D-R).$$

given $\eta_c \leq \eta_u$. □

The following lemma shows that, when we *reflect* some of the feature vectors in a system $(U^t, V^t) = (\{\boldsymbol{u}_j^t\}_{j\in[m]}, \{\boldsymbol{v}_i^t\}_{i\in[n]})$, there is a correspondence between the behaviors of the system with the reflected vectors and the original system.

Lemma 7.7 (Reflection)**.** *Let $(U^t, V^t) = (\{\boldsymbol{u}_j^t\}_{j\in[m]}, \{\boldsymbol{v}_i^t\}_{i\in[n]})$ be a system of m users and n creators with impact functions f, g. Let $a_i, b_j \in \{+1, -1\}$, $\forall i \in [n]$, $\forall i \in [m]$ be some binary constants. Define:*

$$\tilde{\boldsymbol{u}}_j^t = b_j\boldsymbol{u}_j^t = \pm\boldsymbol{u}_j^t, \qquad \tilde{\boldsymbol{v}}_i^t = a_i\boldsymbol{v}_i^t = \pm\tilde{\boldsymbol{v}}_i^t.$$

and impact functions

$$\tilde{f}(\tilde{\boldsymbol{v}}_i, \tilde{\boldsymbol{u}}_j) = a_ib_jf(\boldsymbol{v}_i, \boldsymbol{u}_j), \qquad \tilde{g}(\tilde{\boldsymbol{u}}_j, \tilde{\boldsymbol{v}}_i) = a_ib_jg(\boldsymbol{u}_j, \boldsymbol{v}_i).$$

Then:

- *There is a "correspondence" between the evolution of the system (U^t, V^t) with impact functions f, g and the evolution of the system $(\tilde{U}^t, \tilde{V}^t) = (\{\tilde{\boldsymbol{u}}_j^t\}_{j\in[m]}, \{\tilde{\boldsymbol{v}}_i^t\}_{i\in[n]})$ with impact functions $\tilde{f}, \tilde{g}$. Formally, suppose every user is recommended the same creator in the two systems, then the updated vectors in the two systems still satisfy the relations: $\tilde{\boldsymbol{u}}_j^{t+1} = b_j\boldsymbol{u}_j^{t+1}$, $\tilde{\boldsymbol{v}}_i^{t+1} = a_i\boldsymbol{v}_i^{t+1}$.*
- *If the system $(\tilde{U}^t, \tilde{V}^t)$ is in R-bi-polarization, then the original system (U^t, V^t) is also in R-bi-polarization.*

Proof. Consider the first item. Suppose user i is recommended creator j at time step t in the two systems. Then by definition, the updated user vectors in the two systems satisfy

$$\begin{aligned}\tilde{\boldsymbol{u}}_j^{t+1} &= \mathcal{P}\big(\tilde{\boldsymbol{u}}_j^t + \eta_u\tilde{f}(\tilde{\boldsymbol{v}}_i^t, \tilde{\boldsymbol{u}}_j^t)\tilde{\boldsymbol{v}}_i^t\big) \;=\; \mathcal{P}\big(b_j\boldsymbol{u}_j^t + \eta_ua_ib_jf(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)a_i\boldsymbol{v}_i^t\big) \\ &= \mathcal{P}\big(b_j\boldsymbol{u}_j^t + \eta_ub_jf(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t\big) \;=\; b_j\mathcal{P}\big(\boldsymbol{u}_j^t + \eta_uf(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t\big) \;=\; b_j\boldsymbol{u}_j^{t+1}\end{aligned}$$

Suppose creator i is recommended to the set of users J at time step t in the two systems. Then,

$$\begin{aligned}
\tilde{\boldsymbol{v}}_i^{t+1} &= \mathcal{P}\big(\tilde{\boldsymbol{v}}_i^t + \tfrac{\eta_c}{|J|}\sum_{j\in J} g(\tilde{\boldsymbol{u}}_j^t, \tilde{\boldsymbol{v}}_i^t)\tilde{\boldsymbol{u}}_j^t\big) \\
&= \mathcal{P}\big(a_i\boldsymbol{v}_i^t + \tfrac{\eta_c}{|J|}\sum_{j\in J} a_i b_j g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) b_j \boldsymbol{u}_j^t\big) \\
&= \mathcal{P}\big(a_i\boldsymbol{v}_i^t + \tfrac{\eta_c}{|J|}\sum_{j\in J} a_i g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) \boldsymbol{u}_j^t\big) \\
&= a_i\mathcal{P}\big(\boldsymbol{v}_i^t + \tfrac{\eta_c}{|J|}\sum_{j\in J} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) \boldsymbol{u}_j^t\big) \;=\; a_i\boldsymbol{v}_i^{t+1}.
\end{aligned}$$

This means that the evolution of the system $(\tilde{U}^t, \tilde{V}^t)$ has a correspondence to the evolution of the original system (U^t, V^t).

Consider the second item. Suppose $(\tilde{U}^t, \tilde{V}^t)$ is in R-bi-polarization, so $\tilde{\boldsymbol{v}}_i^t = \pm\boldsymbol{v}_i^t$ is R-close to $\pm\boldsymbol{c}$ and $\tilde{\boldsymbol{u}}_j^t = \pm\boldsymbol{u}_j^t$ is R-close to $\pm\boldsymbol{c}$ with some vector $\boldsymbol{c}\in\mathbb{S}^{d-1}$. This implies that $\boldsymbol{v}_i^t$ is R-close to $\pm\boldsymbol{c}$ and $\boldsymbol{u}_j^t$ is R-close to $\pm\boldsymbol{c}$. So, the system (U^t, V^t) satisfies R-bi-polarization. □

7.9 Proof of Proposition 7.1

Proof. Let $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ be an $(R, \boldsymbol{c})$-bi-polarization state with $R\in[0,1]$ and $\boldsymbol{c}\in\mathbb{S}^{d-1}$, where all $\boldsymbol{u}_j^t$ and $\boldsymbol{v}_i^t$ are within distance R to $+\boldsymbol{c}$ or $-\boldsymbol{c}$. We show that, after one step of update, $\boldsymbol{u}_j^{t+1}$ and $\boldsymbol{v}_i^{t+1}$ are still within distance R to $+\boldsymbol{c}$ or $-\boldsymbol{c}$, so $(\boldsymbol{U}^{t+1}, \boldsymbol{V}^{t+1})$ still satisfies $(R, \boldsymbol{c})$-bi-polarization.

Consider $\boldsymbol{u}_j^t$. Without loss of generality, suppose $\boldsymbol{u}_j^t$ is close to $+\boldsymbol{c}$, so $\|\boldsymbol{u}_j^t - \boldsymbol{c}\|_2 \le R$. Suppose user j is recommended creator i at step t. Let $\tilde{\boldsymbol{v}}_i^t = \boldsymbol{v}_i^t$ if $\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t\rangle \ge 0$ and

$\tilde{\boldsymbol{v}}_i^t = -\boldsymbol{v}_i^t$ if $\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t \rangle < 0$. Then, the user update is

$$\boldsymbol{u}_j^{t+1} = \mathcal{P}\Big(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t\Big) = \mathcal{P}\Big(\boldsymbol{u}_j^t + \eta_u |f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)| \tilde{\boldsymbol{v}}_i^t\Big).$$

Since $\tilde{\boldsymbol{v}}_i^t$ is close to $+\boldsymbol{c}$ or $-\boldsymbol{c}$, $\langle \tilde{\boldsymbol{v}}_i^t, \boldsymbol{u}_j^t \rangle > 0$, and $\boldsymbol{u}_j^t$ is close to $+\boldsymbol{c}$, it must be that $\tilde{\boldsymbol{v}}_i^t$ is close to $+\boldsymbol{c}$, so $\|\tilde{\boldsymbol{v}}_i^t - \boldsymbol{c}\|_2 \le R$. Then, since $\boldsymbol{u}_j^{t+1}$ is the normalization of a vector in the convex cone formed by $\boldsymbol{u}_j^t$ and $\tilde{\boldsymbol{v}}_i^t$, by Lemma 7.2, we have

$$\|\boldsymbol{u}_j^{t+1} - \boldsymbol{c}\|_2 \ \le \ \max\big\{\|\boldsymbol{u}_j^t - \boldsymbol{c}\|_2, \ \|\tilde{\boldsymbol{v}}_i^t - \boldsymbol{c}\|_2\big\} \ \le \ R.$$

Consider $\boldsymbol{v}_i^t$. Suppose $\|\boldsymbol{v}_i^t - \boldsymbol{c}\|_2 \le R$. Let J be the set of users that are recommended creator i at step t. For each $j \in J$, let $\tilde{\boldsymbol{u}}_j^t = \boldsymbol{u}_j^t$ if $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle \ge 0$ and $\tilde{\boldsymbol{u}}_j^t = -\boldsymbol{u}_j^t$ if $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle < 0$. Then, the creator update is

$$\boldsymbol{v}_i^{t+1} = \mathcal{P}\Big(\boldsymbol{v}_i^t + \frac{\eta_c}{|J|}\sum_{j\in J} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t)\boldsymbol{u}_j^t\Big) = \mathcal{P}\Big(\boldsymbol{v}_i^t + \frac{\eta_c}{|J|}\sum_{j\in J} |g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t)| \tilde{\boldsymbol{u}}_j^t\Big).$$

We note that every $\tilde{\boldsymbol{u}}_j^t$ satisfies $\|\tilde{\boldsymbol{u}}_j^t - \boldsymbol{c}\|_2 \le R$ (by the same reasoning as above). Then, since $\boldsymbol{v}_i^{t+1}$ is the normalization of a vector in the convex cone formed by $\boldsymbol{v}_i^t$ and $\{\tilde{\boldsymbol{u}}_j^t\}_{j\in J}$, by Lemma 7.2, we have

$$\|\boldsymbol{v}_i^{t+1} - \boldsymbol{c}\|_2 \ \le \ \min\Big\{\|\boldsymbol{v}_i^t - \boldsymbol{c}\|_2, \ \min_{j\in J}\|\tilde{\boldsymbol{u}}_j^t - \boldsymbol{c}\|_2\Big\} \ \le \ R. \qquad \square$$

7.10 Proof of Lemma 7.1

Lemma 7.1 is proved by induction on the number n of creators. We first show that any system with 1 creator and multiple users must converge to R-bi-polarization in finite steps for any $R > 0$. Using the result for 1 creator, we then construct a finite length path that leads to R-bi-polarization for any system with $n \ge 2$ creators.

7.10.1 Base Case: Convergence Results for $n = 1$ Creator

We prove some convergence results for the special case of only one creator. This will serve as the basis for the proof for $n \geq 2$ creators. Recall that we have the following dynamics update rule:

- User: $\boldsymbol{u}_j^{t+1} = \mathcal{P}(\boldsymbol{u}_j^t + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)\boldsymbol{v}_i^t)$ where $\boldsymbol{v}_i^t$ is the creator recommended to user j; $f(\boldsymbol{v}_i, \boldsymbol{u}_j)$ satisfies:
$$f(\boldsymbol{v}_i, \boldsymbol{u}_j) \text{ is } \begin{cases} > 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle > 0 \\ < 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle < 0 \\ = 0 & \text{if } \langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle = 0. \end{cases} \tag{7.8}$$

- Creator: $\boldsymbol{v}_i^{t+1} = \mathcal{P}(\boldsymbol{v}_j^t + \frac{\eta_c}{|J|} \sum_{j \in J} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t)\boldsymbol{u}_j^t)$ where J is the set of users being recommended creator i.

Lemma 7.8. *Consider a system of* 1 *creator* $\boldsymbol{v}_i^t$ *and* $|J|$ *users* $\{\boldsymbol{u}_j^t\}_{j \in J}$*, where the creator is recommended to all users at every time step. Assume:*

- *Initially,* $\forall j \in J, \langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0 \rangle > 0$.
- *There exists some constant* $L_f > 0$ *such that* $f(\boldsymbol{v}_i, \boldsymbol{u}_j) \geq L_f > 0$ *whenever* $\langle \boldsymbol{v}_i, \boldsymbol{u}_j \rangle > 0$.
- $g(\boldsymbol{u}_j, \boldsymbol{v}_i) = 1$ *when* $\langle \boldsymbol{u}_j, \boldsymbol{v}_i \rangle > 0$.
- $\eta_c \leq \frac{\eta_u L_f}{2}$ *and* $0 \leq \eta_u < \frac{1}{2}$.

Then, for any $R > 0$*, after at most* $\frac{8}{3\eta_u L_f} \ln \frac{2|J|}{R^2}$ *steps,* $\sum_{j \in J} \|\boldsymbol{u}_j^t - \boldsymbol{v}_i^t\|_2^2 \leq R^2$ *will hold forever. In particular, each user vector will satisfy* $\|\boldsymbol{u}_j^t - \boldsymbol{v}_i^t\|_2 \leq R$.

Proof. We first note that, by Lemma 7.5, all user vectors satisfy $\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle > 0$ in all time steps $t > 0$. Hence, the creator update is always $\boldsymbol{v}_i^{t+1} = \mathcal{P}(\boldsymbol{v}_i^t + \frac{\eta_c}{|J|} \sum_{j \in J} g(\boldsymbol{u}_j^t, \boldsymbol{v}_i^t) \boldsymbol{u}_j^t) = \mathcal{P}(\boldsymbol{v}_i^t + \eta_c \frac{1}{|J|} \sum_{j \in J} \boldsymbol{u}_j^t)$.

Let $a_t = 1/(1 - \frac{3\eta_u L_f}{8})^t$. Define the following potential function:

$$\Phi^t \;=\; a_t \sum_{j \in J} \frac{1}{2} \|\boldsymbol{u}_j^t - \boldsymbol{v}_i^t\|_2^2 \;=\; a_t \sum_{j \in J} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big). \tag{7.9}$$

We will show that Φ^t is monotonically decreasing. Take the difference between Φ^{t+1} and Φ^t:

$$\begin{aligned}
\Phi^{t+1} - \Phi^t \;&=\; a_{t+1} \sum_{j \in J} \Big(\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle - \langle \boldsymbol{u}_j^{t+1}, \boldsymbol{v}_i^{t+1} \rangle \Big) \;+\; (a_{t+1} - a_t) \sum_{j \in J} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big) \\
&=\; a_{t+1} \Bigg(\sum_{j \in J} \langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t - \boldsymbol{u}_j^{t+1} \rangle + \sum_{j \in J} \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1} \rangle + \sum_{j \in J} \langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1} \rangle \Bigg) \\
&\qquad +\; (a_{t+1} - a_t) \sum_{j \in J} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big).
\end{aligned}$$

Using Lemma 7.3 with $\boldsymbol{x}^t = \boldsymbol{u}_j^t$, $\boldsymbol{z}^t = \boldsymbol{v}_i^t$, and $\eta = \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)$, we get

$$\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t - \boldsymbol{u}_j^{t+1} \rangle \;\le\; -\frac{\eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)}{1 + \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big) \;\le\; -\frac{\eta_u L_f}{2} \big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t \rangle\big).$$

Using Lemma 7.4 with $\boldsymbol{x}^t = \boldsymbol{u}_j^t$, $\boldsymbol{z}^t = \boldsymbol{v}_i^t$, and $\eta = \eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)$, we get

$$\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t - \boldsymbol{u}_j^{t+1} \rangle \;\le\; -\frac{1}{\eta_u f(\boldsymbol{v}_i^t, \boldsymbol{u}_j^t)} \|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2^2 \;\le\; -\frac{1}{\eta_u} \|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2^2.$$

Using Lemma 7.4 with $\boldsymbol{x}^t = \boldsymbol{v}_i^t$, $\boldsymbol{z}^t = \frac{1}{|J|} \sum_{j \in J} \boldsymbol{u}_j^t$, and $\eta = \eta_c$, we get

$$\sum_{j \in J} \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1} \rangle \;=\; |J| \langle \tfrac{1}{|J|} \sum_{j \in J} \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1} \rangle \;\le\; -\frac{|J|}{\eta_c} \|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2^2.$$

Using the above three inequalities, we can upper bound $\Phi^{t+1} - \Phi^t$:

$$\begin{aligned}
&\Phi^{t+1} - \Phi^t \\
&= \ a_{t+1}\Big(\frac{3}{4}\sum_{j\in J}\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t - \boldsymbol{u}_j^{t+1}\rangle \ + \ \frac{1}{4}\sum_{j\in J}\langle \boldsymbol{v}_i^t, \boldsymbol{u}_j^t - \boldsymbol{u}_j^{t+1}\rangle \\
&\qquad\quad + \sum_{j\in J}\langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1}\rangle \ + \ \sum_{j\in J}\langle \boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t, \boldsymbol{v}_i^t - \boldsymbol{v}_i^{t+1}\rangle\Big) \ + \ (a_{t+1} - a_t)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&\le \ a_{t+1}\Big(- \frac{3}{4}\sum_{j\in J}\frac{\eta_u L_f}{2}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \ - \ \frac{1}{4}\sum_{j\in J}\frac{1}{\eta_u}\|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2^2 \\
&\qquad\quad - \frac{|J|}{\eta_c}\|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2^2 \ + \ \sum_{j\in J}\|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2 \cdot \|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2\Big) \ + \ (a_{t+1} - a_t)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&= \ a_{t+1}\Big(- \frac{3\eta_u L_f}{8}\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&\qquad\quad - \sum_{j\in J}\Big(\underbrace{\frac{1}{4\eta_u}\|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2^2 + \frac{1}{\eta_c}\|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2^2}_{\ge 2\sqrt{\frac{1}{4\eta_u\eta_c}\|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2^2\|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2^2}} - \|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2 \cdot \|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2\Big)\Big) \\
&\qquad + \ (a_{t+1} - a_t)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&\le \ a_{t+1}\Big(- \frac{3\eta_u L_f}{8}\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \ - \ \sum_{j\in J}\Big(\underbrace{\sqrt{\tfrac{1}{\eta_u\eta_c}} - 1}_{\ge 0}\Big)\|\boldsymbol{u}_j^{t+1} - \boldsymbol{u}_j^t\|_2 \cdot \|\boldsymbol{v}_i^{t+1} - \boldsymbol{v}_i^t\|_2\Big) \\
&\qquad + \ (a_{t+1} - a_t)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&\le \ a_{t+1}\Big(- \frac{3\eta_u L_f}{8}\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \ + \ 0\Big) \ + \ (a_{t+1} - a_t)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&= \ \Big(\big(1 - \frac{3\eta_u L_f}{8}\big)a_{t+1} - a_t\Big)\sum_{j\in J}\big(1 - \langle \boldsymbol{u}_j^t, \boldsymbol{v}_i^t\rangle\big) \\
&= \ 0
\end{aligned}$$

where the last step is because $(1 - \frac{3\eta_u L_f}{8})a_{t+1} = a_t$.

We have shown that Φ^t is monotonically decreasing. Thus,

$$\frac{1}{2}\sum_{j\in J}\|\boldsymbol{u}_j^T - \boldsymbol{v}_i^T\|^2 = \frac{\Phi^T}{a_T} \le \frac{\Phi^0}{a_T} \le \frac{\sum_{j\in J} 1}{a_T} = \left(1-\tfrac{3\eta_u L_f}{8}\right)^T |J| \le e^{-\frac{3\eta_u L_f}{8}T}|J| \le \frac{1}{2}R^2$$

whenever $T \ge \frac{8}{3\eta_u L_f}\ln\frac{2|J|}{R^2}$. □

Corollary 7.1 (of Lemma 7.8)**.** *Consider a system of* 1 *creator* $\boldsymbol{v}_i^t$ *and* $|J|$ *users* $\{\boldsymbol{u}_j^t\}_{j\in J}$*, where the creator is recommended to all users at every time step. Assume:*

- *Initially,* $\langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0\rangle \ne 0$ *for every* $j \in J$*.*
- *There exists some constant* $L_f > 0$ *such that* $|f(\boldsymbol{v}_i, \boldsymbol{u}_j)| \ge L_f > 0$ *whenever* $\langle \boldsymbol{v}_i, \boldsymbol{u}_j\rangle \ne 0$*.*
- $g(\boldsymbol{u}_j, \boldsymbol{v}_i) = \text{sign}(\langle \boldsymbol{u}_j, \boldsymbol{v}_i\rangle)$*.*
- $\eta_c \le \frac{\eta_u L_f}{2}$ *and* $0 \le \eta_u < \frac{1}{2}$*.*

Then, for any $R > 0$*, after at most* $\frac{8}{3\eta_u L_f}\ln\frac{2|J|}{R^2}$ *steps, the system will reach* R*-bi-polarization.*

Proof. Let $J^+ = \{j \in J : \langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0\rangle > 0\}$ be the set of users with positive inner products with creator i initially; let $J^- = \{j \in J : \langle \boldsymbol{u}_j^0, \boldsymbol{v}_i^0\rangle < 0\}$. Let $\tilde{\boldsymbol{u}}_j^t = -\boldsymbol{u}_j^t$ for $j \in J^-$ and $\tilde{\boldsymbol{u}}_j^t = \boldsymbol{u}_j^t$ for $j \in J^+$. Then, the system consisting of $\{\tilde{\boldsymbol{u}}_j^t\}_{j\in J}$ and $\boldsymbol{v}_i^t$ satisfies the initial condition $\langle \tilde{\boldsymbol{u}}_j^0, \boldsymbol{v}_i^0\rangle > 0$ in Lemma 7.8. So, by Lemma 7.8, it reaches R-consensus after at most $\frac{8}{3\eta_u L_f}\ln\frac{2|J|}{R^2}$ steps. Then by the reflection lemma (Lemma 7.7), the original system, consisting of $\{\boldsymbol{u}_j^t\}_{j\in J}$ and $\boldsymbol{v}_i^t$, must reach R-bi-polarization. □

7.10.2 Inductive Step: Proof of Lemma 7.1

Lemma 7.9. *Consider a system of $n \geq 1$ creators $\{\boldsymbol{v}_1^t, \ldots, \boldsymbol{v}_n^t\}$ and $|J|$ users $\{\boldsymbol{u}_j^t\}_{j \in J}$. Assume:*

- *Initially, $\langle \boldsymbol{v}_i^0, \boldsymbol{v}_{i'}^0 \rangle > 0$ for every i, i', and $\langle \boldsymbol{v}_i^0, \boldsymbol{u}_j^0 \rangle > 0$ for every i, j.*
- *Assumptions of Lemma 7.8.*

Then, for any $R \in (0, 1)$, there exists a path of finite length that leads the initial state $(\boldsymbol{U}^0, \boldsymbol{V}^0)$ to R-consensus.

Proof. Fix any $R \in (0, 1)$. Choose R_1 such that $\sqrt{(\frac{\eta_u}{\eta_c} + 2)4R_1} = R$. Clearly, $R_1 < R$. We construct a path that leads the state $(\boldsymbol{U}^0, \boldsymbol{V}^0)$ to R-consensus as follows.

Step (1): Consider the subsystem of the first $n-1$ creators and all users J. By induction, there exists a path of length $T_1 = L_{n-1,R_1} < +\infty$ that leads the subsystem to $(R_1, \boldsymbol{c}^{T_1})$-consensus with some $\boldsymbol{c}^{T_1} \in \mathbb{S}^{d-1}$. So, after these T_1 steps, all creators $i \in \{1, \ldots, n-1\}$ and all users $j \in J$ satisfy $\|\boldsymbol{v}_i^{T_1} - \boldsymbol{c}^{T_1}\| \leq R_1$ and $\|\boldsymbol{u}_j^{T_1} - \boldsymbol{c}^{T_1}\| \leq R_1$. Creator n does not update during these T_1 steps, so $\boldsymbol{v}_n^{T_1} = \boldsymbol{v}_n^0$, and it still has positive inner products with the first $n-1$ creators and all users by the convex cone property (Lemma 7.2). Let's then consider the distance between creators n and the consensus center $\boldsymbol{c}^{T_1}$: $\|\boldsymbol{v}_n^{T_1} - \boldsymbol{c}^{T_1}\|$. If $\|\boldsymbol{v}_n^{T_1} - \boldsymbol{c}^{T_1}\| \leq R$, then the system has satisfied $(R, \boldsymbol{c}^{T_1})$-consensus, so our construction is finished. Otherwise, $\|\boldsymbol{v}_n^{T_1} - \boldsymbol{c}^{T_1}\| > R$. We continue the construction as follows:

Step (2): Pick any user $j_0 \in J$, recommend creator n to user j_0 for $T_2 = \frac{8}{3\eta_u L_f} \ln \frac{2}{R_1^2}$ steps, while recommending creator 1 to all other users. From the $(R_1, \boldsymbol{c}^{T_1})$-consensus in step (1) we know $\|\boldsymbol{u}_{j_0}^{T_1} - \boldsymbol{c}^{T_1}\| \leq R_1$, so

$$\langle \boldsymbol{u}_{j_0}^{T_1}, \boldsymbol{c}^{T_1} \rangle = 1 - \tfrac{1}{2}\|\boldsymbol{u}_{j_0}^{T_1} - \boldsymbol{c}^{T_1}\|^2 \geq 1 - \tfrac{R_1^2}{2} > 1 - \tfrac{R^2}{2} \geq 1 - \tfrac{1}{2}\|\boldsymbol{v}_n^{T_1} - \boldsymbol{c}^{T_1}\|^2 = \langle \boldsymbol{v}_2^{T_1}, \boldsymbol{c}^{T_1} \rangle.$$

Thus, we can apply Lemma 7.6 with $\boldsymbol{y} = \boldsymbol{c}^{T_1}$ to derive that, after these T_2 steps,

$$\begin{aligned} \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle &\geq \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\langle \boldsymbol{u}_{j_0}^{T_1}, \boldsymbol{c}^{T_1}\rangle - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle - R_1\Big) \\ &\geq \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle - R_1\Big). \\ \Longrightarrow \quad \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle &\geq \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle - R_1\Big). \qquad (7.10) \end{aligned}$$

For the inner product between creator n and user j_0, by Lemma 7.8 $\|\boldsymbol{v}_n^{T_1+T_2} - \boldsymbol{u}_{j_0}^{T_1+T_2}\| \leq R_1$, so

$$\langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{u}_{j_0}^{T_1+T_2}\rangle \;=\; 1 - \tfrac{1}{2}\|\boldsymbol{v}_n^{T_1+T_2} - \boldsymbol{u}_{j_0}^{T_1+T_2}\|^2 \;\geq\; 1 - \tfrac{R_1^2}{2}. \qquad (7.11)$$

Consider the inner products between creator n and the first $n-1$ creators and the users in $J \setminus \{j_0\}$. Because the first $n-1$ creators and the users in $J \setminus \{j_0\}$ form $(R_1, \boldsymbol{c}^{T_1})$-consensus at time step T_1, by Proposition 7.1, they still form $(R_1, \boldsymbol{c}^{T_1})$-consensus at time step $T_1 + T_2$, so $\|\boldsymbol{v}_i^{T_1+T_2} - \boldsymbol{c}^{T_1}\| \leq R_1$ and $\|\boldsymbol{u}_j^{T_1+T_2} - \boldsymbol{c}^{T_1}\| \leq R_1$. This implies, for $i \neq n$,

$$\begin{aligned} \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{v}_i^{T_1+T_2}\rangle &\geq \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle - \|\boldsymbol{v}_i^{T_1+T_2} - \boldsymbol{c}^{T_1}\| \\ &\geq \langle \boldsymbol{v}_2^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle - R_1 \\ \text{by (7.10)} \quad &\geq \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle - R_1\Big) - R_1, \qquad (7.12) \end{aligned}$$

and for $j \in J \setminus \{j_0\}$,

$$\begin{aligned} \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{u}_j^{T_1+T_2}\rangle &\geq \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle - \|\boldsymbol{u}_j^{T_1+T_2} - \boldsymbol{c}^{T_1}\| \\ &\geq \langle \boldsymbol{v}_2^{T_1+T_2}, \boldsymbol{c}^{T_1}\rangle - R_1 \\ \text{by (7.10)} \quad &\geq \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1}\rangle - R_1\Big) - R_1. \qquad (7.13) \end{aligned}$$

Step (3): Consider the subsystem of the first $n-1$ creators and all users J. By *induction, there exists a path of length* $T_3 = L_{n-1,R_1} < +\infty$ *that leads the subsystem to* $(R_1, \boldsymbol{c}^{T_1+T_2+T_3})$*-consensus with some* $\boldsymbol{c}^{T_1+T_2+T_3} \in \mathbb{S}^{d-1}$*.* So, we have $\|\boldsymbol{v}_i^{T_1+T_2+T_3} - \boldsymbol{c}^{T_1+T_2+T_3}\| \leq R_1$ for every $i \in \{1, \ldots, n-1\}$ and $\|\boldsymbol{u}_j^{T_1+T_2+T_3} - \boldsymbol{c}^{T_1+T_2+T_3}\| \leq R_1$ for every $j \in J$, and $\boldsymbol{v}_n^{T_1+T_2+T_3} = \boldsymbol{v}_n^{T_1+T_2}$. Consider the inner product between creator n and any of the first $n-1$ creators $i \in \{1, \ldots, n-1\}$. By the convex cone property (Lemma 7.2),

$$\begin{aligned}
\langle \boldsymbol{v}_n^{T_1+T_2+T_3}, \boldsymbol{v}_i^{T_1+T_2+T_3} \rangle &= \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{v}_i^{T_1+T_2+T_3} \rangle \\
\text{by Lemma 7.2} &\geq \min\left\{ \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{v}_i^{T_1+T_2} \rangle,\ \min_{j\in J} \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{u}_j^{T_1+T_2} \rangle \right\} \\
\text{by (7.11), (7.12), (7.13)} &\geq \min\left\{ \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle - R_1\Big) - R_1,\ 1 - \tfrac{R_1^2}{2} \right\} \\
&= \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle - R_1\Big) - R_1 \quad (7.14)
\end{aligned}$$

where the last equality is because, under the assumption of $\|\boldsymbol{v}_n^{T_1} - \boldsymbol{c}^{T_1}\| > R = \sqrt{(\frac{\eta_u}{\eta_c} + 2)4R_1}$,

$$\begin{aligned}
&\langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle - R_1\Big) - R_1 \\
&= \tfrac{\eta_u}{\eta_u+\eta_c}\langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - R_1\Big) - R_1 \\
&= \tfrac{\eta_u}{\eta_u+\eta_c}\Big(1 - \tfrac{1}{2}\|\boldsymbol{v}_2^{T_1} - \boldsymbol{c}^{T_1}\|^2\Big) + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - R_1\Big) - R_1 \\
&\leq \tfrac{\eta_u}{\eta_u+\eta_c}\Big(1 - \tfrac{1}{2}(\tfrac{\eta_u}{\eta_c} + 2)4R_1\Big) + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - R_1\Big) - R_1 \\
&\leq \max\{1 - \tfrac{1}{2}(\tfrac{\eta_u}{\eta_c} + 2)4R_1,\ 1 - \tfrac{R_1^2}{2} - R_1\} - R_1 \\
&= 1 - \tfrac{R_1^2}{2} - R_1 - R_1 \ \leq\ 1 - \tfrac{R_1^2}{2}.
\end{aligned}$$

From (7.14) and $\|\boldsymbol{v}_i^{T_1+T_2+T_3} - \boldsymbol{c}^{T_1+T_2+T_3}\| \leq R_1$,

$$\begin{aligned}
\langle \boldsymbol{v}_n^{T_1+T_2+T_3}, \boldsymbol{c}^{T_1+T_2+T_3} \rangle &\geq \langle \boldsymbol{v}_n^{T_1+T_2}, \boldsymbol{v}_i^{T_1+T_2+T_3} \rangle - \|\boldsymbol{v}_i^{T_1+T_2+T_3} - \boldsymbol{c}^{T_1+T_2+T_3}\| \\
&\geq \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \tfrac{R_1^2}{2} - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle - R_1\Big) - 2R_1.
\end{aligned}$$

Using 1 to minus the above inequality, we obtain

$$1 - \langle \boldsymbol{v}_n^{T_1+T_2+T_3}, \boldsymbol{c}^{T_1+T_2+T_3} \rangle \ \le \ \tfrac{\eta_u}{\eta_u+\eta_c}\Big(1 - \langle \boldsymbol{v}_n^{T_1}, \boldsymbol{c}^{T_1} \rangle\Big) + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1.$$

Let $F^t = 1 - \langle \boldsymbol{v}_n^t, \boldsymbol{c}^t \rangle$, then

$$F^{T_1+T_2+T_3} \le \tfrac{\eta_u}{\eta_u+\eta_c} F^{T_1} + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1. \tag{7.15}$$

Repeat steps (2) and (3) for K times. Then, using (7.15) for K times,

$$\begin{aligned}
&F^{T_1+K(T_2+T_3)} \\
&\le \ \tfrac{\eta_u}{\eta_u+\eta_c} F^{T_1+(K-1)(T_2+T_3)} + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1 \\
&\le \ \tfrac{\eta_u}{\eta_u+\eta_c}\bigg(\tfrac{\eta_u}{\eta_u+\eta_c} F^{T_1+(K-2)(T_2+T_3)} + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1\bigg) + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1 \\
&\ \vdots \\
&\le \ \big(\tfrac{\eta_u}{\eta_u+\eta_c}\big)^K F^{T_1} + \Big(1 + \tfrac{\eta_u}{\eta_t+\eta_c} + \cdots + \big(\tfrac{\eta_u}{\eta_t+\eta_c}\big)^{K-1}\Big)\Big(\tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1\Big) \\
&\le \ \big(\tfrac{\eta_u}{\eta_u+\eta_c}\big)^K \cdot 1 + \frac{1}{1 - \frac{\eta_u}{\eta_u+\eta_c}}\Big(\tfrac{\eta_c}{\eta_u+\eta_c}\Big(\tfrac{R_1^2}{2} + R_1\Big) + 2R_1\Big) \\
&= \ \big(\tfrac{\eta_u}{\eta_u+\eta_c}\big)^K + \tfrac{R_1^2}{2} + R_1 + \tfrac{\eta_u+\eta_c}{\eta_c} 2R_1 \\
&\le \ \tfrac{R_1^2}{2} + \tfrac{R_1^2}{2} + R_1 + \tfrac{\eta_u+\eta_c}{\eta_c} 2R_1 \ \le \ \big(\tfrac{\eta_u}{\eta_c} + 2\big)2R_1,
\end{aligned}$$

by choosing $K = \frac{\ln \frac{2}{R_1^2}}{\ln \frac{\eta_u+\eta_c}{\eta_u}} \le \frac{\eta_u+\eta_c}{\eta_c} \ln \frac{2}{R_1^2}$. This means that, after repeating steps (2) and (3) for K times, we must have

$$\begin{aligned}
\|\boldsymbol{v}_n^{T_1+K(T_2+T_3)} - \boldsymbol{c}^{T_1+K(T_2+T_3)}\| \ &= \ \sqrt{2\big(1 - \langle \boldsymbol{v}_n^{T_1+K(T_2+T_3)}, \boldsymbol{c}^{T_1+K(T_2+T_3)} \rangle\big)} \\
&= \ \sqrt{2F^{T_1+K(T_2+T_3)}} \ \le \ \sqrt{2\big(\tfrac{\eta_u}{\eta_c} + 2\big)2R_1} \ = \ R.
\end{aligned}$$

The above inequality, together with the fact that other creators $i \neq n$ and all users in J

already satisfy $(R_1 \leq R, \boldsymbol{c}^{T_1+K(T_2+T_3)})$-consensus after step (3), implies that the whole system has reached $(R, \boldsymbol{c}^{T_1+K(T_2+T_3)})$-consensus.

The length of the path constructed above is at most:

$$T_1 + K(T_2+T_3) \;\leq\; L_{n-1,R_1} + \tfrac{\eta_u+\eta_c}{\eta_c} \ln \tfrac{2}{R_1^2}\Big(\tfrac{8}{3\eta_u L_f} \ln \tfrac{2|J|}{R_1^2} + L_{n-1,R_1}\Big) \;=\; L_{n,R} \;<\; +\infty,$$

which is finite. □

Lemma 7.10. *Consider a subsystem of n creators $\{\boldsymbol{v}_1^t, \ldots, \boldsymbol{v}_n^t\}$ and $|J|$ users $\{\boldsymbol{u}_j^t\}_{j\in J}$. Assume:*

- *Initially, the first $n-1$ creators and all users are in R_0-consensus: $\|\boldsymbol{v}_i^0 - \boldsymbol{c}\| \leq R_0$, $\|\boldsymbol{u}_j^0 - \boldsymbol{c}\| \leq R_0$, with $0 < R_0 < \frac{\eta_c}{5(\eta_c+\eta_u)}$.*
- *$\langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0\rangle > 0$ for some $j_0 \in J$.*
- *$g(\boldsymbol{u}_j, \boldsymbol{v}_i) = \mathrm{sign}(\langle \boldsymbol{u}_j, \boldsymbol{v}_i\rangle)$.*
- *Assumption of Lemma 7.8.*

Then, for any $R \in (0,1)$, there exists a path of finite length that leads the initial state $(\boldsymbol{U}^0, \boldsymbol{V}^0)$ to R-consensus.

Proof. First, we recommend creator n to user j_0 for $T = \frac{8}{3\eta_u L_f} \ln \frac{2}{R_0^2}$ steps, while recommending other creators to other users arbitrarily. Applying Lemma 7.6 with $\boldsymbol{y} = \boldsymbol{u}_{j_0}^0$, we get

$$\begin{aligned}\langle \boldsymbol{v}_n^T, \boldsymbol{u}_{j_0}^0\rangle - \langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0\rangle \;&\geq\; \tfrac{\eta_c}{\eta_u+\eta_c}\Big(\langle \boldsymbol{u}_{j_0}^0, \boldsymbol{u}_{j_0}^0\rangle - \langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0\rangle - R_0\Big) \\ &=\; \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0\rangle - R_0\Big). \qquad (7.16)\end{aligned}$$

On the other hand, because the first $n-1$ creators and all users in $J \setminus \{j_0\}$ form an $(R_0, \boldsymbol{c})$-consensus at time step 0, according to Proposition 7.1, they still form an $(R_0, \boldsymbol{c})$-consensus at time step T, so $\|\boldsymbol{v}_i^T - \boldsymbol{c}\| \le R_0$ for every $i \in \{1, \ldots, n-1\}$. This implies, for every $i \in \{1, \ldots, n-1\}$,

$$\langle \boldsymbol{v}_n^T, \boldsymbol{v}_i^T \rangle - \langle \boldsymbol{v}_n^T, \boldsymbol{u}_{j_0}^0 \rangle \;\ge\; -\|\boldsymbol{v}_i^T - \boldsymbol{u}_{j_0}^0\| \;\ge\; -\|\boldsymbol{v}_i^T - \boldsymbol{c}\| - \|\boldsymbol{c} - \boldsymbol{u}_{j_0}^0\| \;\ge\; -2R_0. \tag{7.17}$$

Adding (7.16) and (7.17) and moving $\langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0 \rangle$ to the right side, we get

$$\begin{aligned} \langle \boldsymbol{v}_n^T, \boldsymbol{v}_i^T \rangle \;&\ge\; \langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0 \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\Big(1 - \langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0 \rangle - R_0\Big) - 2R_0 \\ &=\; \tfrac{\eta_u}{\eta_u+\eta_c}\langle \boldsymbol{v}_n^0, \boldsymbol{u}_{j_0}^0 \rangle + \tfrac{\eta_c}{\eta_u+\eta_c}\big(1 - R_0\big) - 2R_0 \\ &>\; 0 + \tfrac{\eta_c}{\eta_u+\eta_c}\big(1 - R_0\big) - 2R_0 \;>\; 0, \end{aligned}$$

under the condition of $R_0 < \frac{\eta_c}{5(\eta_u+\eta_c)}$. Moreover, for every $j \in J \setminus \{j_0\}$, because $\|\boldsymbol{u}_j^T - \boldsymbol{v}_i^T\| \le \|\boldsymbol{u}_j^T - \boldsymbol{c}\| + \|\boldsymbol{c} - \boldsymbol{v}_i^T\| \le 2R_0$,

$$\langle \boldsymbol{v}_n^T, \boldsymbol{u}_j^T \rangle \;\ge\; \langle \boldsymbol{v}_n^T, \boldsymbol{v}_i^T \rangle - \|\boldsymbol{u}_j^T - \boldsymbol{v}_i^T\| \;\ge\; \tfrac{\eta_c}{\eta_u+\eta_c}\big(1 - R_0\big) - 4R_0 > 0.$$

For j_0, by Lemma 7.8, $\|\boldsymbol{v}_n^T - \boldsymbol{u}_{j_0}^T\| \le R_0$, so

$$\langle \boldsymbol{v}_n^T, \boldsymbol{u}_{j_0}^T \rangle \;=\; 1 - \tfrac{1}{2}\|\boldsymbol{v}_n^T - \boldsymbol{u}_{j_0}^T\|^2 \;\ge\; 1 - \tfrac{R_0^2}{2} \;>\; 0.$$

For the inner product between any creator $i \in \{1, \ldots, n-1\}$ and the users:

$$\begin{aligned} \langle \boldsymbol{v}_i^T, \boldsymbol{u}_{j_0}^T \rangle \;\ge\; \langle \boldsymbol{v}_i^T, \boldsymbol{v}_n^T \rangle - \|\boldsymbol{v}_n^T - \boldsymbol{u}_{j_0}^T\| \;&\ge\; \tfrac{\eta_c}{\eta_u+\eta_c}\big(1 - R_0\big) - 2R_0 - R_0 \\ &=\; \tfrac{\eta_c}{\eta_u+\eta_c}\big(1 - R_0\big) - 3R_0 \;>\; 0; \end{aligned}$$

$$\forall j \in J \setminus \{j_0\}, \quad \langle \boldsymbol{v}_i^T, \boldsymbol{u}_j^T \rangle \ = \ 1 - \tfrac{1}{2}\|\boldsymbol{v}_i^T - \boldsymbol{u}_j^T\|^2 \ \geq \ 1 - \tfrac{1}{2}\big(\|\boldsymbol{v}_i^T - \boldsymbol{c}\| + \|\boldsymbol{c} - \boldsymbol{u}_j^T\|\big)^2$$
$$> \ 1 - \tfrac{1}{2}(2R_0)^2 \ > \ 0.$$

All of the "> 0" inequalities above show that the system of $\{\boldsymbol{v}_i^T\}_{i\in[n]}$ and $\{\boldsymbol{u}_j^T\}_{j\in J}$ satisfies the condition of Lemma 7.9. So, there exists a path of finite length $T_2 < +\infty$ that leads the system to R-consensus by Lemma 7.9. The total length of path $T + T_2 = \frac{8}{3\eta_u L_f} \ln \frac{2}{R_0^2} + T_2 < +\infty$ is finite. □

Lemma 7.1. *Suppose $\eta_c \leq \frac{\eta_u L_f}{2}$ and $\eta_u < \frac{1}{2}$. For any $R > 0$, for almost every state $(\boldsymbol{U}^t, \boldsymbol{V}^t)$ in the state space, there exists a path $(\boldsymbol{U}^t, \boldsymbol{V}^t) \to (\boldsymbol{U}^{t+1}, \boldsymbol{V}^{t+1}) \to \cdots \to (\boldsymbol{U}^{t+T}, \boldsymbol{V}^{t+T})$ of finite length that leads to an R-bi-polarization state $(\boldsymbol{U}^{t+T}, \boldsymbol{V}^{t+T})$.*

Proof. We prove this lemma by induction on the number of creators n. The case for $n = 1$ directly follows from Corollary 7.1 which shows that, for any system of $n = 1$ creator and $|J|$ users with no $\langle \boldsymbol{v}_i^0, \boldsymbol{u}_j^0 \rangle = 0$, there exists a path of length at most $L_1^R = \frac{8}{3\eta_u L_F} \ln \frac{2|J|}{R^2} < +\infty$ that leads to R-bi-polarization.

Consider $n \geq 2$. Consider the subsystem consisting of the first $n-1$ creators $\{\boldsymbol{v}_1^t, \ldots, \boldsymbol{v}_{n-1}^t\}$ and all users. Let $R_0 = \frac{\eta_c}{6(\eta_c + \eta_u)}$. By induction, there exists a path of finite length $T_1 = L_{n-1}^{R_0} < +\infty$ that leads the subsystem to R_0-bi-polarization, with some vector $\boldsymbol{c}_0 \in \mathbb{S}^{d-1}$, so every $\boldsymbol{v}_i^{T_1}$ is R_0-close to $+\boldsymbol{c}_0$ or $-\boldsymbol{c}_0$, for $i \neq n$, and every $\boldsymbol{u}_j^{T_1}$ is R_0-close to $+\boldsymbol{c}_0$ or $-\boldsymbol{c}_0$. Define:

$$\forall i \neq n, \quad \tilde{\boldsymbol{v}}_i^t = \begin{cases} \boldsymbol{v}_i^t & \text{if } \boldsymbol{v}_i^{T_1} \text{ is } R_0\text{-close to } +\boldsymbol{c} \\ -\boldsymbol{v}_i^t & \text{if } \boldsymbol{v}_i^{T_1} \text{ is } R_0\text{-close to } -\boldsymbol{c} \end{cases}$$
$$\forall j \in J, \quad \tilde{\boldsymbol{u}}_j^t = \begin{cases} \boldsymbol{u}_j^t & \text{if } \boldsymbol{u}_j^{T_1} \text{ is } R_0\text{-close to } +\boldsymbol{c} \\ -\boldsymbol{u}_j^t & \text{if } \boldsymbol{u}_j^{T_1} \text{ is } R_0\text{-close to } -\boldsymbol{c} \end{cases}.$$

By definition, we have

$$\|\tilde{\boldsymbol{v}}_i^{T_1} - \boldsymbol{c}_0\| \leq R_0, \quad \forall i \neq n, \qquad\qquad \|\tilde{\boldsymbol{u}}_j^{T_1} - \boldsymbol{c}_0\| \leq R_0, \quad \forall j \in J.$$

This means that $\{\tilde{\boldsymbol{v}}_i^{T_1}\}_{i \neq n}$ and $\{\tilde{\boldsymbol{u}}_j^{T_1}\}_{j \in J}$ form an $(R_0, \boldsymbol{c}_0)$-consensus. Consider creator n. Let

$$\tilde{\boldsymbol{v}}_n^t = \begin{cases} \boldsymbol{v}_n^t & \text{if } \langle \boldsymbol{v}_n^{T_1}, \tilde{\boldsymbol{u}}_{j_0}^{T_1} \rangle > 0 \text{ for some } j_0 \in J \\ -\boldsymbol{v}_n^t & \text{if } \langle \boldsymbol{v}_n^{T_1}, \tilde{\boldsymbol{u}}_j^{T_1} \rangle < 0 \text{ for all } j \in J. \end{cases}$$

(The case where $\langle \boldsymbol{v}_n^{T_1}, \tilde{\boldsymbol{u}}_j^{T_1} \rangle = 0$ for some $j \in J$ is ignored because the initial states that can lead to such states have measure 0.) By definition, we have

$$\langle \tilde{\boldsymbol{v}}_n^{T_1}, \tilde{\boldsymbol{u}}_{j_0}^{T_1} \rangle > 0 \text{ for some } j_0 \in J.$$

Note that, at time step T_1, the system consisting of $\{\tilde{\boldsymbol{v}}_i^{T_1}\}_{i \in [n]}$ and $\{\tilde{\boldsymbol{u}}_j^{T_1}\}_{j \in J}$ satisfies the condition of Lemma 7.10, so there exists a path of length $T_2 = \tilde{L}_n^R < +\infty$ that leads the system to R-consensus. Then by the reflection lemma (Lemma 7.7), the original system $\{\boldsymbol{v}_i^t\}_{i \in [n]}$, $\{\boldsymbol{u}_j^t\}_{j \in J}$ must reach R-bi-polarization. The total length of path that leads to this R-bi-polarization is $L_n^R = T_1 + T_2 = L_{n-1}^{R_0} + \tilde{L}_n^R < +\infty$. □

Part IV

Conclusion

Chapter 8

Conclusion

Incentive design lies at the heart of multi-agent systems, but traditional theories often rely on idealized assumptions — principals with complete knowledge and agents with unbounded rationality. This dissertation challenges those assumptions and proposes a new lens: understanding incentive design in the machine learning age, where both principals and agents learn from data and adapt through experience.

We explored three key directions. First, we studied *incentive design by learning principals*, who interact with agents whose belief systems or private types are unknown and possibly non-Bayesian. In this setting, we introduced learning algorithms that enable the principal to infer agents' subjective prior, quantify agents' behavioral biases, and coordinate agents. We offered theoretical guarantees on regret and sample complexity. This leads to robust information design and mechanism design that remain effective even in the face of epistemic uncertainty about the agent.

Second, we examined *incentive design for learning agents*, where the assumption of perfect rationality is replaced with no-regret learning algorithms. Our results establish a reduction from learning agents to approximately best-responding agents, enabling precise characterizations of the principal's utility in general principal-agent problems. We also provided a complete convergence analysis of multi-agent learning dynamics in first-price auctions with fixed values, making foundational progress towards a long-standing open question in the field.

Finally, we turned our attention to *incentive issues in real-world machine learning systems*. Using recommender systems as a case study, we demonstrated that the strategic behavior of content creators can counteract intended fairness or diversity goals and proposed alternative algorithmic interventions that are more robust to such behaviors. This highlights the importance of integrating incentive-aware design into deployed AI systems.

Future Outlook

This dissertation opens up, and offers tools and insights to, a rich array of future research directions.

A central theme is the *co-evolution of learning and incentives*: as agents and principals continue to rely on data-driven methods, we need a deeper theoretical understanding of how learning dynamics interact with strategic behavior in complex environments. In particular, extending our frameworks to settings with *partial observability*, *limited feedback*, or *strategic manipulation of the learning process itself* is a promising direction. For instance, what happens when agents deliberately game the principal's learning algorithm? How do we design incentive mechanisms that are robust to such strategic adaptation?

For information design specifically, a key direction is *to incorporate richer cognitive models into information design theory.* Human decision-makers often exhibit heuristics, memory constraints, or biases that go beyond simple parametric deviations from Bayesian updating. Building models of belief formation that integrate insights from behavioral economics, psychology, and cognitive science — and then designing incentives under such models — is a crucial step toward making information design truly human-centric.

In relation to the above, another direction is *language-based information design.* Recent advances in generative AI, particularly large language models (LLMs), open up exciting new possibilities for the study of information design. Historically, information design has treated information as abstract signals in mathematical models. Yet in real-world settings, much of communication occurs through language — a domain traditionally hard to analyze formally. The emergence of LLMs enables us to computationally model and experiment with language as a strategic medium, offering a powerful bridge between formal signaling theory and natural language communication. This integration has the potential to yield insights that are inaccessible through conventional models, and to open new frontiers in both information design and AI.

On the practical side, as AI systems are increasingly deployed in high-stakes domains such as healthcare, education, and social media, it becomes imperative to embed *incentive-aligned learning mechanisms* into their design. Whether in coordinating autonomous agents, regulating strategic human users, or training AI models to interact safely with humans, the insights from this dissertation can help bridge the gap between theoretical rigor and real-world deployment.

In sum, this dissertation lays the groundwork for an emerging research agenda at the intersection of machine learning, economics, and algorithmic game theory. It calls for a rethinking of incentive design — not as a static optimization problem — but as a dynamic, adaptive process shaped by learning and interaction. The machine learning age presents new challenges, but also unprecedented opportunities to build systems that are not only intelligent, but also strategically robust and socially responsible.

Bibliography

[AB09] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[AB10] Jean-Yves Audibert and Sébastien Bubeck. Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*, pages 2785–2836, 2010.

[AB23] Arpit Agarwal and William Brown. Online recommendations for agents with discounted adaptive preferences. *arXiv preprint arXiv:2302.06014*, 2023.

[ABB$^+$24] Gagan Aggarwal, Ashwinkumar Badanidiyuru, Santiago R. Balseiro, Kshipra Bhawalkar, Yuan Deng, Zhe Feng, Gagan Goel, Christopher Liaw, Haihao Lu, Mohammad Mahdian, Jieming Mao, Aranyak Mehta, Vahab Mirrokni, Renato Paes Leme, Andres Perlroth, Georgios Piliouras, Jon Schneider, Ariel Schvartzman, Balasubramanian Sivan, Kelly Spendlove, Yifeng Teng, Di Wang, Hanrui Zhang, Mingfei Zhao, Wennan Zhu, and Song Zuo. Auto-Bidding and Auctions in Online Advertising: A Survey. *ACM SIGecom Exchanges*, 22(1):159–183, June 2024.

[AC16a] Ricardo Alonso and Odilon Câmara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706, September 2016.

[AC16b] Ricardo Alonso and Odilon Câmara. Persuading Voters. *American Economic Review*, pages 3590–3605, November 2016.

[ACBFS02] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, pages 48–77, January 2002.

[ACFO13] Daron Acemoğlu, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. Opinion fluctuations and disagreement in social networks. *Mathematics of Operations Research*, 38(1):1–27, 2013.

[ACK$^+$19] Jacob D Abernethy, Rachel Cummings, Bhuvesh Kumar, Sam Taggart, and Jamie H Morgenstern. Learning Auctions with Robust Incentive Guarantees. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS'19, pages 11587–11597, 2019.

[ACS24] Eshwar Ram Arunachaleswaran, Natalie Collina, and Jon Schneider. Pareto-Optimal Algorithms for Learning in Games. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 490–510, New Haven CT USA, July 2024. ACM.

[AFT23] Shipra Agrawal, Yiding Feng, and Wei Tang. Dynamic Pricing and Learning with Bayesian Persuasion. In *Advances in Neural Information Processing Systems*, volume 36, pages 59273–59285. Curran Associates, Inc., 2023.

[AGS20] Guy Aridor, Duarte Goncalves, and Shan Sikdar. Deconstructing the filter bubble: User decision-making recommender systems. RecSys '20, page 82–91, New York, NY, USA, 2020. Association for Computing Machinery.

[AIL23] Jerry Anunrojwong, Krishnamurthy Iyer, and David Lingenbrink. Persuading Risk-Conscious Agents: A Geometric Approach. *Operations Research*, page opre.2023.2438, March 2023.

[AL15] Claudio Altafini and Gabriele Lini. Predictable dynamics of opinion forming for networks with antagonistic interactions. *IEEE Transactions on Automatic Control*, 60(2):342–357, 2015.

[AL18] Mohammad Akbarpour and Shengwu Li. Credible mechanisms. In *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, page 371, 2018.

[AMS95] Robert John Aumann, Michael Bahir Maschler, and Richard E. Stearns. *Repeated games with incomplete information.* MIT Press, Cambridge, 1995.

[ARS13] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning Prices for Repeated Auctions with Strategic Buyers. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 1169–1177, 2013.

[Aum74] Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, March 1974.

[AVGCU19] Enrique Areyan Viqueira, Amy Greenwald, Cyrus Cousins, and Eli Upfal. Learning simulation-based games from data. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1778–1780, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[AVWZ24] Krishna Acharya, Varun Vangala, Jingyan Wang, and Juba Ziani. Producers equilibria and dynamics in engagement-driven recommender systems. *arXiv preprint arXiv:2401.16641*, 2024.

[BBC$^+$24] Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion without a clue. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[BBS18] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The Design and Price of Information. *American Economic Review*, 108(1):1–48, January 2018.

[BBX18] Ashwinkumar Badanidiyuru, Kshipra Bhawalkar, and Haifeng Xu. Targeting and Signaling in Ad Auctions. In *Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Proceedings, pages 2545–2563. Society for Industrial and Applied Mathematics, January 2018.

[BCB$^+$21] Jesse D. Bloom, Yujia Alina Chan, Ralph S. Baric, Pamela J. Bjorkman, Sarah Cobey, Benjamin E. Deverman, David N. Fisman, Ravindra Gupta, Akiko Iwasaki, Marc Lipsitch, Ruslan Medzhitov, Richard A. Neher, Rasmus Nielsen, Nick Patterson, Tim Stearns, Erik Van Nimwegen, Michael Worobey, and David A. Relman. Investigate the origins of COVID-19. *Science*, 372(6543):694–694, May 2021.

[BCD20] Johannes Brustle, Yang Cai, and Constantinos Daskalakis. Multi-Item Mechanisms without Item-Independence: Learnability via Robustness. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 715–761, Virtual Event Hungary, July 2020. ACM.

[BCIV23] Christian Borgs, Jennifer Chayes, Christian Ikeokwu, and Ellen Vitercik. Bursting the Filter Bubble: Disincentivizing Echo Chambers in Social Networks. In *Proceedings of EAAMO'23: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023.

[BCM$^+$22] Francesco Bacchiocchi, Matteo Castiglioni, Alberto Marchesi, Giulia Romano, and Nicola Gatti. Public Signaling in Bayesian Ad Auctions. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 39–45. International Joint Conferences on Artificial Intelligence Organization, July 2022.

[BDO24] Martin Bichler, Julius Durmann, and Matthias Oberlechner. Online Optimization Algorithms in Repeated Price Competition: Equilibrium Learning and Algorithmic Collusion, 2024. _eprint: 2412.15707.

[Ben19] Daniel J Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2:69–186, 2019.

[BFG23] Ashwinkumar Badanidiyuru, Zhe Feng, and Guru Guruganesh. Learning to Bid in Contextual First Price Auctions. In *Proceedings of the ACM Web Conference 2023*, pages 3489–3497, Austin TX USA, April 2023. ACM.

[BFH$^+$21] Martin Bichler, Maximilian Fichtl, Stefan Heidekrüger, Nils Kohring, and Paul Sutterer. Learning equilibria in symmetric auction games using artificial neural networks. *Nature Machine Intelligence*, 3(8):687–695, August 2021.

[BG19] Nikhil Bansal and Anupam Gupta. Potential-Function Proofs for Gradient Methods. *Theory of Computing*, 15(4):1–32, 2019. Publisher: Theory of Computing.

[BGM$^+$19] Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, and Jon Schneider. Contextual Bandits with Cross-Learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[BH05] Avrim Blum and Jason D. Hartline. Near-Optimal Online Auctions. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '05, pages 1156–1163. Society for Industrial and Applied Mathematics, 2005.

[BLO$^+$25] Martin Bichler, Stephan B. Lunowa, Matthias Oberlechner, Fabian R. Pieroth, and Barbara Wohlmuth. Beyond Monotonicity: On the Convergence of Learning Algorithms in Standard Auction Games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):13649–13657, April 2025.

[BM07] Avrim Blum and Yishay Mansour. From External to Internal Regret. *Journal of Machine Learning Research*, pages 1307–1324, 2007.

[BM16] Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games: Bayes correlated equilibrium. *Theoretical Economics*, 11(2):487–522, May 2016.

[BMSW18] Mark Braverman, Jieming Mao, Jon Schneider, and Matt Weinberg. Selling to a No-Regret Buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 523–538, Ithaca NY USA, June 2018. ACM.

[BPT18] Omer Ben-Porat and Moshe Tennenholtz. A game-theoretic approach to recommendation systems with strategic content providers. *Advances in Neural Information Processing Systems*, 31, 2018.

[Bro51] George W. Brown. Iterative Solution of Games by Fictitious Play. In *Activity Analysis of Production and Allocation*. Wiley, New York, 1951.

[BS22] Martino Banchio and Andrzej Skrzypacz. Artificial Intelligence and Auction Design. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 30–31, Boulder CO USA, July 2022. ACM.

[BSC$^+$24] Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Markov persuasion processes: Learning to persuade from scratch. *arXiv preprint arXiv:2402.03077*, 2024.

[BSV16] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Sample complexity of automated mechanism design. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2083–2091, 2016.

[BSV18] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. A general theory of sample complexity for multi-item profit maximization. In *Proceedings of the 2018 ACM Conference on Economics and Computation, Ithaca, NY, USA, June 18-22, 2018*, pages 173–174, 2018.

[BSV19] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Estimating approximate incentive compatibility. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, page 867. ACM, 2019.

[BTCXZ21] Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Regret-Minimizing Bayesian Persuasion. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 128–128, Budapest Hungary, July 2021. ACM.

[BTCXZ24] Yakov Babichenko, Inbal Talgam-Cohen, Haifeng Xu, and Konstantin Zabarnyi. Algorithmic Cheap Talk. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 5–6, New Haven CT USA, July 2024. ACM.

[BV06] Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 36–42, Ann Arbor Michigan USA, June 2006. ACM.

[BV18] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 104–112, Vancouver British Columbia Canada, September 2018. ACM.

[Cam98] Colin Camerer. Bounded rationality in individual decision making. *Experimental economics*, 1:163–183, 1998.

[CAS16] Paul Covington, Jay Adams, and Emre Sargin. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 191–198, Boston Massachusetts USA, September 2016. ACM.

[CBGM15] Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret Minimization for Reserve Prices in Second-Price Auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, January 2015.

[CBL06] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games.* Cambridge University Press, Cambridge, 2006.

[CCG20] Matteo Castiglioni, Andrea Celli, and Nicola Gatti. Persuading Voters: It's Easy to Whisper, It's Hard to Speak Loud. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1870–1877, April 2020.

[CCLP15] Marco Caponigro, Anna Chiara Lai, and Benedetto Piccoli. A nonlinear model of opinion formation on the sphere. *Discrete & Continuous Dynamical Systems - A*, 35(9):4241–4268, 2015.

[CCMG20] Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online Bayesian Persuasion. In *Advances in Neural Information Processing Systems*, pages 16188–16198. Curran Associates, Inc., 2020.

[CG98] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery.

[CHJ20] Modibo K. Camara, Jason D. Hartline, and Aleck Johnsen. Mechanisms for a No-Regret Agent: Beyond the Common Prior. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–270, Durham, NC, USA, November 2020. IEEE.

[CHN17] Shuchi Chawla, Jason D. Hartline, and Denis Nekipelov. Mechanism redesign. *CoRR*, abs/1708.04699, 2017.

[CHS12] Shuchi Chawla, Jason D. Hartline, and Balasubramanian Sivan. Optimal crowdsourcing contests. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 856–868, 2012.

[CLGB24] Qinyi Chen, Jason Cheuk Nam Liang, Negin Golrezaei, and Djallel Bouneffouf. Interpolating Item and User Fairness in Multi-Sided Recommendations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[CLP$^+$24] Yiling Chen, Tao Lin, Ariel D. Procaccia, Aaditya Ramdas, and Itai Shapira. Bias Detection via Signaling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[CMCG21] Matteo Castiglioni, Alberto Marchesi, Andrea Celli, and Nicola Gatti. Multi-Receiver Online Bayesian Persuasion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1314–1323. PMLR, July 2021.

[COZ22] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022.

[CP23] Xi Chen and Binghui Peng. Complexity of Equilibria in First-Price Auctions under General Tie-Breaking Rules. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 698–709, Orlando FL USA, June 2023. ACM.

[CR14] Richard Cole and Tim Roughgarden. The sample complexity of revenue maximization. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 243–252, 2014.

[CS82] Vincent P. Crawford and Joel Sobel. Strategic Information Transmission. *Econometrica*, page 1431, November 1982.

[CWD$^+$23] Yurong Chen, Qian Wang, Zhijian Duan, Haoran Sun, Zhaohua Chen, Xiang Yan, and Xiaotie Deng. Coordinated Dynamic Bidding in Repeated Second-Price Auctions with Budgets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5052–5086. PMLR, July 2023.

[CWM$^+$17] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. Learning to recommend accurate and diverse items. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 183–192, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[CWWZ24] Linda Cai, S. Matthew Weinberg, Evan Wildenhain, and Shirley Zhang. Selling to Multiple No-Regret Buyers. In *Web and Internet Economics*, pages 113–129, Cham, 2024. Springer Nature Switzerland.

[D'A23] Maurizio D'Andrea. Playing against no-regret players. *Operations Research Letters*, 51(2):142–145, March 2023.

[dCZ22] Geoffroy de Clippel and Xu Zhang. Non-Bayesian Persuasion. *Journal of Political Economy*, pages 2594–2642, October 2022.

[DFP$^+$10] Constantinos Daskalakis, Rafael Frongillo, Christos H. Papadimitriou, George Pierrakos, and Gregory Valiant. On Learning Algorithms for Nash Equilibria. In *Algorithmic Game Theory*, pages 114–125. Springer Berlin Heidelberg, 2010.

[DGP20] Francesco Decarolis, Maris Goldmanis, and Antonio Penta. Marketing Agencies and Collusive Bidding in Online Ad Auctions. *Management Science*, 66(10):4433–4454, October 2020.

[DGPS23] Francesco Decarolis, Maris Goldmanis, Antonio Penta, and Ksenia Shakhgildyan. Bid Coordination in Sponsored Search Auctions: Detection Methodology and Empirical Analysis. *The Journal of Industrial Economics*, 71(2):570–592, June 2023.

[DHLZ22] Xiaotie Deng, Xinyan Hu, Tao Lin, and Weiqiang Zheng. Nash Convergence of Mean-Based Learning Algorithms in First Price Auctions. In *Proceedings of the ACM Web Conference 2022*, pages 141–150, Virtual Event, Lyon France, April 2022. ACM.

[DHP16] Nikhil R. Devanur, Zhiyi Huang, and Christos-Alexandros Psomas. The sample complexity of auctions with side information. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 426–439, 2016.

[DHZ$^+$23] Zhijian Duan, Wenhan Huang, Dinghuai Zhang, Yali Du, Jun Wang, Yaodong Yang, and Xiaotie Deng. Is Nash Equilibrium Approximator Learnable? In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 233–241, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[DLL$^+$20] Xiaotie Deng, Ron Lavi, Tao Lin, Qi Qi, Wenwei WANG, and Xiang Yan. A Game-Theoretic Analysis of the Empirical Revenue Maximization Algorithm with Endogenous Sampling. In *Advances in Neural Information Processing Systems*, volume 33, pages 5215–5226, 2020.

[DM22] Sarah Dean and Jamie Morgenstern. Preference Dynamics Under Personalized Recommendations. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 795–816, Boulder CO USA, July 2022. ACM.

[DP18] Constantinos Daskalakis and Ioannis Panageas. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. 2018.

[DP22] Piotr Dworczak and Alessandro Pavan. Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion. *Econometrica*, pages 2017–2051, 2022.

[DPS15] Nikhil R. Devanur, Yuval Peres, and Balasubramanian Sivan. Perfect Bayesian Equilibria in Repeated Sales. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 983–1002, USA, 2015. Society for Industrial and Applied Mathematics.

[DSS19] Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against No-regret Learners. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[DX16] Shaddin Dughmi and Haifeng Xu. Algorithmic Bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, Cambridge MA USA, June 2016. ACM.

[EFG$^+$14] Yuval Emek, Michal Feldman, Iftah Gamzu, Renato PaesLeme, and Moshe Tennenholtz. Signaling Schemes for Revenue Maximization. *ACM Transactions on Economics and Computation*, pages 1–19, June 2014.

[EMR09] Guillaume Escamocher, Peter Bro Miltersen, and Santillan-Rodriguez Rocio. Existence and Computation of Equilibria of First-Price Auctions with Integral Valuations and Bids. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 1227–1228, 2009.

[ENS10] Larry G Epstein, Jawwad Noor, and Alvaro Sandroni. Non-Bayesian Learning. *The B.E. Journal of Theoretical Economics*, 10(1), January 2010.

[EO07] Benjamin Edelman and Michael Ostrovsky. Strategic bidder behavior in sponsored search auctions. *Decision Support Systems*, 43(1):192–198, February 2007.

[ER23] Itay Eilat and Nir Rosenfeld. Performative Recommendation: Diversifying Content via Strategic Incentives, June 2023. arXiv:2302.04336 [cs].

[FG03] Gadi Fibich and Arieh Gavious. Asymmetric First-Price Auctions: A Perturbation Approach. *Mathematics of Operations Research*, 28(4):836–852, 2003.

[FGL$^+$21] Zhe Feng, Guru Guruganesh, Christopher Liaw, Aranyak Mehta, and Abhishek Sethi. Convergence Analysis of No-Regret Bidding Algorithms in Repeated Auctions. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, 2021.

[FHT24] Yiding Feng, Chien-Ju Ho, and Wei Tang. Rationality-robust information design: Bayesian persuasion under quantal response. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 501–546. SIAM, 2024.

[FL98] Drew Fudenberg and David K. Levine. *The theory of learning in games*. MIT Press series on economic learning and social evolution. MIT Press, Cambridge, Mass, 1998.

[FL20] Hu Fu and Tao Lin. Learning Utilities and Equilibria in Non-Truthful Auctions. In *Advances in Neural Information Processing Systems*, volume 33, pages 14231–14242. Curran Associates, Inc., 2020.

[For06] Françoise Forges. Correlated Equilibrium in Games with Incomplete Information Revisited. *Theory and Decision*, 61(4):329–344, December 2006.

[FPS18] Zhe Feng, Chara Podimata, and Vasilis Syrgkanis. Learning to Bid Without Knowing your Value. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 505–522, Ithaca NY USA, June 2018. ACM.

[FRGH$^+$21] Aris Filos-Ratsikas, Yiannis Giannakopoulos, Alexandros Hollender, Philip Lazos, and Diogo Poças. On the Complexity of Equilibrium Computation in First-Price Auctions. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 454–476, Budapest Hungary, July 2021. ACM.

[FRGHK24] Aris Filos-Ratsikas, Yiannis Giannakopoulos, Alexandros Hollender, and Charalampos Kokkalis. On the Computation of Equilibria in Discrete First-Price Auctions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 379–399, New Haven CT USA, July 2024. ACM.

[FTX22] Yiding Feng, Wei Tang, and Haifeng Xu. Online Bayesian Recommendation with No Regret. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 818–819, Boulder CO USA, July 2022. ACM.

[Fuj23] Kaito Fujii. Bayes correlated equilibria and no-regret dynamics, 2023. _eprint: 2304.05005.

[FV97] Dean P. Foster and Rakesh V. Vohra. Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior*, 21(1-2):40–55, October 1997.

[FYC22] Shi Feng, Fang-Yi Yu, and Yiling Chen. Peer Prediction for Learning Agents. In *Advances in Neural Information Processing Systems*, 2022.

[GHTZ21] Chenghao Guo, Zhiyi Huang, Zhihao Gavin Tang, and Xinzhi Zhang. Generalizing complex hypotheses on product distributions: Auctions, prophet inequalities, and pandora's problem. In *Conference on Learning Theory*, pages 2248–2288. PMLR, 2021.

[GHWX23] Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Robust Stackelberg Equilibria. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 735–735, London United Kingdom, July 2023. ACM.

[GHWX24] Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Generalized principal-agency: Contracts, information, games and beyond. In *Proceedings of the 20th Conference on Web and Internet Economics*, WINE '24, 2024.

[GHZ19] Chenghao Guo, Zhiyi Huang, and Xinzhi Zhang. Settling the sample complexity of single-parameter revenue maximization. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 662–673, 2019.

[GJ10] Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49, February 2010.

[GJM21] Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic Incentive-Aware Learning: Robust Pricing in Contextual Auctions. *Operations Research*, 69(1):297–314, January 2021.

[GKS$^+$24] Guru Guruganesh, Yoav Kolumbus, Jon Schneider, Inbal Talgam-Cohen, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Joshua Ruizhi Wang, and S. Matthew Weinberg. Contracting with a Learning Agent. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[GLZ$^+$22] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3953–3957, New York, NY, USA, 2022. Association for Computing Machinery.

[GM87] Daniel A. Graham and Robert C. Marshall. Collusive Bidder Behavior at Single-Object Second-Price and English Auctions. *Journal of Political Economy*, 95(6):1217–1239, 1987.

[GN17] Yannai A. Gonczarowski and Noam Nisan. Efficient empirical revenue maximization in single-parameter auction environments. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 856–868. ACM, 2017.

[GR08] Wayne-Roy Gayle and Jean Francois Richard. Numerical solutions of asymmetric, first-price, independent private values auctions. *Computational Economics*, 32(3):245–278, Oct 2008.

[GS19] Yingni Guo and Eran Shmaya. The Interval Structure of Optimal Disclosure. *Econometrica*, 87(2):653–675, March 2019.

[GW18] Yannai A. Gonczarowski and S. Matthew Weinberg. The sample complexity of up-to-ϵ multi-dimensional revenue maximization. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 416–426, 2018.

[GWMS22] Shumpei Goke, Gabriel Y. Weintraub, Ralph A. Mastromonaco, and Samuel S. Seljan. Bidders' Responses to Auction Format Change in Internet Display Advertising Auctions. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 295–295, Boulder CO USA, July 2022. ACM.

[HHG$^+$13] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, San Francisco California USA, October 2013. ACM.

[HILS25] Keegan Harris, Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. Algorithmic persuasion through simulation, 2025.

[HJMD22] Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. *Advances in Neural Information Processing Systems*, 35:22969–22981, 2022.

[HK15] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[HKJ$^+$23] Jiri Hron, Karl Krauth, Michael I. Jordan, Niki Kilbertus, and Sarah Dean. Modeling Content Creator Incentives on Algorithm-Curated Platforms, July 2023. arXiv:2206.13102 [cs, stat].

[HL17] David Hagmann and George Loewenstein. Persuasion with motivated beliefs. Carnegie Mellon University technical report, 2017.

[HLW18] Zhiyi Huang, Jinyan Liu, and Xiangning Wang. Learning Optimal Reserve Price against Non-Myopic Bidders. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 2042–2052, 2018.

[HMC00] Sergiu Hart and Andreu Mas-Colell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, pages 1127–1150, September 2000.

[HMCG24] Safwan Hossain, Andjela Mladenovic, Yiling Chen, and Gauthier Gidel. A Persuasive Approach to Combating Misinformation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18926–18943. PMLR, July 2024.

[HPT08] Ken Hendricks, Robert Porter, and Guofu Tan. Bidding Rings and the Winner's Curse. *The RAND Journal of Economics*, 39(4):1018–1041, 2008.

[HSMS98] Shlomit Hon-Snir, Dov Monderer, and Aner Sela. A Learning Approach to Auctions. *Journal of Economic Theory*, 82(1):65–88, September 1998.

[HT19] Jason D. Hartline and Samuel Taggart. Sample complexity for non-truthful mechanisms. In *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 399–416, 2019.

[Hur13] Neil J. Hurley. Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 379–382, New York, NY, USA, 2013. Association for Computing Machinery.

[HWZ25] Yanjun Han, Tsachy Weissman, and Zhengyuan Zhou. Optimal No-Regret Learning in Repeated First-Price Auctions. *Operations Research*, 73(1):209–238, January 2025.

[IJS14] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. Mean Field Equilibria of Dynamic Auctions with Learning. *Management Science*, 60(12):2949–2970, December 2014.

[ILPT17] Nicole Immorlica, Brendan Lucier, Emmanouil Pountourakis, and Samuel Taggart. Repeated Sales with Multiple Strategic Buyers. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 167–168. ACM, June 2017.

[IMSW20] Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. Incentivizing exploration with selective data disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, page 647–648, New York, NY, USA, 2020. Association for Computing Machinery.

[Ito20] Shinji Ito. A Tight Lower Bound and Efficient Reduction for Swap Regret. In *Advances in Neural Information Processing Systems*, pages 18550–18559. Curran Associates, Inc., 2020.

[JCL$^+$19] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390, Honolulu HI USA, January 2019. ACM.

[JGS24] Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. Supply-side equilibria in recommender systems. *Advances in Neural Information Processing Systems*, 36, 2024.

[JMvE18] Natali Helberger Judith Möller, Damian Trilling and Bram van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.

[JP24] Atulya Jain and Vianney Perchet. Calibrated Forecasting and Persuasion. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 489–489, New Haven CT USA, July 2024. ACM.

[Kah13] Dan M. Kahan. Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, 8(4):407–424, July 2013.

[KBKW21] Dimitris Kalimeris, Smriti Bhagat, Shankar Kalyanaraman, and Udi Weinsberg. Preference amplification in recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 805–815, New York, NY, USA, 2021. Association for Computing Machinery.

[KG11] Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, pages 2590–2615, October 2011.

[KMZL17] Anton Kolotilin, Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li. Persuasion of a Privately Informed Receiver. *Econometrica*, 85(6):1949–1964, 2017.

[KN19] Yash Kanoria and Hamid Nazerzadeh. Incentive-Compatible Learning of Reserve Prices for Repeated Auctions. In *Companion The 2019 World Wide Web Conference*, 2019.

[KN22] Yoav Kolumbus and Noam Nisan. Auctions between Regret-Minimizing Agents. In *Proceedings of the ACM Web Conference 2022*, pages 100–111, Virtual Event, Lyon France, April 2022. ACM.

[Kos22] Svetlana Kosterina. Persuasion with unknown beliefs. *Theoretical Economics*, pages 1075–1107, 2022.

[KPW$^+$12] Dan M. Kahan, Ellen Peters, Maggie Wittlin, Paul Slovic, Lisa Larrimore Ouellette, Donald Braman, and Gregory Mandel. The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, 2(10):732–735, October 2012.

[KSLK20] Orcun Karaca, Pier Giuseppe Sessa, Anna Leidi, and Maryam Kamgarpour. No-regret learning from partially observed data in repeated auctions. *IFAC-PapersOnLine*, 53(2):14–19, 2020.

[KSS24] Rachitesh Kumar, Jon Schneider, and Balasubramanian Sivan. Strategically-Robust Learning Algorithms for Bidding in First-Price Auctions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 893–893, New Haven CT USA, July 2024. ACM.

[KV12] B. H. Korte and Jens Vygen. *Combinatorial optimization: theory and algorithms*. Algorithms and combinatorics. Springer, Heidelberg ; New York, 5th ed edition, 2012.

[LC25] Tao Lin and Yiling Chen. Generalized Principal-Agent Problem with a Learning Agent. In *The Thirteenth International Conference on Learning Representations*, 2025.

[LCL+23] Dugang Liu, Pengxiang Cheng, Zinan Lin, Xiaolian Zhang, Zhenhua Dong, Rui Zhang, Xiuqiang He, Weike Pan, and Zhong Ming. Bounding System-Induced Biases in Recommender Systems with a Randomized Dataset. *ACM Transactions on Information Systems*, 41(4):1–26, October 2023.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery.

[Leb96] Bernard Lebrun. Existence of an Equilibrium in First Price Auctions. *Economic Theory*, 7(3):421–443, 1996.

[Leb99] Bernard Lebrun. First Price Auctions in the Asymmetric N Bidder Case. *International Economic Review*, 40(1):125–142, February 1999.

[LJE+24] Tao Lin, Kun Jin, Andrew Estornell, Xiaoying Zhang, Yiling Chen, and Yang Liu. User-Creator Feature Polarization in Recommender Systems with Dual Influence. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[LL25] Tao Lin and Ce Li. Information Design with Unknown Prior. In *Proceedings of Innovations in Theoretical Computer Science*, 2025.

[LLZW23] Yue Lin, Wenhao Li, Hongyuan Zha, and Baoxiang Wang. Information design in multi-agent reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[LMS11] Giuseppe Lopomo, Leslie M. Marx, and Peng Sun. Bidder collusion at first-price auctions. *Review of Economic Design*, 15(3):177–211, September 2011.

[LR22] Sagi Levanon and Nir Rosenfeld. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*, pages 12593–12618. PMLR, 2022.

[LS14] Wei Li and Mark W. Spong. Unified cooperative control of multiple agents on a sphere for different spherical patterns. *IEEE Transactions on Automatic Control*, 59(5):1283–1289, 2014.

[Man11] Yishay Mansour. Lecture notes on lower bounds using information theory tools, 2011.

[MM92] R. Preston McAfee and John McMillan. Bidding Rings. *The American Economic Review*, 82(3):579–599, 1992.

[MM07] Robert C. Marshall and Leslie M. Marx. Bidder collusion. *Journal of Economic Theory*, 133(1):374–402, March 2007.

[MM14] Mehryar Mohri and Andres Muñoz Medina. Optimal Regret Minimization in Posted-Price Auctions with Strategic Buyers. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, 2014.

[MMRS94] Robert Marshall, Michael Meurer, Jean-Francois Richard, and Walter Stromquist. Numerical analysis of asymmetric first price auctions. *Games and Economic Behavior*, 7(2):193 – 220, 1994.

[MMSS22] Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against Learners in Bayesian Games. In *Proceedings of Thirty Fifth Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 5221–5252. PMLR, July 2022.

[MP95] Richard D. McKelvey and Thomas R. Palfrey. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, pages 6–38, July 1995.

[MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in Adversarial Regularized Learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2703–2717, USA, 2018. Society for Industrial and Applied Mathematics.

[MR00] Eric Maskin and John Riley. Equilibrium in Sealed High Bid Auctions. *Review of Economic Studies*, 67(3):439–454, July 2000.

[MR15] Jamie Morgenstern and Tim Roughgarden. On the pseudo-dimension of nearly optimal auctions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 136–144, 2015.

[MR16] Jamie Morgenstern and Tim Roughgarden. Learning simple auctions. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1298–1318, 2016.

[MSSW16] Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian Exploration: Incentivizing Exploration in Bayesian Games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 661–661, Maastricht The Netherlands, July 2016. ACM.

[MTG18] Johan Markdahl, Johan Thunberg, and Jorge Goncalves. Almost Global Consensus on the n -Sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, June 2018.

[MTG20] Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Learning Probably Approximately Correct Maximin Strategies in Simulation-Based Games with

Infinite Strategy Spaces. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 834–842, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems.

[MWY+20] Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. Bursting the filter bubble: Fairness-aware network link prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):841–848, Apr. 2020.

[Mye82] Roger B Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of Mathematical Economics*, pages 67–81, June 1982.

[MZ91] George J Mailath and Peter Zemsky. Collusion in second price auctions with heterogeneous bidders. *Games and Economic Behavior*, 3(4):467–486, November 1991.

[NHH+14] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 677–686, New York, NY, USA, 2014. Association for Computing Machinery.

[NRTV07] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani, editors. *Algorithmic Game Theory.* Cambridge University Press, Cambridge, 2007.

[NST15] Denis Nekipelov, Vasilis Syrgkanis, and Eva Tardos. Econometrics for Learning Agents. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation - EC '15*, pages 1–18, Portland, Oregon, USA, 2015. ACM Press.

[PLST20] Renato Paes Leme, Balasubramanian Sivan, and Yifeng Teng. Why Do Competitive Markets Converge to First-Price Auctions? In *Proceedings of The Web Conference 2020*, pages 596–605, Taipei Taiwan, April 2020. ACM.

[PMB23] Siddharth Prasad, Martin Mladenov, and Craig Boutilier. Content prompting: Modeling content provider dynamics to improve user welfare in recommender ecosystems. *arXiv preprint arXiv:2309.00940*, 2023.

[PZMDH20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[Raj19] Suju Rajan. First-price auctions: A key piece of the transparency puzzle. 2019.

[Ren94] James Renegar. Some perturbation theory for linear programming. *Mathematical Programming*, pages 73–91, February 1994.

[RJJX15] Zinovi Rabinovich, Albert Xin Jiang, Manish Jain, and Haifeng Xu. Information Disclosure as a Means to Security. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 645–653, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems.

[Rou16] Tim Roughgarden. Lecture #17: No-Regret Dynamics. In *Twenty lectures on algorithmic game theory*. Cambridge University Press, Cambridge ; New York, NY, 2016.

[RZ24] Aviad Rubinstein and Junyao Zhao. Strategizing against No-Regret Learners in First-Price Auctions. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 894–921, New Haven CT USA, July 2024. ACM.

[SCB$^+$24] Antoine Scheid, Aymeric Capitaine, Etienne Boursier, Eric Moulines, Michael Jordan, and Alain Oliviero Durmus. Learning to mitigate externalities: the coase theorem with hindsight rationality. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[SLL21] Fernando P. Santos, Yphtach Lelkes, and Simon A. Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50):e2102141118, 2021.

[Slu19] Sarah Sluis. Google switches to first price auction. 2019.

[Spe73] Michael Spence. Job Market Signaling. *The Quarterly Journal of Economics*, 87(3):355, August 1973.

[SSL07] Alain Sarlette, Rodolphe Sepulchre, and Naomi Ehrich Leonard. Autonomous rigid body attitude synchronization. In *2007 46th IEEE Conference on Decision and Control*, pages 2566–2571, 2007.

[SYCY13] Ruilong Su, Li'Ang Yin, Kailong Chen, and Yong Yu. Set-oriented personalized ranking for diversified top-n recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 415–418, New York, NY, USA, 2013. Association for Computing Machinery.

[Syr17] Vasilis Syrgkanis. A sample complexity measure with applications to learning optimal auctions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5352–5359, 2017.

[TCK09] Charles S. Taber, Damon Cann, and Simona Kucsova. The Motivated Processing of Political Arguments. *Political Behavior*, 31(2):137–155, June 2009.

[TDM10] Tyler Lu, David Pal, and Martin Pal. Contextual Multi-Armed Bandits. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 485–492. PMLR, March 2010.

[TH21] Wei Tang and Chien-Ju Ho. On the Bayesian Rational Assumption in Information Design. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9:120–130, October 2021.

[TL06] Charles S. Taber and Milton Lodge. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, 50(3):755–769, July 2006.

[TPR20] Ben M. Tappin, Gordon Pennycook, and David G. Rand. Bayesian or biased? Analytic thinking and political belief updating. *Cognition*, 204:104375, November 2020.

[Vit21] Ellen Vitercik. *Automated algorithm and mechanism configuration.* PhD thesis, Carnegie Mellon University, 2021.

[VSZ04] Bernhard Von Stengel and Shmuel Zamir. Leadership with commitment to mixed strategies. Technical report, Citeseer, 2004.

[War07] Edwards Ward. Conservatism in Human Information Processing. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 2007.

[WKBJ23] Yuanhao Wang, Dingwen Kong, Yu Bai, and Chi Jin. Learning Rationalizable Equilibria in Multiplayer Games. In *The Eleventh International Conference on Learning Representations*, 2023.

[WLZL21] Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear Last-iterate Convergence in Constrained Saddle-point Optimization. In *International Conference on Learning Representations*, 2021.

[WPR16] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *Conference on Learning Theory*, pages 1562–1583. PMLR, 2016.

[WRB$^+$18] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. CIKM '18, page 2165–2173, New York, NY, USA, 2018. Association for Computing Machinery.

[WSZ20] Zihe Wang, Weiran Shen, and Song Zuo. Bayesian Nash Equilibrium in First-Price Auction with Discrete Value Distributions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1458–1466, 2020.

[WXY22] Jibang Wu, Haifeng Xu, and Fan Yao. Multi-Agent Learning for Iterative Dominance Elimination: Formal Barriers and New Algorithms. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 543–543. PMLR, July 2022.

[WZF$^+$22] Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I. Jordan, and Haifeng Xu. Sequential Information Design: Markov Persuasion Process and Its Efficient Reinforcement Learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 471–472, Boulder CO USA, July 2022. ACM.

[XFC$^+$16] Haifeng Xu, Rupert Freeman, Vincent Conitzer, Shaddin Dughmi, and Milind Tambe. Signaling in Bayesian Stackelberg Games. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pages 150–158, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[YB21] Chunxue Yang and Xiaohui Bei. Learning Optimal Auctions with Correlated Valuations from Samples. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11716–11726. PMLR, July 2021.

[YCX$^+$18] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 279–287, New York, NY, USA, 2018. Association for Computing Machinery.

[YLN$^+$22] Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. Learning from a learning user for optimal recommendations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25382–25406. PMLR, 17–23 Jul 2022.

[YLN$^+$23] Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How Bad is Top-K Recommendation under Competing Content Creators? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39674–39701. PMLR, July 2023.

[YLW$^+$24] Fan Yao, Yiming Liao, Mingzhe Wu, Chuanhao Li, Yan Zhu, James Yang, Jingzhou Liu, Qifan Wang, Haifeng Xu, and Hongning Wang. User welfare optimization in recommender systems with competing content creators. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3874–3885, New York, NY, USA, 2024. Association for Computing Machinery.

[YZ24] Kunhe Yang and Hanrui Zhang. Computational Aspects of Bayesian Persuasion under Approximate Best Response. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[ZAAZ22] Ziqiao Zhang, Said Al-Abri, and Fumin Zhang. Opinion Dynamics on the Sphere for Stable Consensus and Stable Bipartite Dissensus. *9th IFAC Conference on Networked Systems NECSYS 2022*, 55(13):288–293, January 2022.

[ZH08] Mi Zhang and Neil Hurley. Avoiding monotony: improving the diversity of recommendation lists. RecSys '08, page 123–130, New York, NY, USA, 2008. Association for Computing Machinery.

[ZIX21] You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to Persuade on the Fly: Robustness Against Ignorance. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 927–928, Budapest Hungary, July 2021. ACM.

[ZMKL05] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 22–32, New York, NY, USA, 2005. Association for Computing Machinery.

[ZR12] Morteza Zadimoghaddam and Aaron Roth. Efficiently Learning from Revealed Preference. In *Internet and Network Economics*, volume 7695, pages 114–127. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[ZWL23] Xiaoying Zhang, Hongning Wang, and Hang Li. Disentangled representation for diversified recommendations. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 490–498, 2023.