THE UNIVERSITY OF CHICAGO

TOWARDS A THEORY OF STRATEGIC ALIGNMENT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY

JIBANG WU

CHICAGO, ILLINOIS

JUNE 2025

ABSTRACT

As artificial intelligence systems grow increasingly powerful and influential, their design often overlooks the critical incentive structures embedded within the environments in which they operate. This misalignment can yield unintended and harmful outcomes — such as filter bubbles in recommendation systems and agency problems involving content creators or gig workers. This dissertation frames such phenomena under the unifying lens of strategic alignment in AI, a concept adapted from business management that emphasizes harmonizing AI behavior with the interests of all stakeholders to achieve collectively desirable outcomes.

To address these challenges, this work develops a principled foundation at the intersection of machine learning and algorithmic game theory, advancing both modeling frameworks and algorithmic solutions. We introduce incentive-aware learning algorithms and data-driven mechanisms that offer statistical and computational efficiency guarantees, aiming to enhance the robustness and responsibility of AI systems in strategic, multi-agent environments.

Chapter 2 introduces Markov Persuasion Processes (MPPs), a new model of sequential information design where a sender strategically discloses information to a stream of myopic agents in a Markovian setting. We propose a no-regret learning algorithm, `OP4`, which balances optimism and robustness to achieve sublinear regret and sample efficiency, even in high-dimensional environments via function approximation.

Chapter 3 presents a systematic study of the robust Stackelberg equilibrium (RSE), a solution concept that generalizes the strong Stackelberg equilibrium by accounting for possible suboptimal follower responses. Unlike prior robustness models, RSE accommodates broader uncertainties through a worst-case analysis framework. We establish the existence of RSE and analyze its utility guarantees, showing how the leader's performance varies with robustness levels. Despite the intractability of computing exact solutions, we develop a quasi-polynomial approximation scheme (QPTAS), and further examine the learnability of RSE under utility uncertainty, providing nearly tight sample complexity bounds. As a corollary,

we also improve upon prior results for learning SSE in both accuracy and efficiency.

Chapter 4 presents a truthful and efficient mechanism for improving scientific peer review in large conferences. By eliciting self-reported rankings from authors and applying isotonic regression over partitioned co-authorship blocks, the mechanism balances statistical accuracy with strategic incentives. We prove equilibrium truthfulness and develop a near-linear-time block optimization algorithm, supported by empirical validation on real-world conference data.

Chapter 5 explores the problem of achieving rationalizability—a weaker and epistemically grounded solution concept than Nash equilibrium—under uncoupled learning. We show that standard no-regret algorithms are exponentially inefficient and propose a novel algorithm, `Exp3-DH`, based on diminishing historical rewards. In self-play settings, `Exp3-DH` provably eliminates dominated actions within polynomial time, outperforming existing bandit algorithms.

Together, these contributions offer a unified perspective on how to incorporate strategic considerations into the design of learning algorithms and decision-making systems, fostering more robust, aligned, and socially responsible AI.

To the love, wisdom and joy along this journey.

Research means to search again. Why not?
Sometimes, a new interpretation emerges that is of vast importance.
— Isaac Asimov

The only way of discovering the limits of the possible
is to venture a little way past them into the impossible.
— Arthur C. Clarke

All models are wrong, but some are useful.
— George E. P. Box

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

This PhD journey has been not only an earnest pursuit of knowledge, but also a quiet reckoning of the self. It has meant days filled with curiosity, nights steeped in contemplation, and countless cycles of reading and researching, writing and revising, trying and thinking again. Meaning lies not in the arrival, but in the sincerity of our engagement with the people, the purpose and the process. Above all, I could never have walked this path alone. This dissertation stands not as a solitary achievement, but as a chorus of many voices to whom I owe my deepest gratitude.

First and foremost, I thank my advisor, Prof. Haifeng Xu, whose exceptional mentorship and unwavering support shaped every step of this journey. Your intellectual vitality and thoughtful guidance influenced not only my research but also the way I think. You gave me the freedom to explore, the grace to fail, and the encouragement to persist — all of which made this work possible.

I am grateful to my committee members, Profs. Yuxin Chen and Rad Niazadeh, for the thoughtful feedback and steady support throughout the dissertation. I also thank Profs. Michael Jordan, Emir Kamenica, and László Babai, whose seminal work and teaching left a lasting impression on the development of this research.

Special thanks to career mentors, Profs. Fei Fang, Weijie Su, Zhuoran Yang, Mengdi Wang, and Hongning Wang, for enduring guidance during the most formative stages of this academic path. I am also thankful for Profs. Yiding Feng, Wei Tang, Huazheng Wang, Chi Jin, Cong Ma, Lorenzo Orecchia, Alex Kale, Lily Xu, Ryan Shi, and many others, whose valuable career advice and encouragement helped me navigate key transitions along this journey.

I have been fortunate to work with and learn from many brilliant researchers, including Renqin Cai, Sijun Tan, Xiaohui Bei, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Ashwinkumar Badanidiyuru, Weiran Shen, Jiarui Gan, Minbiao Han, Fan Yao, Siyu Chen, Yifan Wu, Yifan

Guo, Simon Mahns, Chaoqi Wang, Chenghao Yang, and Hao Zhu — along with many more whose names I hold with equal gratitude. To all lab mates, classmates, and colleagues, thank you for the camaraderie, curiosity, and resilience we shared throughout this journey.

I am also thankful to the University of Chicago and University of Virginia Computer Science Departments, the UChicago Center for the Study of Existential Risk, and the Booth Stigler Center, whose generous support enabled the pursuit of this research and the completion of this dissertation.

Finally, I am profoundly grateful to family and friends. Your love has been the foundation, your faith the light through every uncertain moment. This work is, in every sense, a reflection of quiet strength.

Part I

Background and Overview

CHAPTER 1

INTRODUCTION

While intelligent systems are becoming more advanced and influential, their design often overlooks critical incentive structures within their operating environments, risking unintended and potentially harmful consequences. My research aims to **advance the design principles and approaches of intelligent systems towards *strategic alignment*** — a concept centered on *coordinating the interests of all stakeholders to achieve mutually beneficial outcomes*. Examining the theoretical foundations of machine learning and algorithmic economics, I study the strategic alignment problem over two key components of intelligent systems, illustrated in Figure 1.1:

Figure 1.1: The modeling paradigms of value alignment (left) and strategic alignment (right).

Decision Alignment: Learning *for* Strategic Decision-Making. The outcomes of data-driven decisions can be subject to the strategic responses from stakeholders under conflicting interest and asymmetric information. For example, buyers would avoid items paid to be promoted by e-commerce websites, and freelance workers would reject orders from the gig platforms, when they perceive the system's recommendations to be suboptimal. My work models the strategic interactions in the multi-agent decision-making processes and adopts the online learning framework to analyze the adaptive decision optimization problems in a complex, unknown environment.

Feedback Alignment: Learning *from* Strategic Data Sources. The data empowering machine learning systems can be strategically withheld or manipulated by the data providers. For example, applicants might selectively conceal their information, and reviewers would deliberately submit misleading feedbacks, when the system's fully informed decisions are not in their best interest. My work models data providers' incentives on the learning outcomes and adopts a mechanism design perspective to analyze how different design of statistical methods (e.g., ranking, classification or calibration) or monetary incentives could induce more desirable equilibrium outcomes.

1.1 Decision Alignment: Learning for Strategic Decision-Making

> *"Every individual... intends only his own gain, and is led by an invisible hand to promote an end which was no part of his intention."* — Adam Smith

The "invisible hand" metaphor illustrates how properly designed incentive structures can guide self-interested individuals to inadvertently promote the greater social good. This concept is increasingly relevant in the realm of machine learning, as the scale of applications expands and the conflict of economic interests intensifies. For example, a content platform wants to estimate the ad revenues from serving different types of content, but it is up to the creators to decide what content to produce. While the platform seeks high-quality content to boost its long-term growth, creators may opt to minimize their production costs. This misalignment has prompted platforms to implement revenue-sharing models, fueling the growth of the creator economy, projected to exceed half a trillion by 2027 [Bhargava, 2022]. However, current algorithm designs are inadequate, especially in light of their roles in the proliferation of misinformation and filter bubbles on the Internet. To address these problems, I proposed a general modeling framework of *principal-agent decision-making processes* and designed provably efficient learning algorithms for adaptive decision-making towards strategic alignment.

A General Framework of Principal-Agent Problems [Gan et al., 2024]. To formalize the decision optimization problem in the presence of misaligned interests and asymmetric information, I develop a general framework based on the well-established principal-agent model in microeconomics. Under this framework, one key observation is that various forms of strategic decisions in practice, such as monetary contracts, information signals and commitments, can be unified into a single mathematical representation, of which the optimal strategies can be solved in polynomial time.

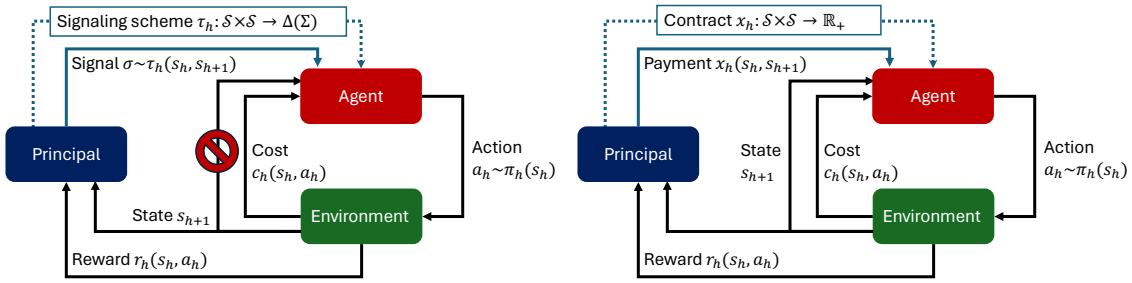


Figure 1.2: An illustration of canonical models under the principal-agent decision-making framework: Markov persuasion processes (left), contractual reinforcement learning (right).

Principal-Agent Reinforcement Learning [Wu et al., 2022c, 2024]. To tackle more real world challenges faced in decision optimization, I further extend the modeling framework into the unknown, dynamic environments. Specifically, I have developed two canonical models, as illustrated in Figure 1.2, and designed a class of efficient online reinforcement learning algorithms that provably approaches the principal's optimal strategy given the strategic agent's optimal responses. In *Markov persuasion processes* [Wu et al., 2022c], the principal has the information advantage to observe the state of the environment, and agents decide which action to take. This models the typical situations where Internet platforms hold massive datasets (both historical and real-time) and would like to persuade their users to take certain action that aligns with the platforms' objective through information signaling. In *contractual reinforcement learning* [Wu et al., 2024], the principal incentivizes agents to take certain actions through contracts binding on the realization of the next state. This model

captures another kind of common problems where Internet platforms specify some revenue-sharing policies, yet the task of creating content or providing services is delegated to its contractors or developers.

Robustifying Decisions under Bounded Rationality [Wu et al., 2022a, Gan et al., 2023]. I also consider situations where agents' responses are suboptimal. With no behavioral assumption on agent's rationality, I studied the principal's worst-case optimal decision, *δ-robust Stackelberg equilibrium*, and influences from the robustness measure δ. For its computation, I proved its hardness in general and designed a quasi-polynomial approximation scheme. This result improves existing bound of statistical and computational complexity on learning strong Stackelberg equilibria. I also consider the problem under agents' quantal responses, a common behavioral model where suboptimal actions are taken with probabilities proportional to their suboptimality. Even though the principal's optimal decision under this setting is intractable to compute, I show that the agent's bounded rationality enables sample-efficient algorithms for learning the agent's utility and thereby predicting the responses.

Multi-Agent Learning towards Rationalizability [Wu et al., 2022b]. In the case when agents are also learning to make theirs best responses, I study the multi-agent learning dynamics in strategic games and seek to understand whether they can reach any desirable outcomes, such as the basic solution concept known as *rationalizability*, where all agents eliminate their iteratively dominated actions. I show that this multi-agent learning objective is not always possible for a large class of no-regret learning algorithms, while a new type of strategic learning algorithm have provably efficient convergence guarantee by carefully reweighing the influence of historical rewards.

1.2 Feedback Alignment: Learning from Strategic Data Sources

"When a measure becomes a target, it ceases to be a good measure"

— Goodhart's law

Goodhart's law highlights that metrics lose their effectiveness when they become optimization targets. This powerful observation also characterizes the challenges faced by today's data-driven systems, where the strategic behaviors of data providers can undermine the system's overall integrity and efficacy. For example, a recent work [Bevendorff et al., 2024] investigates the question, "Is Google Getting Worse?" due to search engine optimization (SEO) — a common practice for a website or a product to boost its search ranking and Internet traffic. While certain SEO techniques are encouraged to improve search quality, the manipulative tactics used to game the system have caused several harmful incidents. As a result, search engines are caught in a constant struggle to counteract SEO spam and preserve the quality and integrity of their services. Similar concerns arise in cases such as loan application and college admission, as more and more high-stake decisions have become data-driven.

Isotonic Mechanism for Self-Evaluations [Tan et al., 2021, Wu et al., 2023]. One crucial data-driven system under stress is the academic conference peer review, where the supply of qualified reviewers is failing to meet an ever-growing volume of submissions. The typical approach as in my earlier work [Tan et al., 2021] is to design calibration schemes to mitigate reviewers' perception noises. My recent work [Wu et al., 2023] proposes *Isotonic Mechanism* to utilize authors' self-evaluation to assist large-scale peer reviews. Ironically, it is often against authors' best interest to share candid assessment of their own work, despite a valuable source of opinions. The mechanism resolves the conflict of interests by partitioning the network of co-authorships, eliciting ranking information from authors in each partition block and adjusting their papers' peer review scores through the isotonic regression.

I show that under natural conditions, truthful reporting by each author is strategically advantageous, forming a payoff-dominant Nash equilibrium, regardless of their overlapped interests in co-authored papers. In simulations based on public data from ICLR, about 10% of the mistakenly rejected papers are rectified to acceptance. Pilot experiments based on our method are undergoing in ICML. This result demonstrates a successful balance between statistical efficiency and incentive compatibility.

Auctioning with Strategic Data Providers [Wu et al., 2021]. Another example is the auction design problem when each bidder may hold a new kind of private information that can be strategically withheld, e.g., proprietary data that can improve the ad platform's estimation of conversion rates (such as user activities on the advertisers' websites). I show that it is still possible to design truthful mechanisms where the bidders would fully reveal their data. Specifically, I develop two different black-box transformations, which convert truthful mechanisms in classic setup to a truthful mechanism in this new setup. Based on this reduction approach, I propose new auction mechanisms that elicits full information from bidders and maximizes the social welfare or the auctioneer's revenue under natural conditions. Along the way, I also show that properly regulating auctioneer's usage of bidders' information can lead to more robust mechanisms with strategy-proofness.

Learning to Incentivize Information Acquisition [Chen et al., 2023]. I also study the information acquisition problem, where a principal (e.g., content platforms) hires an agent (e.g., data annotators) to gather information on her behalf and the agent can choose different effort level that determines the quality of information. First, I show that the optimal mechanism can be described a scoring rule that specifies the agent's payment based on the discrepancy between realized outcomes and the belief based on agent's reported information. Moreover, when the agent's capability and utility structure are unknown, I show that the sample-efficient learning of optimal scoring rules is impossible unless the agent's potential

effort levels are known to some extent. In this case, we can construct an online learning algorithm to approach the optimal objective with low regret guarantees.

Part II

Decision Alignment

CHAPTER 2

MARKOV PERSUASION PROCESSES

2.1 Introduction

Most sequential decision models assume that there is a sole agent who possesses and processes all relevant (online or offline) information and takes an action accordingly. However, the economic literature on information design [Kamenica and Gentzkow, 2011b, Bergemann and Morris, 2019] highlights the importance of considering information asymmetry in decision making, where the decision maker and information possessor may be two parties with different interests and goals. Complications of such kind are seen pervasively from freelance drivers canceling orders from ride-sharing platforms to buyers rejecting the product recommendation by online shopping websites, in that the recommended action is not the optimal option for the decision maker. Given the large data sets being collected by corporations and governments, with avowed goals that relate data analysis to social welfare, it is timely to pursue formal treatments of sequential information design and to understand how a social planner can strategically inform sequential agents (e.g., users, clients or citizens) to induce desirable equilibrium outcomes.

2.1.1 Our Model and Scope

In this paper, we consider an abstract framework, namely, the *Markov persuasion process* (MPP). A single privately-informed social planner (henceforth, *sender*) interacts with the self-interested, myoptic decision-makers (henceforth, *receivers*) over time in a Markovian environment. In each round, a newly arriving receiver chooses one action among several alternatives; the *state* of environment evolves accordingly and is observed by the sender and the receiver. Meanwhile, before a receiver makes his choice, the sender is able to privately observe a realized *outcome* sampled from some prior and send messages to the receiver

conditioned on this information. The utilities of the sender and receiver are different, both of which evolve with the environment and are functions of the chosen action and realized outcome. Through information design, the sender could persuade the receiver to take certain action in her interest. A standard revelation principal argument allows the sender to focus on designing signaling schemes that directly recommend action to the follower with the persuasiveness constraint (see Section 2.3.1 and 2.4.1). In the reinforcement learning problem of MPPs, we further relax the assumption such that the sender does not initially know how the environment state transits, the prior or his reward function at each step.

A Motivating Example

On today's popular ride-sharing platforms, the drivers are no longer obligated to follow the dispatched order from their platforms. Instead, the drivers have their own utilities calculated based on factors such as trip distance, time, the driver's and rider's type. Meanwhile, a platform's utility is usually some fraction of the trip charges, in addition to the operating costs and a series of performance metrics such as the riders' overall rating and wait time. The MPP captures those factors by allowing different utility functions for the platform and drivers, both of which are functions of driver's action and features of the environment.

One key ingredient of this problem is the information asymmetry. The platform nowadays holds a massive amount of historical data and has access to real-time information on active riders and driver types in different locations, based on which it could persuade the drivers to take actions in its own interest. MPPs adopt the standard information design framework to model the platform's strategic information revelation. Let us consider a concrete example. For simplicity, suppose there are two possible types of riders: *generous tippers* who consist of 1/3 of the rider population and *ordinary tippers* who are the rest. A generous tipper always tips, whereas an ordinary tipper leaves a tip in 10% of the trips. A typical driver tends to decline long-distance trips unless they are assured to receive a tip in good chances, say at least

50%. Then clearly if the platform reveals each individual rider's type, those drivers would only pick up generous tippers. In addition, if the platform reveals no information, those drivers would always decline long-distance trips, as there is only $\frac{1}{3} \times 100\% + \frac{2}{3} \times 10\% = 40\%$ of chance getting a tip according to the common prior belief on the rider type distribution. Interestingly, it turns out that the platform can significantly increase the chance for drivers to pick up long-distance trips via a simple prompt message in its App interface that reveals partial information about the rider type. For instance, it can pop up a message "$\geq 50\%$ chance of tips" whenever the rider is a generous tipper. Moreover, when the rider is an ordinary tipper, the app will pop up the same message a random half of the time, and remain silent otherwise (which is essentially a signal of low chance of tips). In this way, when the driver receives the prompt message, there is $\frac{1}{2} \times 100\% + \frac{1}{2} \times 10\% = 55\%$ of chance getting a tip and thus picking up the rider would be his optimal action. With the above *signaling scheme*, 2/3 of the riders will be picked up for their long-distance trip requests. We remark that similar signaling schemes already exist in various forms of demand heat map [Chen et al., 2015, Yang et al., 2019, Guda and Subramanian, 2019], which can be viewed as revealing partial information about the riders' willingness to pay at different regions,[1] though it is unclear whether these signaling schemes are optimally designed.

The other part of this problem is the evolution of the environment. The drivers' actions, such as whether to relocate to a certain location, or whether to accept a certain order, affects the state of the environment in many different ways: a large number of drivers moving to the same locations can cause traffic jams; a rider who was declined to be picked-up may seek alternative transportation. This requires the platform to plan accordingly and to optimize its cumulative utility in the long run. Indeed, various combinatorial optimization algorithms and reinforcement learning algorithms for vehicle repositioning, routing and order matching have been developed to optimize their operational efficiency and profit [Li et al., 2019a,

1. The commitment assumption is naturally enforced by the deployed software, which cannot be altered easily.

Qin et al., 2020, Liang et al., 2021, Qin et al., 2021]. However, those centralized planning algorithms ignore the drivers' incentives, so algorithms developed under the MPP framework have the potential to substantially improve the platform's utility in practice.

The problem above is just an exemplary case where information design meets reinforcement learning in Markovian environments. We refer the interested readers to Appendix A.1 for a few other potential applications including recommendation services for Ad keywords or online shopping.

Key Modeling Assumptions and Rationale

Based on the above motivating examples, below we clarify and justify several important modeling assumptions we made in this paper.

The MPPs models two types of information in each round: outcome and environment state. The notion of *outcome* characterizes the sender's private information in face of each receiver, such as the features of riders or information of their trip requests in the above example. The outcome follows a prior distribution such as the general demographics of riders. The platform can thus leverage such fine-grained knowledge on each rider's information, matching with the current location and preferences of each driver, to persuade the drivers to take the trip orders. Meanwhile, the notion of *state* characterizes the Markovian state of the environment, e.g., the availability of drivers in each areas. The state is affected by the receiver's action, as the availability changes after a driver decides to provide the ride. We also follow the modeling convention in reinforcement learning to let the utility functions evolve by itself over time instead of subsuming such change into states. This helps limit the state space while still having a flexible model of utility at different time steps of each episode. In practice, the Markov state captures the general driver supply or rider demand at locations that are affected by the drivers' decisions, whereas the utility evolves by itself over time, as the drivers' day rate is generally different from night rate.

We assume that each receiver *myopically* maximizes his utility at that particular step, whereas the sender is a system planner who aims to maximize her long-term accumulated expected utility. On one hand, this reduces the design complexity, as the platforms may serve thousands or millions of receivers every day and could hardly strategize on the interaction history with each individual receiver; Gan et al. [2021] shows that it is NP-hard to even approximately compute the sender's optimal policy when the receiver is far-sighted. On the other hand, we believe the assumption of myopic receivers also make realistic sense. Between two consecutive actions of the same receiver (e.g., two trips provided by the same driver), the platform may have already served thousands of other receivers in between. In such Markovian environments with many receivers, each receiver's previous action has almost no effect to the new state at which he is about to take the next action. So it may instead posit that each receiver is a fresh new receiver.

In the learning problem, we assume the true distribution of the outcome, environment's transition kernel and the sender's reward functions are all unknown to the sender. We assume that the sender knows exactly the receivers' utility functions. This is primarily a technical limit due to the well-known challenges to learn receivers' utility functions from best response in sublinear regret without additional structural assumption. However, we believe this assumption appears realistic in many applications, since the utility for a driver on the platform or for an advertiser on an ad exchange platform can usually be calculated based on some preset price rules and thus is predictable by the platform itself whereas the platform's utility may involve many unpredictable factors like users' satisfaction and cost of business operation.

Finally, we remark that many real world problems might not require our model in its full generality. As such, we introduce a tabular case (finite Markov states) in Section 2.4.1, a contextual case (no state transition) in Section 2.5.4 and showcase the simplified algorithm in Section 2.5.2 for high level intuitions. At the same time, the generalized model in Section

2.4.2 aims to introduce techniques and relax assumptions to make our proposed solution applicable to as many complicated real world problems as possible.

2.1.2 Our Results and Contributions

To provide a formal foundation for the study of sequential information design, we introduce the Markov persuasion process, where a sender, with informational advantage, seeks to persuade a stream of myopic receivers to take actions that maximize the sender's cumulative utility in a finite-horizon Markovian environment with varying prior and utility functions. We need to address a key challenge regarding the planning problem in MPPs, specifically, how to find persuasive signaling policies that are also optimized for the sender's long-term objective. Moreover, in face of the uncertainty for both the environment and receivers, there is a dilemma that the optimal policy based on estimated prior is not necessarily persuasive and thus cannot induce the desired trajectory, whereas a full information revelation policy is always persuasive but usually leads to suboptimal cumulative utility. So the reinforcement learning algorithm in MPPs has to ensure optimality under the premise of robust persuasiveness. This makes our algorithm design non-trivial and regret analysis highly challenging.

We show how to surmount these analysis and design challenges, and present a no-regret learning algorithm, which we refer to as Exp3 with Diminishing History (`OP4`), that provably achieves a $\widetilde{O}\big(\sqrt{d_\phi^2 d_\psi^3 H^4 T}\big)$ regret with high probability, where d_ϕ, d_ψ are dimensions of the feature spaces, H is the horizon length in each episode, T is the number of episodes, and $\widetilde{O}(\cdot)$ hides logarithmic factors as well as problem-dependent parameters. To establish this result, we start in Section 2.4.3 to construct a modified formulation of the Bellman equation that can efficiently determine the optimal (resp. ϵ-optimal) policy with finite (resp. infinite) states and outcomes. In Section 2.5, we then move to the learning problem and introduce the design of the `OP4` that adopts both the optimistic principle in utility estimation to incentivize exploration and the pessimism principle in prior estimation to prevent a detrimental

equilibrium for the receiver. In Sections 2.5.3 and 2.5.4, we showcase `OP4` in the tabular MPPs and contextual Bayesian persuasion problem, respectively, both of which are practical special cases of MPPs. In Section 2.8, we then generalize these positive results to MPPs with large outcome and state spaces via linear function approximation and generalized linear models.

In summary, our contributions are threefold. At the conceptual level, we identify the need for sequential information design in real-world problems and accordingly formulate a novel model, the MPP, to capture the misaligned incentives between the (sequential) decision makers and information possessors. At the methodological level, our key insight is a new algorithmic principle—optimism to encourage exploration and pessimism to induce robust equilibrium behavior. Finally, at the technical level, we develop a novel regret decomposition tailored to this combination of optimism and pessimism in the design of online learning algorithms. The fact that the combined optimism-pessimism concept can still lead to $O(\sqrt{T})$ regret for strategic setups was not clear before our new regret decomposition lemma. We expect this design principle and our proof techniques can be useful for other strategic learning problems.

2.2 Related Work

Our work is built on the foundation of information design and reinforcement learning. We refer the readers to Section 2.3.1 and 2.3.2 for background and formal introductions. Here we focus on the technical and modeling comparisons with related work from dynamic Bayesian persuasion and provably efficient reinforcement learning. In Section 2.2.3, we also distinguish the unique challenges and modeling capacities in our problem from the previous studies on incentivized exploration.

2.2.1 Dynamic Bayesian persuasion

Starting from seminal work by Kamenica and Gentzkow [2011b], the study of Bayesian persuasion looks at the design problem to influence an uninformed decision maker through strategic information revelation. Many variants of this model have been studied, with applications in security, advertising, finance, etc. [Rabinovich et al., 2015, Xu et al., 2015, Goldstein and Leitner, 2018, Badanidiyuru et al., 2018a]. More recently, several dynamic Bayesian persuasion frameworks have been proposed to model the long-term interest of the sender. Many papers [Ely, 2017, Renault et al., 2017, Farhadi and Teneketzis, 2021, Lehrer and Shaiderman, 2021] consider the setting where the sender observes the evolving states of a *Markov chain*, seeks to influence the receiver's belief of the state through signaling and thereby persuade him to take certain actions. In contrast to our setting, the receiver's actions in their models have no influence on the evolution of the Markov process and thus can only maximize his utility on his belief of current state, given all the historical signals received from the sender. In [Ely, 2017, Farhadi and Teneketzis, 2021], the Markov chain has two states (one is absorbing); the receiver is interested in detecting the jump to the absorbing state, whereas the sender seeks to prolong the time to detection of such a jump. Renault et al. [2017] show a greedy disclosure policy that ignores its influence to the future utility can be optimal in the Markov chain with special utility functions. Lehrer and Shaiderman [2021] characterize optimal strategies under different discount factors as well as the optimal values the sender could achieve. Castiglioni et al. [2020b] consider the persuasion problem in an online learning setup where the receivers' types are unknown and chosen adversarially from a finite set beforehand. In this case, effective learning is computationally intractable but does admit an $O(\sqrt{T})$ regret learning algorithm (which runs in exponential time). Closer to our model is that of Gan et al. [2021]—we both assume the Markovian environment with state transition influenced by receiver's action, as well as a separate persuasion state drawn from a prior independent of receiver's action. However, Gan et al. [2021] focus on the planning

problem for the infinite-horizon MDP, solving sender's optimal signaling policy when the environment is known in cases when the receiver is myoptic or far-sighted. In particular, it is shown as NP-hard to approximate an optimal policy against a far-sighted receiver, which also justifies our interest in the myoptic receiver. Another related work by Zu et al. [2021] studies the learning problem in the repeated persuasion setting (without Markov state transition) between a stream of myopic receivers and a sender without initial knowledge of the prior. It introduces the notion of regret as well as the robustness principle to this learning problem that we adopt and generalize to our model.

2.2.2 Efficient Reinforcement Learning

Reinforcement learning has seen its successful applications in various domains, such as robotics, finance and dialogue systems [Kober et al., 2013, Zheng et al., 2020, Li et al., 2016]. Along with the empirical success, we have seen a growing quest to establish provably efficient RL methods. Classical sample efficiency results focus on tabular environments with small, finite state spaces [Auer et al., 2008, Osband et al., 2016, Azar et al., 2017, Dann et al., 2017, Strehl et al., 2006, Jin et al., 2018, Russo, 2019]. Notably, through the design principle, known as optimism in the face of uncertainty [Lattimore and Szepesvári, 2020], an RL algorithm would provably incur a $\Omega(\sqrt{|\mathcal{S}||\mathcal{A}|T})$ regret under the tabular setting, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces respectively [Jin et al., 2018, Azar et al., 2017]. More recently, there have been advances in RL with function approximation, especially the linear case. Jin et al. [2020] proposed an efficient algorithm for a setting where the transition kernel and the utility function are both linear functions with respect to a feature mapping: $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. A similar assumption has been studied for different settings and has led to sample efficiency results [Yang and Wang, 2019, Du et al., 2019, Neu and Pike-Burke, 2020, Zanette et al., 2020, He et al., 2021]. Moreover, other general function approximations have been studied in parallel, including generalized linear function approximation [Wang et al.,

2019], linear mixture MDPs based on a ternary feature mapping [Ayoub et al., 2020, Zhou et al., 2021b, Cai et al., 2020, Zhou et al., 2021a], kernel approximation [Yang et al., 2020] as well as models based on the low Bellman rank assumption [Jiang et al., 2017, Dann et al., 2018]. We make use of these function approximation techniques to model our conditional prior, and we show how to address information design within these efficient reinforcement learning frameworks, thereby obtaining provably sample-efficient algorithms for MPPs.

2.2.3 Incentivized Exploration

Our model is also related to the recent series of work on incentivized exploration [Kremer et al., 2014, Mansour et al., 2021, Simchowitz and Slivkins, 2021]. In a similar way, their models assume myopic and strategic receivers who would only follow the recommended actions if indeed in their best interests. The sender (recommendation system) also exploits the information asymmetry through information design to incentivize the agents to take recommended (often explorative) actions that serve the sender's interest. Nevertheless, our work differs from incentivized exploration in at least two fundamental ways. First, the two frameworks have fundamentally different types of information asymmetry. In incentivized exploration, the sender's information advantage is her refined posterior belief on the environment parameter based on the observed history; all receivers in the stream have no knowledge on the history but the prior. In contrast, akin to the standard Bayesian persuasion, the sender's information advantage in MPP is the exact knowledge of the environment parameter, i.e., the realized outcomes at this round. Thus the sender in MPP does *not* need to learn the underlying environment parameter (she directly observes it) but instead needs to learn the true distribution of the outcomes as well as the enviroment's transition kernel. Second, the MPP is natively designed to model the misaligned objectives between sender and receivers, whereas standard models in incentivized exploration assumes the same sender-receiver utility at each round but the tension is between the sender's long-term objective and

receivers' short-term objectives.[2] From this perspective, MPPs can handle arbitrarily misaligned incentives whereas incentivized exploration (as the model name suggests) typically addresses the tension between long-term and short-term interests. The above differences also lead to very different results for the two problems. For instance, we are able to design learning algorithms with no-regret guarantees even for MPPs with infinite number of states, whereas Simchowitz and Slivkins [2021] show that incentivized exploration in Markovian environment requires exponential sample complexity dependence on $O(|\mathcal{S}||\mathcal{A}|H)$, due to the necessity to hide exploration under its Bayesian incentivize-compatible constraints. That said, both models are based on well-motivated problems in real life, different technical insights are developed and thus the studies on one side should not subsume the other.

2.3 Preliminaries

This section provides some necessary background in information design and Markov decision processes, as preparation for our model of Markov persuasion processes presented in the next section.

2.3.1 Basics of Information Design

Classic information design [Kamenica and Gentzkow, 2011b] considers the persuasion problem between a single sender (she) and receiver (he). The receiver is the only actor, and looks to take an action $a \in \mathcal{A}$ which results in receiver utility $u(\omega, a)$ and sender utility $v(\omega, a)$. Here $\omega \in \Omega$ is the realized *outcome* of certain environment uncertainty, which is drawn from a prior distribution $\mu \in \Delta(\Omega)$, and $\mathcal{A}$ is a finite set of available actions for the receiver. While $u, v : \Omega \times \mathcal{A} \to [0, 1]$ and the prior distribution μ are all common knowledge,

2. Mansour et al. [2021] discussed the setting where the sender has misaligned utility with the agents. However, such utility is only modeled as the auxiliary feedback, and there is no guarantee for the sender's recommendation policy to optimize its own utility while satisfying BIC.

the sender possesses an informational advantage and can privately observe the realized outcome ω. The persuasion problem studies how the sender can selectively reveal her private information about ω to influence the receiver's decisions and ultimately maximize her own expected utility v.

To model the sender's strategic revelation of information, it is standard to use a *signaling scheme*, which essentially specifies the conditional distribution of a random variable (namely the *signal*), given the outcome ω. Before the realization of the outcome, the sender commits to such a signaling scheme. Given the realized outcome, the sender samples a *signal* from the conditional distribution according to the *signaling scheme* and reveals it to the receiver. Upon receiving this *signal*, the receiver infers a posterior belief about the outcome via Bayes' theorem (based on the correlation between the signal and outcome ω as promised by the signaling scheme) and then chooses an action a that maximizes the expected utility.

A standard revelation-principle-style argument shows that it is without loss of generality to focus on *direct* and *persuasive* signaling schemes [Kamenica and Gentzkow, 2011b]. A scheme is direct if each signal corresponds to an action recommendation to the receiver, and is persuasive if the recommended action indeed maximizes the receiver's a posteriori expected utility. More formally, in a direct signaling scheme, $\pi = (\pi(a|\omega) : \omega \in \Omega, a \in \mathcal{A})$, $\pi(a|\omega)$ denotes the probability of recommending action a given realized outcome ω. Upon receiving an action recommendation a, the receiver computes a posterior belief for ω: $\mathbf{Pr}(\omega|a) = \frac{\mu(\omega)\pi(a|\omega)}{\sum_{\omega'}\mu(\omega')\pi(a|\omega')}$. Thus, the action recommendation a is persuasive if and only if a maximizes the expected utility w.r.t. the posterior belief about ω; i.e., $\sum_\omega \mathbf{Pr}(\omega|a)\cdot u(\omega,a) \geq \sum_\omega \mathbf{Pr}(\omega|a)\cdot u(\omega,a')$ for any $a' \in \mathcal{A}$. Equivalently, we define *persuasiveness* as $\sum_{\omega\in\Omega}\mu(\omega)\pi(a|\omega)\cdot[u(\omega,a)-u(\omega,a')] \geq 0, \forall a,a' \in \mathcal{A}$. Let $\mathcal{P} = \{\pi : \pi(\cdot|\omega) \in \Delta(\mathcal{A}) \text{ for each } \omega \in \Omega\}$ denote the set of all signaling schemes. To emphasize that the definition of persuasiveness depends on the prior μ, we denote the set of

persuasive schemes on prior μ by

$$\mathrm{Pers}(\mu) \coloneqq \left\{\pi \in \mathcal{P} : \sum_{\omega\in\Omega} \mu(\omega)\pi(a|\omega)\left[u(\omega,a) - u\left(\omega,a'\right)\right] \geq 0, \quad \forall a, a' \in \mathcal{A}\right\}.$$

Given a persuasive signaling scheme $\pi \in \mathrm{Pers}(\mu)$, it is in the receiver's best interest to take the recommended action and the sender's expected utility is $V(\mu,\pi) \coloneqq \sum_{\omega\in\Omega}\sum_{a\in\mathcal{A}} \mu(\omega)\pi(a|\omega)v(\omega,a)$.

Therefore, given full knowledge of the persuasion instance, the sender can solve for an optimal persuasive signaling scheme that maximizes her expected utility through the following linear program (LP), (see, e.g., Dughmi and Xu [2019] for details):

$$\text{Persuasion as an LP:} \qquad \mathrm{OPT}\left(\mu\right) \coloneqq \max_{\pi\in\mathrm{Pers}(\mu)} \quad V(\mu,\pi).$$

2.3.2 Basics of Reinforcement Learning and Markov Decision Processes

The Markov decision process (MDP) [Puterman, 2014, Sutton and Barto, 2018] is a classic mathematical framework for the sequential decision making problem. In this work, we focus on the model of episodic MDP. Specifically, at the beginning of the episode, the environment has an initial state s_1 (possibly picked by an adversary). Then, at each step $h \geq 1$, the agent takes some action $a_h \in \mathcal{A}$ to interact with environment at state $s_h \in \mathcal{S}$. The state s_h obeys a Markov property and thus captures all relevant information in the history $\{s_i\}_{i<h}$. Accordingly, the agent receives the utility $v_h(s_h, a_h) \in [0,1]$ and the system evolves to the state of the next step $s_{h+1} \sim P_h(\cdot|s_h, a_h)$. Such a process terminates after $h = H$, where H is also known as the horizon length. Here, $\mathcal{A}$ is a finite set of available actions for the agent, $\mathcal{S}$ is the (possibly infinite) set of MDP states. The utility function $v_h : \mathcal{S}\times\mathcal{A} \to [0,1]$ and transition kernel $P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ may vary at each step. A policy of the agent $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ characterizes her decision making process at step h—after observing the state s, the agent takes action a with probability $\pi_h(a|s)$.

In an episodic MDP with H steps, under policy $\boldsymbol{\pi} = \{\pi_h\}_{h\in[H]}$, we define the value function as the expected value of cumulative utilities starting from an arbitrary state,

$$V_h^{\pi}(s) \coloneqq \mathbb{E}_{P,\boldsymbol{\pi}}\bigg[\sum_{h'=h}^{H} v_h(s_{h'}, a_{h'})\bigg| s_{h'} = s\bigg], \quad \forall s \in \mathcal{S}, h \in [H].$$

Here $\mathbb{E}_{P,\boldsymbol{\pi}}$ means that the expectation is taken with respect to the trajectory $\{s_h, a_h\}_{h\in[H]}$, which is generated by policy $\boldsymbol{\pi}$ on the transition model $P = \{P_h\}_{h\in[H]}$. Similarly, we define the action-value function as the expected value of cumulative utilities starting from an arbitrary state-action pair,

$$Q_h^{\pi}(s,a) \coloneqq v_h(s_h, a_h) + \mathbb{E}_{P,\boldsymbol{\pi}}\bigg[\sum_{h'=h+1}^{H} v_h(s_{h'}, a_{h'})\bigg| s_{h'} = s, a_{h'} = a\bigg], \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h \in [H].$$

The optimal policy is defined as $\boldsymbol{\pi}^* \coloneqq \arg\max_{\boldsymbol{\pi}} V_h^{\boldsymbol{\pi}}(s_1)$, which maximizes the (expected) cumulative utility. Since the agent's action affects both immediate utility and next states that influences its future utility, it thus demands careful planning to maximize total utility. Notably, $\boldsymbol{\pi}^*$ can solved by dynamic programming based on the Bellman equation [Bellman, 1957]. Specifically, with $V_{H+1}^*(s) = 0$ and for each h from H to 1, iteratively update $Q_h^*(s,a) = v_h(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V_{h+1}^*(s', a)$, $V_h^*(s) = \max_{a\in\mathcal{A}} Q_h^*(s,a)$, and determine the optimal policy $\boldsymbol{\pi}^*$ as the greedy policy with respect to $\{Q_h^*\}_{h\in[H]}$.

In online reinforcement learning, the agent has no prior knowledge of the environment, namely, $\{v_h, P_h\}_{h\in[H]}$, but aims to learn the optimal policy by interacting with the environment for T episodes. For each $t \in [T]$, at the beginning of the t-th episode, after observing the initial state s_1^t, the agent chooses a policy $\boldsymbol{\pi}^t$ based on the observations before t-th episode. The discrepancy between $V_1^{\boldsymbol{\pi}^t}(s_1^t)$ and $V_1^*(s_1^t)$ serves as the suboptimality of the agent at the t-th episode. The performance of an online learning algorithm is measured by the expected regret, $\mathrm{Reg}(T) \coloneqq \sum_{t=1}^{T}[V_1^*(s_1^t) - V_1^{\boldsymbol{\pi}^t}(s_1^t)]$.

2.4 Markov Persuasion Processes

This section introduces the Markov Persuasion Process (MPP), a model for sequential information design in *Markovian environments*. We start by introducing a basic, tabular model of MPP that captures a set of basic problems in real life. We then consider the generalized MPP model with richer structure such as large state spaces by incorporating function approximation techniques.

2.4.1 A Model of Markov Persuasion Processes (MPPs)

We define the basic model of the Markov persuasion process. In a nutshell, an MPP inherits the notion of outcome space Ω and prior μ from a Bayesian persuasion model, as well as the notion of state space $\mathcal{S}$, action space $\mathcal{A}$, and transition kernel P from a standard episodic MDP.[3]

At a high level, there are two major differences between MPPs and MDPs: (1) In an MPP, the planner (i.e., sender) *cannot* directly take an action but instead can leverage its information advantage and "persuade" a receiver to take a desired action a_h at each step $h \in [H]$. (2) In an MPP, the state transition is affected not only by the current action a_h and state s_h, but also by the realized outcome ω_h of Nature's probability distribution. Specifically, the state transition kernel at step h is denoted as $P_h(s_{h+1}|s_h, \omega_h, a_h)$. To capture the sender's persuasion of the receiver at step h, we assume that a fresh receiver arrives at each time h with a prior μ_h over the outcome ω_h. The sender can observe the realized outcome ω_h and would like to strategically reveal information about ω_h in order to persuade the receiver to take a certain action a_h.

Meanwhile, differing from classical single-shot information design, the immediate utility functions $u_h, v_h : \mathcal{S} \times \Omega \times \mathcal{A} \to [0, 1]$ for the receiver and sender vary not only at each step

3. In this paper, we restrict our attention to *finite-horizon* (i.e., episodic) MPPs with H steps denoted by $[H] = \{1, \cdots, H\}$, and leave the study of infinite-horizon MPPs as a future direction.

h but also additionally depend on the commonly observed state s_h of the environment.

Formally, an MPP with a horizon length H proceeds as follows at each step $h \in [H]$:

1. A fresh receiver with prior distribution $\mu_h \in \Delta(\Omega)$ and utility $u_h : \mathcal{S} \times \Omega \times \mathcal{A} \to [0, 1]$ arrives.
2. The sender commits to a *persuasive* signaling policy $\pi_h : \mathcal{S} \to \mathcal{P}$, which is public knowledge.
3. After observing the realized state s_h and outcome ω_h, the sender accordingly recommends the receiver to take an action $a_h \sim \pi_h(\cdot|s_h, \omega_h)$.
4. Given the recommended action a_h, the receiver takes an action a'_h, receives utility $u_h(s_h, \omega_h, a'_h)$ and then leaves the system. Meanwhile, the sender receives utility $v_h(s_h, \omega_h, a'_h)$.
5. The next state $s_{h+1} \sim P_h(\cdot|s_h, \omega_h, a'_h)$ is generated according to $P_h : \mathcal{S} \times \Omega \times \mathcal{A} \to \Delta(\mathcal{S})$, the state transition kernel at the h-th step.

Here we coin the notion of a *signaling policy* π_h as a mapping from state to a signaling scheme at the h-th step. It captures a possibly multi-step procedure in which the sender commits to a signaling scheme after observing the realized state and then samples a signal after observing the realized outcome. For notational convenience, we denote $\pi(a|s, \omega)$ as the probability of recommending action a given state s and realized outcome ω. We can also generalize the notion of persuasiveness from classic information design to MPPs and define $\mathrm{Pers}(\mu, u)$ as the persuasive set that contains all signaling policies that are persuasive to the receiver with utility u and prior μ for any state $s \in \mathcal{S}$:

$$\mathrm{Pers}(\mu, u) \coloneqq \left\{ \pi : \mathcal{S} \to \mathcal{P} : \int_{\omega \in \Omega} \mu(\omega)\pi(a|s, \omega)\big[u(s, \omega, a) - u(s, \omega, a')\big]\mathrm{d}\omega \geq 0, \quad \forall a, a' \in \mathcal{A}, s \in \mathcal{S} \right\}.$$

Recall that $\mathcal{P}$ consists of all mappings from Ω to $\Delta(\mathcal{A})$. As such, the sender's persuasive signaling scheme $\pi_h \in \text{Pers}(\mu_h, u_h)$ is essentially a stochastic policy as defined in standard MDPs — π_h maps a state s_h to a stochastic action a_h — except that here the probability of suggesting action a_h by policy π_h depends additionally on the realized outcome ω_h that is only known to the sender.

We say $\boldsymbol{\pi} \coloneqq \{\pi_h\}_{h\in[H]}$ is a *feasible* policy if $\pi_h \in \text{Pers}(\mu_h, u_h), \forall h \in [H]$, because the state transition trajectory would otherwise be infeasible if the receiver is not guaranteed to take the recommended action, i.e., $a'_h \neq a_h$. We denote the set of all *feasible* policies as $\mathcal{P}^H \coloneqq \prod_{h\in[H]} \text{Pers}(\mu_h, u_h)$.

2.4.2 General MPPs with Contexts and Linear Parameterization

To provide a broadly useful modeling concept, we also study a generalized setting of the MPPs with contextual prior and a possibly large space of states, outcomes and contexts.

Contextual Prior. At the beginning of each episode, a sequence of contexts $C = \{c_h \in \mathcal{C}\}_{h\in[H]}$ is realized by Nature and becomes public knowledge. And we allow the prior μ_h to be influenced by the context c_h at each step h, and thus denote it by $\mu_h(\cdot|c_h)$. Specifically, the contextual information is able to model the uncertainty such as the varying demographics of active user group affected by events at different time of the day. For example, the scheduled concerts or sport games affects the rider demand in certain locations. And in the case of online shopping platforms, the prior of consumer interests may be affected by the different holidays or seasons at different time of year. Here we allow the sequence of contexts to be adversarially generated.

Linear Parameterization. We also relax the state, context and outcome space $\mathcal{S}, \mathcal{C}, \Omega$ to be continuous and additionally assume that the transition kernels and utility functions are linear, and the conditional priors of outcomes are generalized linear models (GLM) of the

context at each steps. More formally, for each step $h \in [H]$, our linearity condition assumes:

- The sender's utility is $v_h \coloneqq v_h^*(s_h, \omega_h, a_h) = \psi(s_h, \omega_h, a_h)^\top \gamma_h^*$, where (1) $\psi(\cdot, \cdot, \cdot) \in \mathbb{R}^{d_\psi}$ is a known feature vector; (2) $\gamma_h^* \in \mathbb{R}^{d_\psi}$ is the unknown linear parameter at step h.

- The next state s_{h+1} is drawn from the distribution $P_{M,h}^*(\cdot|s_h, \omega_h, a_h) = \psi(s_h, \omega_h, a_h)^\top M_h^*(\cdot)$, where $M_h^* = (M_h^{(1)}, M_h^{(2)}, \ldots, M_h^{(d_\psi)})$ is a vector of d_ψ unknown measures over $\mathcal{S}$ at step h.

- The outcome $\omega_h \in \mathbb{R}$ subjects to a generalized linear model (GLM), which models a wider range of hypothesis function classes.[4] Given the context c_h, there exists a link function $f : \mathbb{R} \to \mathbb{R}$ such that $\omega_h = f(\phi(c_h)^\top \theta_h^*) + z_h$, where $\phi(\cdot) \in \mathbb{R}^{d_\phi}$ is a feature vector and $\theta_h^* \in \mathbb{R}^{d_\phi}$ is an unknown parameter. The noises $\{z_h\}_{h \in [H]}$ are independent σ-sub-Gaussian variables with zero mean. We denote the prior of ω_h with parameter θ at context c as $\mu_\theta(\cdot|c)$.

Without loss of generality, we assume that there exist Φ, Ψ such that $\|\phi(s)\| \leq \Phi$,[5] $\|\psi(s, \omega, a)\| \leq \Psi$ for all $s \in \mathcal{S}, \omega \in \Omega$ and $a \in \mathcal{A}$. We also assume that $\|\theta_h^*\| \leq L_\theta$, $\|\gamma_h^*\| \leq L_\gamma$, $\|M_h^*\| \leq L_M$, $|\mathcal{A}| \geq 2$, $|\Omega| \geq 2$. Such a regularity condition is common in the RL literature.

2.4.3 Optimal Signaling Policy in MPPs

In order to maximize the sender's utility, we study the optimal policy in MPPs, in analogy to that of standard MDPs. We start by considering the value of any feasible policy $\boldsymbol{\pi}$. For

4. We note that GLM is a strictly generalization of the linear model assumption that we have for the distribution of transition kernel P. While we could use similar technique to extend the distribution of P to GLM using techniques similar to that in Wang et al. [2019], but we save such an extension for simplicity, since it is not the primary focus of our work.

5. To simplify notation, we will always omit the subscript of the L_2 norm in this paper.

each step $h \in [H]$, we define the sender's value function, $V_h^\pi : \mathcal{S} \to \mathbb{R}$, as the expected value of cumulative utilities under $\boldsymbol{\pi}$ when starting from an arbitrary state at the h-th step. Mathematically,

$$V_h^\pi(s) \coloneqq \mathbb{E}_{P,\mu,\pi}\left[\sum_{h'=h}^{H} v_{h'}\big(s_{h'}, \omega_{h'}, a_{h'}\big)\middle| s_h = s\right],$$

where the expectation $\mathbb{E}_{P,\mu,\pi}$ is taken with respect to the randomness of the trajectory (i.e., randomness of state transition), realized outcome and the stochasticity of $\boldsymbol{\pi}$. Accordingly, we define the Q-function (action-value function) $Q_h^\pi : \mathcal{S} \times \Omega \times \mathcal{A} \to \mathbb{R}$ as the expected value of cumulative utilities under $\boldsymbol{\pi}$ when starting from an arbitrary outcome and state-action pair at the h-th step,

$$Q_h^\pi(s, \omega, a) \coloneqq v_h(s, \omega, a) + \mathbb{E}_{P,\mu,\pi}\left[\sum_{h'=h+1}^{H} v_{h'}\big(s_{h'}, \omega_{h'}, a_{h'}\big)\middle| s_h = s, \omega_h = \omega, a_h = a\right].$$

By definition, $Q_h(\cdot,\cdot,\cdot), V_h(\cdot) \in [0, h]$, since $v_h(\cdot,\cdot,\cdot) \in [0, 1]$. To simplify notation, for any Q-function Q_h and any distributions μ_h and π_h over Ω and $\mathcal{A}$, we additionally denote

$$\big\langle Q_h, \mu_h \otimes \pi_h \big\rangle_{\Omega\times\mathcal{A}}(s) \coloneqq \mathbb{E}_{\omega\sim\mu_h, a\sim\pi_h(\cdot|s,\omega)}\left[Q_h(s, \omega, a)\right].$$

Using this notation, the Bellman equation associated with signaling policy $\boldsymbol{\pi}$ becomes

$$Q_h^\pi(s, \omega, a) = (v_h + P_h V_{h+1}^\pi)(s, \omega, a), \quad V_h^\pi(s) = \big\langle Q_h^\pi, \mu_h \otimes \pi_h \big\rangle_{\Omega\times\mathcal{A}}(s), \quad V_{H+1}^\pi(s) = 0, \tag{2.1}$$

which holds for all $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$. Similarly, the Bellman optimality equation is

$$Q_h^*(s, \omega, a) = (v_h + P_h V_{h+1}^*)(s, \omega, a),\ V_h^*(s) = \max_{\pi'_h \in \mathrm{Pers}(\mu_h, u_h)} \big\langle Q_h^*, \mu_h \otimes \pi'_h \big\rangle_{\Omega\times\mathcal{A}}(s),\ V_{H+1}^*(s) = 0. \tag{2.2}$$

We remark that the optimal policy of an MPP instance could be similarly derived from its ground-truth parameters, $\{P_h^*, v_h^*, \mu_h^*\}_{h\in[H]}$. Meanwhile, the above equations implicitly assume the context $C = \{c_h\}_{h\in[H]}$ (and thus the priors) are determined in advance. To emphasize the values' dependence on context which will be useful for the analysis of later learning algorithms, we extend the notation to $V_h^\pi(s; C), Q_h^\pi(s, \omega, a; C)$ to specify that the value (resp. Q) function is estimated based on which prior μ conditioned on which sequence of context C.

A Note on Computational Efficiency. We note that the above Bellman Optimality Equation in (2.2) also implies an efficient dynamic program to compute the optimal policy $\boldsymbol{\pi}^*$ in the basic tabular model of MPP in Section 2.4.1, i.e., when $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$ are all discrete. This is because the maximization problem in equation (2.2) can be solved in polynomial time by a linear program. The generalized MPP of Section 2.4.2 imposes some computational challenges due to infinitely many outcomes and states. Fortunately, we can efficiently solve the optimal policy in the infinite state MDP with linear function approximation:[6] we can determine $Q_h^*(\cdot,\cdot,\cdot)$ through a linear function of $q_h^* \in \mathbb{R}^{d_\psi}$ with the observed feature $\psi(\cdot,\cdot,\cdot)$, and the dominating operation is to compute $\max_{\pi\in\mathrm{Pers}(\mu_h,u_h)}\langle Q_h^*, \mu_h \otimes \pi_h\rangle_{\Omega\times\mathcal{A}}(s)$ at each step. Let the sender utility function be Q_h^*; such an optimization is exactly the problem of optimal information design with infinitely many outcomes but finitely many actions, which has been studied in previous work [Dughmi and Xu, 2019]. It turns out that there is an efficient algorithm that can signal on the fly for any given outcome ω and obtains an ϵ-optimal persuasive signaling scheme in $\mathrm{poly}(1/\epsilon)$ time. Therefore, in our later studies of learning, we will take these algorithms as given and simply assume that we can compute

6. We refer the interested reader to the description of Algorithm 1 in [Jin et al., 2020]. In the learning setup, the optimal policies can be computed efficiently based on the realized states in previous episodes using the update rule by the Sherman-Morrison formula. Nevertheless, the planning problem may not be efficient when it is required to compute the integral w.r.t. the entire probability mass of the transition kernel. In this case, it is natural to assume additional structures on the probability distribution for computational tractability.

the optimal signaling scheme efficiently at any given state s. One caveat is that our regret guarantee will additionally lose an additive ϵ factor at each step due to the availability of only an ϵ-optimal algorithm, but this loss is negligible if we set $\epsilon = O(1/(TH))$ by using a poly(TH) time algorithm.

2.5 Reinforcement Learning in MPPs and the Optimism-Pessimism Principle

In this section, we study the online reinforcement learning (RL) problem for the optimal signaling policy in an MPP. Here the learner only knows the utility functions of the receivers and has no prior knowledge about the prior distribution, the sender's utility function, and the transition kernel. While the computation of optimal policy in MPPs in Section 2.4.3 may appear analogous to that of a standard MDP, as we will see that the corresponding RL problem turns out to be significantly different, partially due to the presence of the stream of receivers, whose decisions are *self-interested* and not under the learner's control. This makes the learning challenging because if the receivers' incentives are not carefully addressed, they may take actions that are extremely undesirable to the learner. Such concern leads to the integration of the *pessimism* principle into our learning algorithm design. Specifically, our learner will be optimistic to the estimation of the Q-function, similar to many other RL algorithms, in order to encourage exploration. But more interestingly, it will be pessimistic to the uncertainty in the estimation of the prior distributions in order to prepare for detrimental equilibrium behavior. Such dual considerations lead to an interesting *optimism-pessimism principle* (OPP) for learning MPPs under the online setting. From a technical point of view, our main contribution is to prove how the mixture of optimism and pessimism principle can still lead to no regret algorithms, and this proof crucially hinges on a robust property of the MPP model which we develop and carefully apply to the regret analysis. To the best of our knowledge, this is the first time that OPP is employed to learn the optimal information

design in an online fashion. We prove that it can not only satisfy incentive constraints but also guarantees efficiency in terms of both sample complexity and computational complexity.

In order to convey our key design ideas before diving into the intricate technicalities, this section singles out two representative special cases of the online sequential information design problem. In a nutshell, we present a learning algorithm `OP4` that combines the principle of optimism and pessimism such that the sender can learn to persuade without initially knowing her own utility or the prior distribution of outcomes. In the *tabular MPP*, we illustrate the unique challenges of learning to persuade arising from the dynamically evolving environment state according to a Markov process. Through the *contextual Bayesian persuasion*, we showcase the techniques necessary for learning to persuade with infinitely many states (i.e., contexts) and outcomes. We shall omit most proofs in this section to focus on the high-level ideas, because the proof for the general setting presented in Section 2.8 suffices to imply all results for the two special cases here.

2.5.1 Learning Optimal Policies in MPPs: Setups and Benchmarks

We consider the episodic reinforcement learning problem in finite-horizon MPPs. Different from the full knowledge setting in Section 2.4.3, the transition kernel, the sender's ground-truth utility function and the outcome prior at each step of the episode, $\{P_h^*, v_h^*, \mu_h^*\}_{h\in[H]}$, are all unknown. The sender has to learn the optimal signaling policy by interacting with the environment as well as a stream of receivers in T number of episodes. For each $t \in [T] = \{1, \cdots, T\}$, at the beginning of t-th episode, given the data $\{(c_h^\tau, s_h^\tau, \omega_h^\tau, a_h^\tau, v_h^\tau)\}_{h\in[H],\tau\in[t-1]}$, the adversary picks the context sequence $\{c_h^t\}_{h\in[H]}$ as well as the initial state s_1^t, and the agent accordingly chooses a signaling policy $\boldsymbol{\pi}^t = \{\pi_h^t\}_{h\in[H]}$. Here v_h^τ is the utility collected by the sender at step h of episode τ.

Regret To evaluate the online learning performance, given the ground-truth outcome prior $\boldsymbol{\mu}^* = \{\mu_h^*\}_{h\in[H]}$, we define the sender's total (expected) regret over the all T episodes as

$$\mathrm{Reg}(T, \boldsymbol{\mu}^*) \coloneqq \sum_{t=1}^{T} \left[V_1^*(s_1^t; C^t) - V_1^{\boldsymbol{\pi}^t}(s_1^t; C^t) \right]. \tag{2.3}$$

Note that if $\boldsymbol{\pi}^t$ is not always feasible under $\boldsymbol{\mu}^*$, but is only persuasive with high probability, so the corresponding regret under $\boldsymbol{\pi}^t$ should be also in high probability sense.

It turns out that in certain degenerate cases it is impossible to achieve a sublinear regret. For example, if the set of possible posterior outcome distributions that induce some $a \in \mathcal{A}$ as the optimal receiver action has zero measure, then such posterior within a zero-measure set can never be exactly induced by a signaling scheme without a precise knowledge of the prior. Thus, the regret could be $\Omega(T)$ if receiver cannot be persuaded to play such action a. Therefore, to guarantee no regret, it is necessary to introduce certain regularity assumption on the MPP instance. Towards that end, we shall assume that the receivers' utility u and prior μ at any step of the MPP instance always satisfies a minor assumption of (p_0, D)-regularity as defined below.

Regularity Conditions An instance satisfies (p_0, D)-*regularity*, if for any feasible state $s \in \mathcal{S}$ and context $c \in \mathcal{C}$, we have

$$\mathbb{P}_{\omega\sim\mu^*(\cdot|c)} \left[\omega \in \mathcal{W}_{s,a}(D)\right] \geq p_0, \quad \forall a \in \mathcal{A},$$

where μ^* is the ground-truth prior of outcomes and $\mathcal{W}_{s,a}(D) \triangleq \{\omega : u(s,\omega,a) - u(s,\omega,a') \geq D, \forall a' \in \mathcal{A}/\{a\}\}$ is the set of outcomes ω for which the action a is optimal for the receiver by at least D at state s. In other words, an instance is (p_0, D)-regular if every action a has at least probability p_0, under randomness of the outcome, to be strictly better than other actions by at least D. This regularity condition is analogous to a regularity condition of Zu

et al. [2021] but is generalizable to infinite outcomes as we consider here.

In addition, we make two remarks about the learning setup. First, in the learning problem, we assume the prior may be unknown to the sender, but the receiver has full knowledge of prior. This assumption is not essential but just for technical rigor. Even if receivers have limited knowledge or computational power to accurately determine the utility-maximizing actions, the sender should have sufficient ethical or legal reasons to comply with the persuasive constraints in practice (e.g., in the example of subsection 2.1.1, it is mathematically equivalent to send the message " $\geq 60\%$ chance of tips" instead since all that matters is the posteriors the message induces, but this message may suffer legal or ethical issues). From a long-term perspective, receivers would only take the recommendation if the platform has a good reputation (i.e., persuasive with high probability).

Second, we assume the sender can observe the exact outcome at each round. This is slightly different from (in fact, stronger than) some economic studies on information design [Bergemann and Morris, 2019] in which the sender does not need to observe the exact outcome but only designs *experiments* to reveal information about the underlying outcome. In the planning problem of a known MPP, the two choices of assumption makes no difference for the search of the optimal signaling policy. However, in the reinforcement learning problem where the sender does not know the prior, the two assumptions will make a difference. Under our assumption, the sender can simply rely on the realized outcome to estimate the prior. Otherwise, the sender would face the additional challenge to balance the optimality of signaling scheme and its informativeness on estimating the prior; this is an interesting open research question beyond the scope of our paper. We do point out that, in many real world applications, it is natural for the platform to observe the exact outcomes such as the profiles of riders requesting a trip, or the cookies of the Internet users, and accordingly make recommendations to the receivers.

2.5.2 Algorithm: Optimism-Pessimism Principle for MPPs

The learning task in MPPs involves two intertwined challenges: (1) How to persuade the receiver to take desired actions under unknown μ_h^*? (2) Which action to persuade the receiver to take in order to explore the underlying environment? For the first challenge, due to having finite data, it is impossible to perfectly recover μ_h^*. We can only hope to construct an approximately accurate estimator of μ_h^*. To guard against potentially detrimental equilibrium behavior of the receivers due to the prior estimation error, we propose to adopt the pessimism principle. Specifically, before each episode, we conduct uncertainty quantification for the estimator of the prior distributions, which enables us to construct a confidence region containing the true prior with high probability. Then we propose to find the signaling policy within a pessimistic candidate set—signaling policies that are simultaneously persuasive with respect to all prior distributions in the confidence region. When the confidence region is valid, such a pessimism principle ensures that the executed signaling policy is always persuasive with respect to the true prior. Furthermore, to address the second challenge, we adopt the celebrated principle of optimism in the face of uncertainty [Lattimore and Szepesvári, 2020], which has played a key role in the online RL literature. The main idea of this principle is that, the uncertainty of the Q-function estimates essentially reflects our uncertainty about the underlying model. By adding the uncertainty as a bonus function, we encourage actions with high uncertainty to be recommended and thus taken by the receiver when persuasiveness is satisfied. We then fuse the two principles into the Optimism-Pessimism Principle for Markov Persuasion Process (`OP4`) in Algorithm 2.1.

Pessimism to Induce Robust Equilibrium Behavior From the data in the past episode, the sender can estimate the mean of the prior as well as obtain a confidence region through concentration inequalities. Given this partial knowledge of the prior distribution, the sender needs to design a signaling scheme that works in the face of any possible priors in

ALGORITHM 2.1: OP4 Overview

for *episode* $t = 1 \dots T$ **do**
  Receive the initial state $\{s_1^t\}$ and context $C^t = \{c_h^t\}_{h=1}^H$.
  For each step $h \in [H]$, estimate prior μ_h^t along with the confidence region $\mu_{\mathcal{B}_h^t}$, and construct an optimistic Q-function Q_h^t iteratively with the value function V_h^t.
  for *step* $h = 1, \dots, H$ **do**
    Choose robust signaling scheme $\pi_h^t \in \arg\max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}} (s_h^t; C^t)$.
    Observe state s_h, outcome ω_h and recommend action $a \sim \pi_h^t(\cdot | s_h, \omega_h)$ to the receiver.

the confidence region in order to ensure the receiver will take its recommended action with high probability. Specifically, we let $\mathrm{B}_\Sigma(\theta, \beta) \coloneqq \{\theta' : \|\theta' - \theta\|_\Sigma \leq \beta\}$ denote the closed ball in $\|\cdot\|_\Sigma$ norm of radius $\beta > 0$ centered at $\theta \in \mathbb{R}^{d_\theta}$. For any set $\mathcal{B} \subseteq \mathbb{R}^{d_\theta}$, we let $\mathrm{Pers}(\mu_\mathcal{B}, u)$ denote the set of signaling policies that are simultaneously persuasive under all weigh vectors $\theta \in \mathcal{B}$: $\mathrm{Pers}(\mu_\mathcal{B}, u) \coloneqq \bigcap_{\theta \in \mathcal{B}} \mathrm{Pers}(\mu_\theta, u)$. For any non-empty set $\mathcal{B}$, the set $\mathrm{Pers}(\mu_\mathcal{B}, u)$ is convex since it is an intersection of convex sets $\mathrm{Pers}(\mu_\theta, u)$, and is non-empty since it must contain the full-information signaling scheme. We note that since $\mathrm{Pers}(\mu_\mathcal{B}, u)$ is a convex set, we can solve the linear optimization among the policies in $\mathrm{Pers}(\mu_\mathcal{B}, u)$ in polynomial time (see e.g., Zu et al. [2021]).

Optimism to Encourage Exploration In order to balance exploration and exploitation, we adopt the principle of optimism in face of uncertainty to the value iteration algorithm based on Bellman equation, following in a line of work in online RL such as Q-learning with UCB exploration [Jin et al., 2018], UCBVI [Azar et al., 2017], LSVI-UCB [Jin et al., 2020] (also see Wang et al. [2020], Yang et al. [2020], Wang et al. [2021b] and the references therein). The additional UCB bonus on the Q-value encourages exploration and has been shown to be a provably efficient online method to improve policies in MDPs. Moreover, this method not only works for the simple tabular setting, but also generalizes to settings with infinite state spaces by exploiting linearity of the Q-function and a regularized least-squares program to determine the optimal estimation of Q-value. In fact, within our framework, we could obtain

efficient learning result in the infinite state space setting through other optimism-based online RL methods and general function approximators, such as linear mixture MDPs [Ayoub et al., 2020, Zhou et al., 2021b, Cai et al., 2020, Zhou et al., 2021a], or kernel approximation [Yang et al., 2020] or bilinear classes [Du et al., 2021].

To provide a concrete picture of the learning process, we instantiate the `OP4` algorithm in two special cases and showcase our key ideas and techniques before delving into the more involved analysis of the generalized MPP setting. Nevertheless, we remark that whether the problem instance is tabular or in the form of linear or generalized linear approximations is not essential and not the focus of our study. `OP4` itself only relies on two things, i.e., the uncertainty quantification for Q-function and prior estimation. So even the model-free RL framework can be replaced by model-based RL, as we can just construct confidence region for the transition models.

2.5.3 Warm-up I: Reinforcement Learning in the Tabular MPP

We first consider MPPs in tabular setting with finite states and outcomes, as described in Section 2.4.1. In this case, the prior on outcomes at each step degenerates to an unknown but fixed discrete distribution independent of context. As linear parameterization is not required for discrete probability distribution, the algorithm can simply update the empirical estimation of μ_h^t through counting. Similarly, the transition kernel P_h^* is estimated through the occurrence of observed samples, and we uses this estimated transition to compute the Q-function $\widehat{Q}_h^t$ from the observed utility and estimated value function in the next step, according to the Bellman equation. To be specific, for each $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}$, μ_h^t and $\widehat{Q}_h^t$

are estimated through the following equations:

$$\mu_h^t(\omega) \leftarrow \frac{\lambda/|\Omega| + N_{t,h}(\omega)}{\lambda + t - 1},$$
$$\widehat{Q}_h^t(s,\omega,a) \leftarrow \frac{1}{\lambda + N_{t,h}(s,\omega,a)} \sum_{\tau\in[t-1]} \big\{\mathbb{I}(s_h^\tau = s, \omega_h^\tau = \omega, a_h^\tau = a)\big[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau)\big]\big\},$$

where $N_{t,h}(\omega) = \sum_{\tau\in[t-1]} \mathbb{I}(\omega_h^\tau = \omega)$ and $N_{t,h}(s,\omega,a) = \sum_{\tau\in[t-1]} \mathbb{I}(s_h^\tau = s, \omega_h^t = \omega, a_h^t = a)$ respectively count the effective number of samples that the sender has observed arriving at ω, or the combination $\{s,\omega,a\}$), and $\lambda > 0$ is a constant for regularization.

In our learning algorithm, we determine the radius of confidence region $\mathcal{B}_h^t$ for μ_h^t according to confidence bound $\epsilon_h^t = O(\sqrt{\log(HT)/t})$. Moreover, we add a UCB bonus term of form $\rho/\sqrt{N_{t,h}(s,\omega,a)}$ to $\widehat{Q}_h^t$ to obtain the *optimistic Q-function* Q_h^t. Then, it selects a robustly persuasive signaling scheme that maximizes an optimistic estimation of Q-function with respect to the current prior estimation μ_h^t. Finally, it makes an action recommendation a_h^t using this signaling scheme, given the state and outcome realization $\{s_h^t, \omega_h^t\}$.

Theorem 2.1. *Let $\epsilon_h^t = \widetilde{O}(\sqrt{1/t})$, and $\rho = \widetilde{O}(|S|\cdot|\Omega|\cdot|A|H)$. Under (p_0, D)-regularity, with prob. at least $1 - 3H^{-1}T^{-1}$, `OP4` has regret $\widetilde{O}\big(|C|(|S|\cdot|\Omega|\cdot|A|)^{3/2}\cdot H^2\sqrt{T}/(p_0 D)\big)$ in tabular MPPs.*

To obtain the regret of `OP4`, we have to consider the regret arising from different procedures. Formal decomposition of the regret is described in Lemma A.1. Separately, we upper bound errors incurred from estimating Q-function (Lemma A.2), the randomness of of choosing the outcome, action and next state (Lemma A.8) as well as estimating the prior of outcome and choosing a persuasive signaling scheme that is robustly persuasive for a subset of priors (Lemmas A.3 and A.4). As the two warm-up models are special cases of the general MPP, the proof of the above properties follows from that of the general MPP setting, and thus is omitted here.

2.5.4 Warm-up II: Reinforcement Learning in Contextual Bayesian Persuasion

We now move to another special case with $H = 1$, such that the MPP problem reduces to a contextual-bandit-like problem, where transitions no longer exist. Given a context c and a persuasive signaling policy π, the value function is simply the sender's expected utility for any $s \in \mathcal{S}$,

$$V^{\pi}(s;c) \coloneqq \int_{\omega} \sum_{a \in A} \mu(\omega|c)\pi(a|s,\omega)v(s,\omega,a)\mathrm{d}\omega.$$

The sender's optimal expected utility is defined as $V^*(s;c) \coloneqq \max_{\pi \in \mathrm{Pers}(\mu_{\theta^*}(\cdot|c),u^*)} V^{\pi}(s;c)$.

Meanwhile, we consider the general setting where outcome ω is a continuous random variable that subjects to a generalized linear model. To be specific, the prior μ is conditioned on the context c with the mean value $f(\phi(c)^{\top}\theta)$. For the prior μ and link function f, we assume the smoothness of the prior and the bounded derivatives of the link function:

Assumption 2.6. *There exists a constant $L_{\mu} > 0$ such that for any parameter θ_1, θ_2, we have $\left\|\mu_{\theta_1}(\cdot|c) - \mu_{\theta_2}(\cdot|c)\right\|_1 \leq L_{\mu}\left\|f\left(\phi(c)^{\top}\theta_1\right) - f\left(\phi(c)^{\top}\theta_2\right)\right\|$ for any given context c.*

Assumption 2.7. *The link function f is either monotonically increasing or decreasing. Moreover, there exists absolute constants $0 < \kappa < K < \infty$ and $0 < M < \infty$ such that $\kappa \leq |f'(z)| \leq K$ and $|f''(z)| \leq M$ for all $|z| \leq \Phi L_{\theta}$.*

It is natural to assume a Lipschitz property of the distribution in Assumption 2.6. For instance, Gaussian distributions and uniform distributions satisfy this property. Assumption 2.7 is standard in the literature [Filippi et al., 2010, Wang et al., 2019, Li et al., 2017b]. Two example link functions are the identity map $f(z) = z$ and the logistic map $f(z) = 1/(1+e^{-z})$ with bounded z. It is easy to verify that both maps satisfy this assumption.

Different from tabular setting, we are now unable to use the counting-based estimator to keep track of the distribution of the possibly infinite states and outcomes. Instead, we resort

to function approximation techniques and estimate the linear parameters θ^*, γ^*. In each episode, OP4 respectively updates the estimation and confidence region of θ^t, γ^t, with which it can determine the outcome prior under pessimism and sender's utility under optimism. To be specific, θ^t is solved by a constrained least-squares problem and q^t is solved by a regularized least-squares problem:

$$\theta^t \leftarrow \arg \min_{\|\theta\| \le L_\theta} \sum_{\tau \in [t-1]} \left[\omega^\tau - f(\phi(c^\tau)^\top \theta_h)\right]^2,$$
$$\gamma^t \leftarrow \arg \min_{\gamma \in \mathbb{R}^\psi} \sum_{\tau \in [t-1]} \left\| v^\tau - \psi(s^\tau, \omega^\tau, a^\tau)^\top \gamma \right\|^2 + \lambda \|\gamma\|^2 .$$

We then estimate the prior by setting $\mu^t(\cdot|c)$ to the distribution of $f\big(\phi(c)^\top \theta^t\big) + z$ and estimate the sender's utility by setting $v^t(\cdot,\cdot,\cdot) = \psi(\cdot,\cdot,\cdot)^\top \gamma^t$. On one hand, to encourage exploration, OP4 adds the UCB bonus term of form $\rho \|\psi(\cdot,\cdot,\cdot)\|_{(\Gamma^t)^{-1}}$ to the Q-function, where $\Gamma^t = \lambda I_{d_\psi} + \sum_{\tau \in [t]} \psi(s^\tau, \omega^\tau, a^\tau)\psi(s^\tau, \omega^\tau, a^\tau)^\top$ is the Gram matrix of the regularized least-squares problem and ρ is equivalent to a scalar. This is a common technique for linear bandits. On the other hand, OP4 determines the confidence region of θ^t with radius β, and ensures that signaling scheme is robustly persuasive for any possible (worst case) prior induced by linear parameters θ in this region. Combining optimism and pessimism, OP4 picks the signaling scheme among the robust persuasive set that maximizes the sender's optimistic utility.

Theorem 2.2. *Under (p_0, D)-regularity and Assumption 2.6 and 2.7, there exist absolute constants $C_1, C_2 > 0$ such that, for $\lambda = \max\{1, \Psi^2\}$, $\beta = C_1(1+\kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(T)})$, and $\rho = C_2 d_\psi \sqrt{\log(4 d_\psi \Psi^2 T^3)}$, with prob. at least $1-3T^{-1}$, OP4 has regret $\widetilde{O}\big(d_\phi \sqrt{d_\psi^3}\sqrt{T}/(p_0 D)\big)$ in contextual Bayesian persuasion problems.*

Since we estimate the prior by computing an estimator θ^t, we evaluate the persuasiveness of OP4 through the probability that θ^* lies in the confidence region centered at θ^t with the

radius $\beta = O(\sqrt{d_\phi \log(T)})$ in weighted norm. Due to the smoothness of the prior and the assumption of link function, the error of the estimated prior is bounded by the product of β and the weighted norm of feature vector $\|\phi(c^t)\|_{\Sigma^t} = O(1/\sqrt{t})$, which yields the same conclusion for ϵ^t in the tabular MPP case. Also compared to Li et al. [2017a], we do not require any regularity for Σ^t, since we add a constant matrix $\Phi^2 I$ to the Gram matrix Σ^t. This ensures that Σ^t is always lower bounded by the constant $\Phi^2 > 0$. The proof of the persuasiveness and sublinear regret of contextual bandit can be viewed as a direct reduction of the MPP case when the total step $H = 1$. We decompose the regret in the same way as that in Lemma A.1 for MPPs and then estimate the upper bound for each item to measure the regret loss.

2.8 No-Regret Learning in the General Markov Persuasion Process

In this section, we present the full version of the `OP4` algorithm for MPPs and show that it is persuasive with high probability and meanwhile achieves average regret $\widetilde{O}\big(d_\phi \cdot d_\psi^{3/2} H^2 \sqrt{T}/(p_0 D)\big)$.

In the general MPP setting with the linear utility and transition, a crucial property is that the Q-functions under any signaling policy is always linear in the feature map ψ, similar to linear MDPs [Jin et al., 2020]. Therefore, when designing learning algorithms, it suffices to focus on linear Q-functions. `OP4` iteratively fits the optimal Q-function, which is parameterized by q_h^* as $\psi(\cdot,\cdot,\cdot)^\top q_h^*$ at each step $h \in [H]$. When learning the Q-functions of MPPs and the prior of persuasion states simultaneously, it operates similarly as that in tabular MPPs and contextual Bayesian persuasion. At the t-th episode, given historical data $\{(c_h^\tau, s_h^\tau, \omega_h^\tau, a_h^\tau, v_h^\tau)\}_{h\in[H],\tau\in[t-1]}$, we can estimate unknown vectors $\theta_h^*, q_h^*, \forall h \in [H]$ by solving the following constrained or regularized least-squares problems:

$$\theta_h^t \leftarrow \underset{\|\theta_h\| \le L_\theta}{\operatorname{argmin}} \sum_{\tau \in [t-1]} \left[\omega_h^\tau - f(\phi(c_h^\tau)^\top \theta_h)\right]^2,$$

$$q_h^t \leftarrow \underset{q \in \mathbb{R}^{d_\psi}}{\operatorname{argmin}} \sum_{\tau \in [t-1]} \left[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau; C^t) - \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)^\top q\right]^2 + \lambda \|q\|^2.$$

Additionally, V_{h+1}^t is the estimated value function with the observed context C^t at the episode t, which we describe formally later. This estimator is used to replace the unknown transition P_h and distribution ν_h in equation (2.2). Moreover, we can update the estimate of outcome prior μ_h^t and Q-function Q_h^t respectively. Here `OP4` adds the UCB bonus to Q_h^t to encourage exploration. The formal description is given in Algorithm A.1 in Appendix A.2.1. Below we show that the `OP4` is persuasive and guarantees sublinear regret with high probability in this general setup.

Theorem 2.3. *Under (p_0, D)-regularity and Assumption 2.6 and 2.7, there exist absolute constants $C_1, C_2 > 0$ such that, for $\lambda = \max\{1, \Psi^2\}$, $\beta = C_1(1+\kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(HT)})$ and $\rho = C_2 d_\psi H \sqrt{\log(4 d_\psi \Psi^2 H^2 T^3)}$, with prob. at least $1 - 3H^{-1}T^{-1}$, `OP4` has regret $\widetilde{O}\big(d_\phi d_\psi^{3/2} H^2 \sqrt{T}/(p_0 D)\big)$ in general MPPs.*

Proof Sketch

The regret analysis of `OP4` in MPPs turns out to be challenging for its combination of pessimism and optimism simultaneously. The key to our proof is a novel way of regret decomposition to decouple the effect of pessimism and optimism. Specifically, we decompose the regret into four terms below,

$$\mathrm{Reg}(T,\mu^*) = \underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\left\{\mathbb{E}_{\mu_h^*,\pi_h^*}[\delta_h^t(s_h,\omega_h,a_h^t)|s_1=s_1^t]-\delta_h^t(s_h^t,\omega_h^t,a_h^t)\right\}}_{\text{(i)}} + \underbrace{\sum_{t\in[T]}\sum_{h\in[H]}(\zeta_{t,h}^1+\zeta_{t,h}^2)}_{\text{(ii)}}$$
$$+ \underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\mathbb{E}_{\mu_h^*,\pi_h^*}\left[\left\langle Q_h^t,\mu_h^*\otimes\pi_h^*-\mu_h^t\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h;C^t)|s_1=s_1^t\right]}_{\text{(iii)}}$$
$$+ \underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\left\langle Q_h^t,(\mu_h^t-\mu_h^*)\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h^t,C^t)}_{\text{(iv)}}.$$

We now examine and explain the conceptual meaning of each term at a high level. The technical part of decomposing the regret and bounding each term is deferred to the Appendix A.2.2.

Term (i) captures the loss due to optimism. This is measured by the temporal-difference (TD) error, $\delta_h^t(s,\omega,a) = (v_h^t + P_h V_{h+1}^t - Q_h^t)(s,\omega,a;C^t)$, which intuitively quantifies how far the Q-functions $\{Q_h^t\}_{h\in[H]}$ are from satisfying the Bellman optimality equation in equation (2.2). Due to the optimistic Q-value estimation, δ_h^t in term (i) is always non-positive. Hence, even without observing the trajectories under ground-truth prior μ^* and the optimal signaling policy π^*, we can upper bound both δ_h^t and $-\delta_h^t$, as formally stated in Lemma A.2.

Term (ii) captures the two bounded martingale difference sequences, $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$, which respectively corresponds to the randomness of realizing the outcome $\omega_h^t \sim \mu_h^*(\cdot|c_h)$ and signaling the action $a_h^t \sim \pi_h^t(s_h^t,\omega_h^t,\cdot)$, as well as the randomness of drawing the next state s_{h+1}^t from $P_h(\cdot|s_h^t,\omega_h^t,\cdot)$. We formally define those terms in equation (A.2) and prove their bound in Lemma A.8.

Term (iii) captures the loss due to pessimism. As in Lemma A.12, the key observation is that the utility gap between a robustly persuasive signaling policy and the optimal policy has a bound independent of the randomness from the trajectories under μ^* and π^*.

Term (iv) captures the bounded difference between the estimated prior μ_h^t and ground

truth prior μ_h^* by construction. The concentration bound is proved in Lemma A.4.

Finally, besides the sublinear regret bound result, we also need to show that the signaling policy of `OP4` guarantees persuasiveness w.r.t. the true prior with high probability. This is due to the concentration of the prior estimation, stated in Lemma A.5.

A note on tightness of regret bound It has been shown that the regret lower bound in linear MDP is $\Omega(\sqrt{dH^2T})$, the optimism based RL algorithm, LSVI-UCB [Jin et al., 2020], has a $\widetilde{O}(\sqrt{d^3H^3T})$ regret, and recent work, LSVI-UCB-Restart [Zhou et al., 2020], has a tighter regret bound $\widetilde{O}(\sqrt{d^{8/3}H^2T})$ under certain conditions. `OP4`, compared to LSVI-UCB, has an additional $\sqrt{H}$ term, primarily due to the pessimism for the outcome prior estimation. And it remains an open question to derive the lower bound and to develop algorithms with tight regret bound for learning in MPPs.

CHAPTER 3

ROBUST STACKELBERG EQUILIBRIUM

3.1 Introduction

The Stackelberg game stands as a cornerstone in the realm of hierarchical decision-making processes, serving as a canonical model that underpins fundamental economic models. The game in its normal form (see Section 3.2 for details) involves two players, a leader and a follower, with utility function $u_l, u_f : [m] \times [n] \to \mathbb{R}$. The leader moves first by committing to a (possibly randomized) strategy $\boldsymbol{x} \in \Delta^m$, the follower then responds with an action $j \in [n]$. Denote $u(\boldsymbol{x}, j) = \mathbf{E}_{i \sim \boldsymbol{x}}[u(i, j)]$ as the player's expected payoff under utility function u and strategy profile $(\boldsymbol{x}, j)$. A conventional design of the optimal leader strategy can be formulated as a solution to the following (optimistic) bi-level optimization problem,

$$\max_{\boldsymbol{x} \in \Delta^m} \max_{j \in [n]} u_l(\boldsymbol{x}, j) \quad \text{s.t.} \quad u_f(\boldsymbol{x}, j) \geq \max_{j' \in [n]} u_f(\boldsymbol{x}, j'), \tag{3.1}$$

which solves for the leader payoff-maximizing strategy given that the follower would respond optimally and break ties optimistically. This solution concept is known as the strong Stackelberg equilibrium (SSE). In particular, as long as the game instance is non-degenerated, the "strong" assumption of the follower's optimistic tie-breaking behavior is without loss of generality, and SSE forms the subgame perfect Nash equilibria of the underlying extensive-form game [Von Stengel and Zamir, 2010]. The SSE is a fundamental game-theoretical concept, as it reduces to the minimax strategy in zero-sum games and is viewed as another natural extension of minimax strategy to general sum games, analogous to the Nash equilibrium [Conitzer, 2016].

However, the SSE also exhibits a critical limitation that if the follower responds (even slightly) *suboptimally* to the leader, the quality of the leader's SSE strategy may deteriorate substantially. These scenarios commonly occur when there is uncertainty regarding the

follower's utilities [Letchford et al., 2009, Kiekintveld et al., 2013, Kroer et al., 2018], or when the follower is boundedly rational [Yang et al., 2012], or when the follower has imperfect perception of the leader's strategies [An et al., 2012, Muthukumar and Sahai, 2019], or even a mixture of some of these factors [Pita et al., 2010]. In this paper, we take an agnostic stance on the cause of suboptimal follower responses and adopt a worst-case analysis for the robust design of the optimal leader strategy. This leads to the following (pessimistic) bi-level optimization problem,

$$\max_{\boldsymbol{x}\in\Delta^m} \min_{j\in[n]} u_l(\boldsymbol{x}, j) \quad \text{s.t.} \quad u_f(\boldsymbol{x}, j) > \max_{j'\in[n]} u_f(\boldsymbol{x}, j') - \delta, \tag{3.2}$$

which solves for the leader payoff-maximizing strategy given that the follower's response is the worst to the leader's utility among all actions whose utilities are up to a small difference δ from the optimal response. We refer to this solution concept as the δ-robust Stackelberg equilibrium (δ-RSE). The parameter δ reflects the follower's level of suboptimality and can be chosen according to the designer's knowledge of the game, e.g., the confidence bound on utility estimation or the degree of irrationality. This solution concept naturally extends the SSE in the sense that, as $\delta \to 0$, the leader utility of δ-RSE approaches (continuously under certain conditions) until the leader utility of SSE. This property notably allows δ-RSE to serve as the pessimistic design choice in the process of learning SSE, where δ is determined by the confidence bound in parameter estimation. We defer the detailed discussion to Section 3.3 and 3.5.

Despite its simplicity and power to capture a wide range of suboptimal behaviors, there appears to be a lack of thorough analysis in the literature for such a natural solution concept of robust Stackelberg equilibrium. For simultaneous move games, somewhat similar concepts of robust Nash equilibrium have been proposed and thoroughly analyzed [Tijs, 1981, Aghassi and Bertsimas, 2006]. However, the analysis in simultaneous-move games is quite different from that of Stackelberg games. For Stackelberg games, most relevant to ours is perhaps

an applied study by Pita et al. [2010] who proposed a mixed integer linear program for computing a subtle variant of our RSE solution in order to find robust leader strategies. However, the running time of their program is exponential in the worst case; moreover, as we will elaborate in Section 3.3, the RSE variant they considered may not always exist. Many fundamental questions still remain: How to define the equilibrium concept so that it always exists? What is the computational complexity of finding the RSE? Are there provably sub-exponential algorithms for computing the RSE? What are the connections between RSE and other equilibrium concepts? These are the questions we aim to answer in this paper.

3.1.1 Our Contributions

This paper is dedicated to a principled study of δ-RSE from its analytic properties, to its computational and statistical complexity. We start by formalizing the notion of δ-RSE, where $\delta > 0$ is an upper bound on the follower's utility loss due to suboptimal behavior. We prove that under a properly chosen boundary condition, a δ-RSE exists in every Stackelberg game for any $\delta > 0$. Consequently, the leader's utility in an δ-RSE can be viewed as a well-defined function $u_{\mathrm{RSE}}(\delta)$ of $\delta > 0$. By analyzing the function $u_{\mathrm{RSE}}(\delta)$, we show several analytical properties of the δ-RSE and compare it with other solution concepts, including the classic SSE and the maximin equilibrium. In particular, under a minor non-degeneracy assumption, the function $u_{\mathrm{RSE}}(\delta)$ is proved to exhibit Lipschitz continuity within a regime of small δ, and approaches the leader utility under SSE as $\delta \to 0$. Such continuity is a "surprisingly" nice property, as an agent's equilibrium utility typically is *not* continuous in the other agents' parameters in discrete strategic games. For instance, in SSE, the leader's utility may drop significantly if the follower's utility function changes even slightly or the follower's response is slightly sub-optimal because these may cause the follower to switch to a response that is dramatically worse to the leader. Interestingly, our added layer of worst-case analysis somehow smoothed the follower response and installed the continuity property. This

continuity property has multiple interesting implications. First, it implies that regardless of what causes the follower's up-to-δ suboptimal behavior,[1] the leader's utility will always be $O(\delta)$ off from the SSE utility for small δ. Second, this property turns out to be very useful for learning both the RSE and SSE, which we will further elaborate on below.

Next, we investigate the complexity of computing a δ-RSE. In sharp contrast to the tractability of computing an SSE, we show that for any $\delta > 0$, it is NP-hard to even approximate (the leader's strategy in) a δ-RSE; this inapproximability result rules out the possibility of a fully polynomial time approximation scheme (FPTAS), assuming P $\neq$ NP. Our proof employs a highly nontrivial reduction from the EXACT 3-SET COVER (X3C) problem. The reduction is combinatorial in nature despite computing (continuous) mixed strategies, and a key technical challenge in the proof is to relate the continuous game strategy space to the combinatorial solution space of X3C. On the positive side, we present a quasi-polynomial approximation scheme (QPTAS) to compute an approximate RSE. Our proof employs the *probabilistic method* [Alon and Spencer, 2016] to prove the existence of an approximate δ-RSE, which is similar to [Lipton et al., 2003] for proving the existence of an approximate Nash equilibrium with a simple format termed *uniform strategy* (which can be enumerated in quasi-polynomial time). However, our algorithm for identifying the approximate δ-RSE is significantly different — in fact, the approximate δ-RSE is not even a uniform strategy, but instead is some strategy nearby. This is due to the nature of bi-level optimization in Stackelberg games. While a uniform leader strategy $\overline{\boldsymbol{x}}$ and any nearby strategy $\boldsymbol{x}$ will lead to similar leader utilities, they will lead to different follower responses, which in turn affects what leader utility is induced in the equilibrium. This challenge forces us to efficiently search the nearby region of a uniform strategy $\overline{\boldsymbol{x}}$ for a leader strategy $\boldsymbol{x}$ that induces the most favorable follower response. This challenge brought by follower responses is not present in

1. It is easy to see that any δ perception error on follower's payoffs [Letchford et al., 2009, Kiekintveld et al., 2013] or leader's mixed strategy [An et al., 2012, Muthukumar and Sahai, 2019] will lead to $O(\delta)$ suboptimal follower responses when payoffs are bounded.

computing an approximate Nash equilibrium. We remark that it is an intriguing yet highly non-trivial open question to close the gap between the above hardness result and QPTAS.[2]

Last but not least, we turn to the learnability of δ-RSE in a setting where the payoff functions are not known in advance but need to be learned from samples of the players' utilities. Such a learning paradigm is crucial to today's common practice of "centralized training, decentralized execution" in multi-agent learning [Lowe et al., 2017, Bai et al., 2021]. We obtain almost tight results on the learnability of δ-RSE. Specifically, we construct a learning algorithm that, with high probability, outputs a strategy with leader's utility $O(\epsilon)$ or $O(1)$ away from the δ-RSE by using $O(1/\epsilon^2)$ samples, depending on whether a continuity condition is satisfied or not. We then present hard instances with sample complexity lower bounds for each case. As a corollary of this learnability result and the continuity property mentioned, we immediately obtain an algorithm for learning SSE. This algorithm strictly improves a recent learning algorithm for SSE by [Bai et al., 2021] on both utility guarantee and computational efficiency.

3.1.2 Related Work

Stackelberg games have a wide range of applications in economics, finance and security [Von Stengel and Zamir, 2010, Van Long and Sorger, 2010, Roth et al., 2016, Kiekintveld et al., 2009, Paruchuri et al., 2008, Tambe, 2011]. These previous works considered the equilibrium concept, often known as the *strong Stackelberg equilibrium*. The theory of computing the SSE starts from the seminal work by Conitzer and Sandholm [2006] and has led to a series of algorithmic studies on Stackelberg games [Blum et al., 2019, Conitzer and Korzhyk, 2011, Korzhyk et al., 2011, Letchford and Conitzer, 2010]. These computation problems of

2. Familiar readers may recall that there was a similar gap in the complexity landscape of computing a Nash equilibrium, which was an open problem for about 10 years. Specifically, around 2005, Lipton et al. [2003] developed a QPTAS for finding an ϵ-Nash whereas Daskalakis et al. [2009], Chen et al. [2009] ruled out FPTAS for two-player Nash (assuming PPAD$\not\subseteq$P). The gap between these two results was open for about 10 years until Rubinstein [2016] settled the lower bound that rules out PTAS for Nash and matches the QPTAS of [Lipton et al., 2003], assuming the Exponential Time Hypothesis for PPAD.

SSE are fundamental to the field of bi-level optimization [Dempe, 2002] and are viewed as extensions of the classic minimax optimization problem [v. Neumann, 1928] to general sum games. A recent work by Goktas and Greenwald [2021] studied Stackelberg games with dependent strategy sets, based on the constrained maximin model [Wald, 1945], and designed first-order methods to efficiently solve the Stackelberg equilibrium under the condition of convex-concave utility and constraints. It turns out that the RSE problem is equivalent to solving a constrained maximin problem where the follower's strategy set is determined by the leader's strategy according to the δ-best response set function. However, as the δ-best response set function exhibits discontinuity without the convex-concave structure, computing RSEs is shown by our Theorem 3.6 to be computationally intractable.

The robust design problem of Stackelberg strategies has been studied in many recent works, in contexts with uncertain follower utilities [Letchford et al., 2009, Kiekintveld et al., 2013, Kroer et al., 2018] or follower's uncertain perception of leader's mixed strategies [An et al., 2012, Muthukumar and Sahai, 2019]. Another approach is to explicitly model the follower's suboptimal decision-making process with probabilistic modeling, such as the well-known quantal response model [McKelvey and Palfrey, 1995, Yang et al., 2012]. The solution concept of RSE is different from the previous robustness notion in that the RSE does not make any specific assumptions about the underlying cause of suboptimal follower behavior. Thus, it is applicable to a wider range of scenarios to address suboptimal follower behavior. To our knowledge, [Pita et al., 2010] is the only work that studied a very close RSE solution concept as us. Nevertheless, they focused mostly on experimentally verifying the performance of algorithms based on mixed-integer linear programming for computing robust solutions.

The utility uncertainties have also been considered in machine learning contexts. These results take a learning-theoretic approach and show efficient algorithms to learn the strong Stackelberg equilibrium. One line of research considers the case where the leader is only able to observe follower's responses but not the payoffs [Balcan et al., 2015, Blum et al., 2014,

Letchford et al., 2009, Peng et al., 2019]. All these works focused on a setting where the follower always optimally responds to the leader's strategy, whereas the follower in our model may respond with any approximately optimal action. Meanwhile, Bai et al. [2021] studied a setting where the learner can query any action profile and directly observe bandit feedback of the follower's payoffs. We analyze the learnability of the RSE in the same learning setup and present strengthened results as a side product of our result for learning the RSE. Our work reveals the intrinsic connections between the robust solution concept and learnability, which in some sense echoes the reduction result from online learning regret to robust mechanism design [Camara et al., 2020].

More generally, our work subscribes to the rich literature on robust game theory [Aghassi and Bertsimas, 2006, Bielefeld, 1988, Camerer, 2011, Lin et al., 2008, Tijs, 1981, Tan et al., 1995, Crespi et al., 2017]. In simultaneous games, [Tijs, 1981, Tan et al., 1995] studied the existence of δ-equilibrium point as an extension of the Nash equilibrium, where each player's strategy, given other players' strategy profile, has suboptimality at most δ. Meanwhile, Aghassi and Bertsimas [2006] proposed another equilibrium concept, known as the robust-optimization equilibrium, under which each player, given other players' strategy profile, takes the strategy that maximizes her worst-case utility under the uncertainty of her utility function. This concept is a distribution-free solution concept in contrast to the ex-post equilibrium [Crémer and McLean, 1985] in Bayesian games [Harsanyi, 1967]. Among others, Aghassi and Bertsimas [2006], Crespi et al. [2017], Perchet [2020] studied the existence and computation of robust-optimization equilibrium as well as its sensitivity to the robustness parameters, through its connection with the Nash equilibrium of a nominal game (based on worst case payoff function of the original game). Our paper also considers a robust solution concept based on the worst-case optimization in Stackelberg games and studies similar aspects of this solution concept. However, a critical difference between RSE and the robust-optimization equilibrium [Aghassi and Bertsimas, 2006] is that the source of uncertainty in

RSE primarily comes from the opponent's (possibly suboptimal) responses, instead of each player's utility function — as shown in Section 3.5, it is relatively straightforward to accommodate the latter type of uncertainty assuming the continuity of utility over the robustness parameters in our setup.

3.2 Preliminaries

Stackelberg Games. Throughout this paper, we focus on the normal-form Stackelberg game between two players, who are referred to as the leader and the follower, respectively. The leader has the first-mover advantage and commitment power. For a Stackelberg game instance (u_l, u_f), $u_l, u_f \in \mathbb{R}^{m\times n}$ denote the leader's and follower's utility matrix, where m (resp. n) is the number of leader's (resp. follower's) actions. As a convention in bimatrix games, the (i, j) entry of utility matrix $u_l(i, j)$ (resp. $u_f(i, j)$) is the leader's (resp. follower's) payoff under the action profile (i, j). We also use the standard notation $[m] := \{1, \ldots, m\}$ for the set of m actions and $\Delta^m := \{\boldsymbol{x} | \sum_{i=1}^m x_i = 1, x_i \geq 0, \forall i \in [m]\}$ for the m-dimensional simplex.

The game has two stages. The leader moves first by committing to a *mixed strategy*, $\boldsymbol{x} = (x_1, \cdots, x_m) \in \Delta^m$, where each x_i represents the probability the leader playing action i. The follower then responds to the leader's committed strategy $\boldsymbol{x}$ with some action $j \in [n]$.[3] At the end, the leader and follower receive the payoff $u_l(\boldsymbol{x}, j), u_f(\boldsymbol{x}, j)$, respectively, where $u_l(\boldsymbol{x}, j) := \sum_{i\in[m]} x_i \cdot u_l(i, j)$ (and similarly $u_f(\boldsymbol{x}, j)$) is the expected payoff over the randomness of $\boldsymbol{x}$.

A Remark on Approximation. This paper works extensively with approximate solutions for both computation and learning parts. To be consistent, all approximated solutions

3. Restricting the follower's response to the pure action set $[n]$ is without loss of generality for our analysis, because in both RSE and SSE, fixing to any leader strategy, even if the follower's optimization problem is relaxed to the mixed strategy space Δ^n, it will always admit vertex solutions.

we provide are in an additive sense, unless otherwise clarified. As a convention for studying additive approximation, we normalize the entries in all players' utility matrices u_f, u_l to be within the interval $[0, 1]$. This is without loss of generality since rescaling and shifting the utilities will not change a game's equilibrium.

Strong Stackelberg Equilibrium. The conventional solution concept of Stackelberg games is the Strong Stackelberg Equilibrium (SSE), in which the follower always optimally responds to the leader's strategy and breaks ties in favor of the leader, when there is more than one action that maximizes the follower's utility. While we have already sketched SSE as the solution to the bilevel optimization program (3.1), it is more convenient to define SSE based on the notion of follower's best response set function as follows.

Definition 1 (Strong Stackelberg Equilibrium)**.** *In a Stackelberg game (u_l, u_f), we say a strategy profile $(\boldsymbol{x}^*, j^*)$ is a strong Stackelberg equilibrium if it holds that:*

$$\boldsymbol{x}^* \in \operatorname*{argmax}_{\boldsymbol{x}\in\Delta^m} \max_{j\in \mathrm{BR}(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \quad \textit{and} \quad j^* \in \operatorname*{argmax}_{j\in \mathrm{BR}(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j). \tag{3.3}$$

where $\mathrm{BR}(\boldsymbol{x}) := \{j \in [n] | u_f(\boldsymbol{x}, j) \geq \max_{j'\in[n]} u_f(\boldsymbol{x}, j')\}$ *denotes the follower's best response set under the leader strategy* $\boldsymbol{x}$*.*

Robust Stackelberg Equilibrium. The best response set function $\mathrm{BR}(\cdot)$ defines an ideal situation, where the follower can always identify her optimal response(s) without any error. In practice, the follower may pick suboptimal responses due to various reasons, such as bounded rationality and limited observations [Pita et al., 2010]. A natural extension of the best response set, therefore, allows a small error $\delta > 0$ in the follower's choice of actions; we define the δ-optimal response set of the follower as

$$\mathrm{BR}_\delta(\boldsymbol{x}) := \left\{ j \in [n] \,\middle|\, u_f(\boldsymbol{x}, j) > \max_{j'\in[n]} u_f(\boldsymbol{x}, j') - \delta \right\} \tag{3.4}$$

for any leader strategy $\boldsymbol{x} \in \Delta^m$ and $\delta > 0$. For the completeness of definition, we denote $\mathrm{BR}_\delta(\boldsymbol{x}) = \mathrm{BR}(\boldsymbol{x})$ when $\delta = 0$.[4] As such, we can rewrite the optimization program (3.2) into a more convenient definition of the δ-robust Stackelberg equilibrium (δ-RSE) as follows.

Definition 2 (δ-Robust Stackelberg Equilibrium)**.** *In a Stackelberg game (u_l, u_f), for any $\delta > 0$, we say a strategy profile $(\boldsymbol{x}^*, j^*)$ is a δ-robust Stackelberg equilibrium if it holds that:*

$$\boldsymbol{x}^* \in \operatorname*{argmax}_{\boldsymbol{x}\in\Delta^m} \min_{j\in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \qquad \text{and} \qquad j^* \in \operatorname*{argmin}_{j\in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j). \tag{3.5}$$

In addition, we denote $u_{\mathrm{RSE}}(\delta) := u_l(\boldsymbol{x}^*, j^*)$ as the leader's utility obtained in a δ-RSE.

Inducibility Gap and Non-Degeneracy. In this paper, we also introduce a notion called the *inducibility gap*, denoted by Δ. This quantity relates to the hardness of robustifying Stackelberg leader strategy, and it plays an important role in our analysis of δ-RSE in this paper.

Definition 3 (Inducibility Gap)**.** *In a Stackelberg game (u_l, u_f), the inducibility gap is the largest constant Δ such that for any follower actions $j \in [n]$, there exists a leader strategy $\boldsymbol{x}^j$ with $u_f(\boldsymbol{x}^j, j) \geq \max_{j'\neq j} u_f(\boldsymbol{x}^j, j') + \Delta$.*

Namely, the inducibility gap is an intrinsic property of the game on how easy it is for the leader to incentivize the follower to play any action. More specifically, let $\Delta(\boldsymbol{x}, j) := u_f(\boldsymbol{x}, j) - \max_{j'\neq j} u_f(\boldsymbol{x}, j')$ denote the follower's utility margin of taking action j under leader's strategy $\boldsymbol{x}$ — a measure of how inducible is j under $\boldsymbol{x}$ (e.g., if $\Delta(\boldsymbol{x}, j) > \delta$, then $\mathrm{BR}_\delta(\boldsymbol{x}) = \{j\}$). Then, we can alternatively define the inducibility gap through the "minimax" lens:

$$\Delta = \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in[n]} \Delta(\boldsymbol{x}, j).$$

4. As we will discuss below, this definition of δ-optimal response set extends the standard best response set in the sense that $\lim_{\delta\to 0^+} \mathrm{BR}_\delta(\boldsymbol{x}) = \mathrm{BR}(\boldsymbol{x})$ under some non-degeneracy assumption.

We remark that under the classic solution concept of SSE, it is without loss of generality to assume $\Delta > 0$. First, any generic game has $\Delta \neq 0$, because for randomly generated game instances, the events that two actions always have the same payoff have zero measure [Von Stengel and Zamir, 2010]. Moreover, if a game has $\Delta < 0$, then there exists some follower action j that can never be the best follower response (thus will not be played): more formally through contraposition of Definition 3, there exists follower action j such that for any $\boldsymbol{x}$ we have $u_f(\boldsymbol{x}, j) < u_f(\boldsymbol{x}, j') + \Delta < u_f(\boldsymbol{x}, j')$ for some $j' \in [n]$. Therefore, it is without loss of generality to ignore and remove this never-to-be-played follower action j (regardless of solving or learning the game); consequently, we obtain a Stackelberg game in which every follower action at least can possibly be a best response, and such a game has $\Delta > 0$. However, for δ-RSE, it is only without loss of generality to assume $\Delta > -\delta$, since δ-suboptimal follower actions will affect the δ-RSE. Some of our results about δ-RSE hold under assumption $\Delta > 0$, which slightly loses generality compared to $\Delta > -\delta$. This discrepancy becomes smaller as δ decreases (i.e., the follower becomes less suboptimal). Overall we believe it is a reasonable assumption to pursue when δ is small.

We conclude this section with an example of Stackelberg game motivated by the real world application and illustrate the both solution concepts of SSE and δ-RSE.

Example 1 (Stackelberg Competition)**.** *The Stackelberg game origins from the duopoly competition model studied by Von Stackelberg [2010]. There are two firms who are selling the same type of product. The leader firm is able to enter a market earlier than the follower firm. The leader firm has two options, to set either a high or low price for the product. The follower firm also have two options, which is to either compete with the leader or leave this market. Their payoff matrices can be formulate as follows.*

u_l, u_f	j_1 *(compete)*	j_2 *(leave)*
i_1 *(high)*	$3, 2$	$6, 1$
i_2 *(low)*	$2, 0$	$4, 1$

Under this pair of payoff matrices, we can observe that if the leader prices the product at a high price, then the follower is willing to compete. Otherwise, the follower would choose to leave the market. To design the Stackelberg strategy, the leader has the power to randomize her actions (e.g., to offer a price in the middle of high and low). Suppose the follower chooses the best response and break tie optimistically, we can plot the leader's utility function $\max_{j \in \mathrm{BR}(\boldsymbol{x})} u(\boldsymbol{x}, j)$ *by the red shaded line and its maximum gives the SSE leader strategy,* $\boldsymbol{x}^* = (\frac{1}{2}, \frac{1}{2})$ *with the follower response* j_1*. For various reasons in practice, the follower could choose any action from his* δ*-best response set and break tie pessimistically. In this case, the leader's utility function becomes* $\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u(\boldsymbol{x}, j)$*, plotted as the green shaded line and its maximum gives the* δ*-RSE leader strategy,* $\boldsymbol{x}^* = (\frac{1+\delta}{2}, \frac{1-\delta}{2})$ *with the follower response* j_1*.*

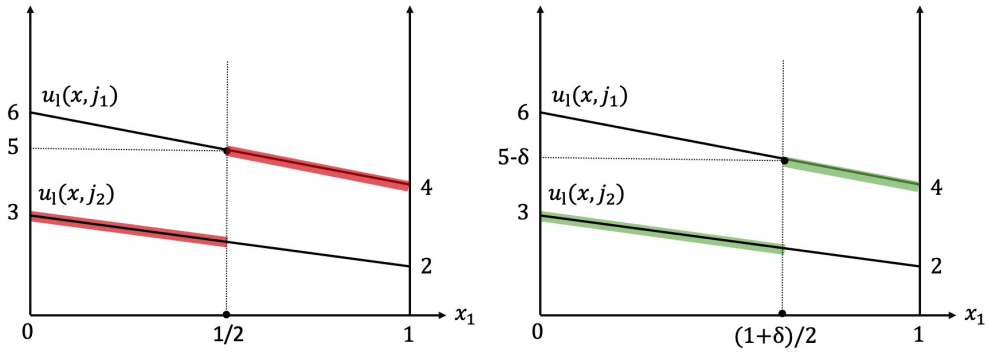


Figure 3.1: Plots of leader utility functions under follower response models in SSE (left) and δ-RSE (right). The x-axis is the probability of the leader choosing i_1, while the y-axis is the leader payoff.

3.3 Analytic Properties of RSE

3.3.1 On the Alternative Definitions of δ-RSE

In this section, we show how our current definition of δ-RSE is an inevitable choice by comparing it with several plausible alternative choices. The first objection one may raise is that the notion of δ-RSE may not be well-defined, in the sense that $\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j)$ (as

a function of $\boldsymbol{x}$) may not always have a maximum. This has been a concern for some variants of the Stackelberg equilibrium such as the *weak Stackelberg equilibrium*, where the follower is assumed to break tie pessimistically [Von Stengel and Zamir, 2010]. It turns out that δ-RSE is indeed well-defined, due to a delicate choice of the "$>$" in Equation (3.4). Below, we present a simple example to illustrate the subtlety of defining the δ-optimal response set, as the δ-RSE does not exist when the exactly δ-optimal responses are included.

Example 2 (Non-existence of δ-RSE under an Alternative Definition)**.** *We claim that in the following game instance, the δ-RSE would not exist under a slightly revised definition of δ-optimal response set,* $\mathrm{BR}_\delta(\boldsymbol{x}) = \{j | u_f(\boldsymbol{x}, j) \geq u_f(\boldsymbol{x}, j') - \delta, \forall j' \neq j\}$.

u_l, u_f	j_1	j_2
i_1	$0, -1$	$0, 0$
i_2	$0, -\delta$	$1, 0$
i_3	$0, 1$	$0, 0$

Notice that the leader's optimal strategy should ensure that $j_1 \notin \mathrm{BR}_\delta(\boldsymbol{x})$*, because the leader's utility would otherwise always be zero due to the follower's pessimistic tie-breaking. This means that i_3 can be dropped, since it only encourages the follower to take j_1 while gives the leader zero utility. Hence, it suffices to set the leader strategy as* $\boldsymbol{x} = (\xi, 1-\xi, 0)$*, under which the follower's δ-optimal response set (under the modified definition) would be* $\mathrm{BR}_\delta(\boldsymbol{x}) = \{j_2\}$*, resulting leader utility* $u_l(\boldsymbol{x}, j_2) = 1 - \xi$*. The smaller the ξ is, the larger the leader utility is. It is now clear that the optimal leader strategy is to play i_1 with infinitely small probability* $\xi > 0$ *while play i_2 with probability* $1-\xi$*. However, ξ can't be equal to* 0*, as this would include j_1 into the δ-best follower response set that makes the leader's utility become* 0*. Therefore, since ξ needs to be infinitely small but greater than* 0*, the δ-RSE is not well defined and does not exist for this example.*

Example 2 shows that the leader may optimize over an *open* set of mixed strategies that induces the desirable follower response behaviors and thus cannot achieve the exact optimal

strategy, but it is unclear how our definition of $\mathrm{BR}_\delta(\cdot)$ in Equation (3.4) can avoid this problem. Below, we present a constructive proof to show that the δ-RSE we defined via Equation (3.4) and (3.5) above exists in every game for any $\delta > 0$.

Proposition 3.1. *The δ-RSE under Definition 2 exists in every game for any $\delta > 0$ and can be computed in $O(2^n \operatorname{poly}(m, n))$ time.*

The proof idea of Proposition 3.1 is to explicitly provides an algorithm for computing the δ-RSE, though it runs in exponential time in the worst case. We refer readers to Appendix B.1.1 for the detailed proof. The following are a few remarks about Proposition 3.1. First, it implies that for instances with a small n, a δ-RSE can be computed efficiently. Nevertheless, the exponential dependency on n in the time complexity appears to be inevitable, as we will show in the next section that computing a δ-RSE is NP-hard. Second, while one may be willing to compromise and settle with the "supremum" of leader strategies, instead of using the exact "maximum" in Equation (3.5), this is generally viewed as being undesirable. Such a wrinkle of the non-existence of a solution concept is always a concern for game-theoretical analysis. Hence, we believe it is helpful to figure out the right way so that we do not always need to worry about the existence and how to tweak the solution to make it achievable.[5] This also paves the way for many of our following analysis.

Another reasonable concern here is that given the existing equilibrium concepts, is it truly necessary to define and study the δ-RSE solution concept? In particular, might it be that the SSE leader strategy or the maximin leader strategy will already perform well, i.e., achieve ϵ-optimal $u_{\mathrm{RSE}}(\delta)$ assuming the suboptimal follower responses? Unfortunately, Proposition 3.2 shows that both the SSE leader strategy and maximin leader strategy are highly suboptimal strategies with a gap of $\Omega(1)$ loss for approximating the δ-RSE. As a result, we cannot simply

5. For example, Pita et al. [2010] used non-strict inequalities for both the δ-optimal follower response and non-δ-optimal follower response. This choice actually leads to strange inconsistency in the follower behavior modeling, in which the follower will effectively break ties *against* the leader among actions strictly within δ-optimal response region but then *in favor of* the leader at the boundary of δ-optimal response regions.

apply the leader strategy from other equilibrium (e.g. SSE) to expect robust performance on par with δ-RSE. See Appendix B.1.2 for the proof of Proposition 3.2.

Proposition 3.2 (Suboptimality of Standard Equilibrium Strategies)**.** *For any $\delta > 0$, there exists game instances in which both the SSE leader strategy $\boldsymbol{x}_1$ and maximin leader strategy $\boldsymbol{x}_2$ have a constant suboptimality gap of at least $1/2$ for approximating the δ-RSE, i.e.,*

$$\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}_1)} u_l(\boldsymbol{x}_1, j) < u_{\mathrm{RSE}}(\delta) - \frac{1}{2} \quad \text{and} \quad \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}_2)} u_l(\boldsymbol{x}_2, j) < u_{\mathrm{RSE}}(\delta) - \frac{1}{2}.$$

The other common question of the δ-RSE definition is regarding the follower's pessimistic tie-breaking behavior. That is, how does the tie-breaking behavior affect the performance of δ-RSE against the suboptimal follower responses? Similar question is asked about the SSE, and Von Stengel and Zamir [2010] showed that the SSE enjoys generically unique leader payoff regardless of the follower's tie-breaking rules. Does δ-RSE share similar property? Would the leader payoff under δ-RSE strategy $\boldsymbol{x}^*$ stay the same if the follower could switch to different tie-breaking rules? The question becomes non-trivial if we were to restrict to reasonably small δ such that $\delta < \Delta$. It turns out that the answer is "No", as shown in Proposition 3.3 below. This result confirms the necessity of our worst case analysis under pessimistic tie-breaking.

Proposition 3.3 (Tie-breaking Rule Matters for δ-RSE)**.** *For any $\delta > 0$, there exist game instances with inducibility gap $\Delta > \delta$ in which the leader payoffs of δ-RSE leader strategy $\boldsymbol{x}^*$ under the optimistic and pessimistic tie-breaking rule have a difference of at least δ, i.e.,*

$$\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) < \max_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) - \delta.$$

See Appendix B.1.3 for the constructed instance with $m = 3$. Meanwhile, we show in Appendix B.1.5 that the δ-RSE strategy $\boldsymbol{x}^*$ does have unique leader payoff regardless of the follower's tie-breaking rules in any generic game with $m = 2$, which leads to another efficient

algorithm to compute δ-RSE when m is constant.

3.3.2 The Price of Robustness: δ-RSE Leader Utility Curve over δ

Given the analysis above, we know the leader's utility in a δ-RSE is a well-defined function in the domain of $\delta > 0$, for the existence and uniqueness of its value. We denote this function as $u_{\mathrm{RSE}}(\delta)$. Next, we derive the main results of this section about characteristics of δ-RSE through $u_{\mathrm{RSE}}(\delta)$. The δ interval with continuity property is especially crucial for our study of δ-RSE in the following sections. To provide an intuitive understanding of Theorem 3.4, we depict a typical shape of the function $u_{\mathrm{RSE}}(\delta)$ in Figure 3.2.

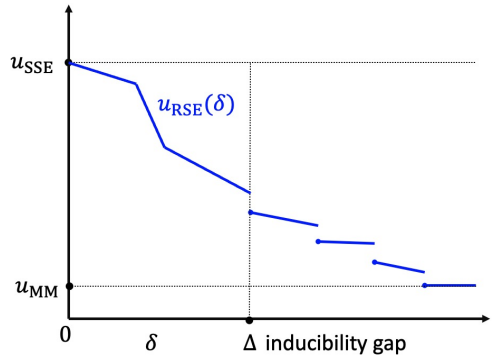


Figure 3.2: An illustration of $u_{\mathrm{RSE}}(\delta)$, the leader utility in δ-RSE w.r.t. $\delta > 0$.

Theorem 3.4. *Denote the leader's SSE payoff as u_{SSE}, and the leader's maximin payoff as $u_{\mathrm{MM}} = \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in[n]} u_l(\boldsymbol{x}, j)$, the following properties of δ-RSE hold for any game:*

1. *For any δ, δ' such that $0 < \delta \leq \delta'$, $u_{\mathrm{RSE}}(\delta)$ is monotone non-increasing and is bounded as,*

$$u_{\mathrm{SSE}} \geq u_{\mathrm{RSE}}(\delta) \geq u_{\mathrm{RSE}}(\delta') \geq u_{\mathrm{MM}}, \tag{3.6}$$

   where $u_{\mathrm{RSE}}(0^+) := \lim_{\delta\to 0^+} u_{\mathrm{RSE}}(\delta)$ exists; moreover, if $\Delta > 0$, $u_{\mathrm{RSE}}(0^+) = u_{\mathrm{SSE}}$.

2. *For any δ such that $0 < \delta < \Delta$, it holds that*

$$u_{\mathrm{RSE}}(\delta) \geq u_{\mathrm{SSE}} - \delta/\Delta. \tag{3.7}$$

3. *For any $\Delta > 0$, $u_{\mathrm{RSE}}(\delta)$ is L-Lipschitz continuous when $\delta \in (0, \Delta - \frac{1}{L}]$ and $L > 1/\Delta$. Meanwhile, $u_{\mathrm{RSE}}(\delta)$ can be discontinuous when $\delta \geq \Delta$.*

Before proceeding to the proof, we make a few remarks about Theorem 3.4. First, Property 1 suggests that the equality for $u_{\mathrm{SSE}} \geq u_{\mathrm{RSE}}(\delta)$ can be attained, if the non-degeneracy condition $\Delta > 0$ is satisfied. Note that the asumption $\Delta > 0$ is necessary as in Appendix B.1.4 we present an instance with $\Delta = 0$ such that $u_{\mathrm{RSE}}(0^+) < u_{\mathrm{SSE}}$. Property 2 shows that with the inducibility gap Δ, the lower bound of the leader's utility under δ-RSE can be improved from u_{MM} to be $u_{\mathrm{SSE}} - \delta/\Delta$. Later in Section 3.4 of algorithmic studies, we will show how this property leads to a simple approximation algorithm for δ-RSE. Lastly, Property 3 shows that $u_{\mathrm{RSE}}(\delta)$ is Lipschitz continuous when $\delta < \Delta$. This Lipschitz continuity turns out to be very useful for learning the δ-RSE in contexts where the follower utility is not known in advance. We will demonstrate its applicability to learning in Section 3.5.

Proof of Theorem 3.4.

Property 1. We begin with the monotonicity of $u_{\mathrm{RSE}}(\delta)$. According to the definition of follower's δ-optimal response set in Equation (3.4), $\mathrm{BR}_\delta(\boldsymbol{x})$ expands with δ, i.e., for any leader strategy $\boldsymbol{x}$, we have

$$\mathrm{BR}_\delta(\boldsymbol{x}) \subseteq \mathrm{BR}_{\delta'}(\boldsymbol{x}), \quad \forall 0 < \delta \leq \delta'.$$

Recall that $u_{\mathrm{RSE}}(\delta) = \max_{\boldsymbol{x} \in \Delta^m} \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j)$. Hence, for any $\boldsymbol{x} \in \Delta^m$, $0 < \delta \leq \delta'$, we have

$$u_{\mathrm{RSE}}(\delta) = \max_{\boldsymbol{x} \in \Delta^m} \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \geq \max_{\boldsymbol{x} \in \Delta^m} \min_{j \in \mathrm{BR}_{\delta'}(\boldsymbol{x})} u_l(\boldsymbol{x}, j) = u_{\mathrm{RSE}}(\delta').$$

For the lower bound of $u_{\mathrm{RSE}}(\delta)$, let $\delta \geq \max_{i \in [m], j, j' \in [n]} u(i, j) - u(i, j')$. Notice that we

have $\mathrm{BR}_\delta(\boldsymbol{x}) = [n]$, for any $\boldsymbol{x}$. This readily implies,

$$u_{\mathrm{RSE}}(\delta) = \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in\mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) = \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in[n]} u_l(\boldsymbol{x}, j) = u_{\mathrm{MM}}.$$

For the upper bound of $u_{\mathrm{RSE}}(\delta)$, notice that $\mathrm{BR}(\boldsymbol{x}) \subseteq \mathrm{BR}_\delta(\boldsymbol{x})$ for any $\delta > 0, \boldsymbol{x} \in \Delta^m$. It then follows that

$$\begin{aligned} u_{\mathrm{SSE}} = \max_{\boldsymbol{x}\in\Delta^m} \max_{j\in\mathrm{BR}(\boldsymbol{x})} u_l(\boldsymbol{x}, j) &\geq \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in\mathrm{BR}(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \\ &\geq \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in\mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) = u_{\mathrm{RSE}}(\delta). \end{aligned}$$

Moreover, $u_{\mathrm{RSE}}(0^+)$ exists because $u_{\mathrm{RSE}}(\delta)$ is monotone non-increasing in δ and is upper bounded by u_{SSE}. $u_{\mathrm{RSE}}(0^+) = u_{\mathrm{SSE}}$ is implied by the squeeze theorem, based on the bounds from Equations (3.6) and (3.7) from Property 2 when $\Delta > 0$. In Appendix B.1.4, we show an instance with $u_{\mathrm{RSE}}(0^+) < u_{\mathrm{SSE}}$ when $\Delta = 0$.

Property 2. To prove Equation (3.7), we construct a leader strategy $\widehat{\boldsymbol{x}}$ and show that playing $\widehat{\boldsymbol{x}}$ against a δ-rational follower yields a utility of at least $u_{\mathrm{SSE}} - \frac{\delta}{\Delta}$ for the leader. Let $(\boldsymbol{x}^*, j^*)$ be the SSE of the game. Let $\boldsymbol{x}^{j^*}$ be a strategy such that $u_f(\boldsymbol{x}^{j^*}, j^*) \geq u_f(\boldsymbol{x}^{j^*}, j') + \Delta$ for any $j' \neq j^*$. By definition of the inducibility gap, such a strategy must exist.

Since $\Delta > \delta$, we can set $\widehat{\boldsymbol{x}} = (1 - \frac{\delta}{\Delta})\boldsymbol{x}^* + \frac{\delta}{\Delta}\boldsymbol{x}^{j^*}$, which is a valid leader strategy. We have $\mathrm{BR}_\delta(\widehat{\boldsymbol{x}}) = \{j^*\}$, since the following inequality holds for any $j' \neq j^*$,

$$\begin{aligned} u_f(\widehat{\boldsymbol{x}}, j^*) &= (1 - \frac{\delta}{\Delta})u_f(\boldsymbol{x}^*, j^*) + \frac{\delta}{\Delta}u_f(\boldsymbol{x}^{j^*}, j^*) \\ &\geq (1 - \frac{\delta}{\Delta})u_f(\boldsymbol{x}^*, j') + \frac{\delta}{\Delta}\big(u_f(\boldsymbol{x}^{j^*}, j') + \Delta\big) \\ &= (1 - \frac{\delta}{\Delta})u_f(\boldsymbol{x}^*, j') + \frac{\delta}{\Delta}u_f(\boldsymbol{x}^{j^*}, j') + \delta \\ &= u_f(\widehat{\boldsymbol{x}}, j') + \delta. \end{aligned}$$

Hence, we have $\min_{j\in \mathrm{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) = u_l(\widehat{\boldsymbol{x}}, j^*)$, which can be bounded from below as

$$u_l(\widehat{\boldsymbol{x}}, j^*) \;=\; (1-\frac{\delta}{\Delta})u_l(\boldsymbol{x}^*, j^*) + \frac{\delta}{\Delta}u_l(\boldsymbol{x}^{j^*}, j^*) \;\geq\; (1-\frac{\delta}{\Delta})u_l(\boldsymbol{x}^*, j^*) \;=\; (1-\frac{\delta}{\Delta})u_{\mathrm{SSE}},$$

where we used $u_l(\boldsymbol{x}^{j^*}, j^*) \geq 0$. Since $u_{\mathrm{SSE}} \leq 1$, we have

$$u_{\mathrm{RSE}}(\delta) \geq \min_{j\in \mathrm{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) \geq (1-\delta/\Delta)u_{\mathrm{SSE}} \geq u_{\mathrm{SSE}} - \delta/\Delta.$$

Property 3. We demonstrate with examples in Appendix B.1.4 on how $u_{\mathrm{RSE}}(\delta)$ may be discontinuous when $\delta \geq \Delta$. In what follows we prove its Lipschitz continuity for $\delta \in (0, \Delta)$. Pick arbitrary $L > 1/\Delta$ and two arbitrary numbers δ and δ' such that $0 < \delta < \delta' \leq \Delta - 1/L$. We show that $|u_{\mathrm{RSE}}(\delta) - u_{\mathrm{RSE}}(\delta')| \leq L(\delta' - \delta)$ to complete the proof.

Let $(\boldsymbol{x}^*, j^*)$ be a δ-RSE. Pick arbitrary $\widetilde{j} \in \mathrm{argmax}_{j\in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_f(\boldsymbol{x}^*, j)$, and let $\widetilde{\boldsymbol{x}}$ be a strategy such that $u_f(\widetilde{\boldsymbol{x}}, \widetilde{j}) \geq u_f(\widetilde{\boldsymbol{x}}, j) + \Delta$ for all $j \neq \widetilde{j}$ (which exists according to the definition of the inducibility gap). We construct a leader strategy $\widehat{\boldsymbol{x}} = \frac{\Delta-\delta'}{\Delta-\delta}\boldsymbol{x}^* + \frac{\delta'-\delta}{\Delta-\delta}\widetilde{\boldsymbol{x}}$. We have $j \notin \mathrm{BR}_{\delta'}(\widehat{\boldsymbol{x}})$ if $j \notin \mathrm{BR}_\delta(\boldsymbol{x}^*)$ because

$$\begin{aligned}
u_f(\widehat{\boldsymbol{x}}, \widetilde{j}) &= \frac{\Delta-\delta'}{\Delta-\delta}u_f(\boldsymbol{x}^*, \widetilde{j}) + \frac{\delta'-\delta}{\Delta-\delta}u_f(\widetilde{\boldsymbol{x}}, \widetilde{j}) \\
&\geq \frac{\Delta-\delta'}{\Delta-\delta}\big(u_f(\boldsymbol{x}^*, j) + \delta\big) + \frac{\delta'-\delta}{\Delta-\delta}\big(u_f(\widetilde{\boldsymbol{x}}, j) + \Delta\big) \\
&= \frac{\Delta-\delta'}{\Delta-\delta}u_f(\boldsymbol{x}^*, j) + \frac{\delta'-\delta}{\Delta-\delta}u_f(\widetilde{\boldsymbol{x}}, j) + \delta' \\
&= u_f(\widehat{\boldsymbol{x}}, j) + \delta',
\end{aligned} \tag{3.8}$$

where $u_f(\boldsymbol{x}^*, \widetilde{j}) \geq u_f(\boldsymbol{x}^*, j) + \delta$ because $\widetilde{j} \in \mathrm{argmax}_{j\in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_f(\boldsymbol{x}^*, j)$, $j \notin \mathrm{BR}_\delta(\boldsymbol{x}^*)$.

Hence, $\mathrm{BR}_{\delta'}(\widehat{\boldsymbol{x}}) \subseteq \mathrm{BR}_{\delta}(\boldsymbol{x}^*)$ and

$$\begin{aligned} u_{\mathrm{RSE}}(\delta') = \min_{j \in \mathrm{BR}_{\delta'}(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) &= \min_{j \in \mathrm{BR}_{\delta'}(\widehat{\boldsymbol{x}})} \left(\frac{\Delta - \delta'}{\Delta - \delta} u_l(\boldsymbol{x}^*, j) + \frac{\delta' - \delta}{\Delta - \delta} u_l(\widetilde{\boldsymbol{x}}, j) \right) \\ &\geq \min_{j \in \mathrm{BR}_{\delta}(\boldsymbol{x}^*)} \frac{\Delta - \delta'}{\Delta - \delta} u_l(\boldsymbol{x}^*, j) = \frac{\Delta - \delta'}{\Delta - \delta} u_{\mathrm{RSE}}(\delta). \end{aligned}$$

This means that

$$u_{\mathrm{RSE}}(\delta) - u_{\mathrm{RSE}}(\delta') \leq \frac{\delta' - \delta}{\Delta - \delta} \cdot u_{\mathrm{RSE}}(\delta) \leq \frac{\delta' - \delta}{\Delta - \delta} \leq L(\delta' - \delta),$$

where the last inequality is due to $\delta \leq \Delta - 1/L$. Since u_{RSE} is non-increasing as we argued for Property (1), we then have $|u_{\mathrm{RSE}}(\delta) - u_{\mathrm{RSE}}(\delta')| = u_{\mathrm{RSE}}(\delta) - u_{\mathrm{RSE}}(\delta') \leq L(\delta' - \delta)$.

□

We conclude this section with a discussion on the convexity of $u_{\mathrm{RSE}}(\delta)$. Intuitively, one would conjecture that $u_{\mathrm{RSE}}(\delta)$ is convex non-increasing, as the leader's payoff should have diminishing margin as δ increases and follower chooses the worse responses. However, this intuition is only correct when $m = 2$. The function in general is neither convex nor concave, as illustrated in the plot in Figure 3.2.

Proposition 3.5 ((non-)convexity of $u_{\mathrm{RSE}}(\delta)$)**.** *When $m = 2$, $u_{\mathrm{RSE}}(\delta)$ is convex when $\delta \in (0, \Delta)$. When $m > 2$, there exist game instances where $u_{\mathrm{RSE}}(\delta)$ is neither convex nor concave, even when $\delta \in (0, \Delta)$.*

The proof hinges on the fact that, when $m = 2$, $u_{\mathrm{RSE}}(\delta)$ is the pointwise maximum of a set of linear functions, each of which corresponds to the optimal leader utility in a δ-optimal response region. We also construct an explicit example of neither convex nor concave $u_{\mathrm{RSE}}(\delta)$ with $m = 3$. See Appendix B.1.6 for the full proof of Proposition 3.5.

3.4 Computational Complexity of RSE

In this section, we study the complexity of computing and approximating a δ-RSE. We first show that NP-hardness of the computation problem in general and then propose a QPTAS algorithm to compute the approximated δ-RSE.

3.4.1 Hardness of Approximating δ-RSE

We start with the result that it is NP-hard, in general, to obtain an ϵ-optimal δ-RSE leader strategy. We remark that, though δ-RSE appears to be a natural solution concept, we are not aware of any previous results on its computational complexity. The closest result we can find is the NP-hardness of computing an optimal leader strategy that is robust with respect to uncertainty about the follower's utility matrix, studied by [Letchford et al., 2009]. Despite some similarity in spirit, these two problems can not be seen as special cases of each other. Moreover, from a technical point of view, our hardness result also sheds light on the inapproximability of the problem whereas the proof technique of [Letchford et al., 2009] only implies the hardness of exact computation and leaves inapproximability an open problem from their problem.

Theorem 3.6. *It is NP-hard to compute a $\frac{1}{2n}$-optimal δ-RSE leader strategy.*

Proof of Theorem 3.6. We show a reduction from the EXACT 3-SET COVER (X3C) problem. An X3C instance is given by an integer k, a collection of m subsets $S_1, \ldots, S_m \subseteq [3k]$, each of size 3. It is a yes-instance if there exists $J \subseteq [m]$, such that $|J| = k$ and $\bigcup_{j \in J} S_j = [3k]$; we call such a J an *exact cover*. Otherwise, it is a no-instance. We reduce an instance of X3C to a game with the following utility matrices. The leader has m actions to choose from, each corresponding to a subset in the X3C instance. The follower has $n = m + 3k + 1$ actions $\{a\} \cup \{b_j : j \in [m]\} \cup \{c_i : i \in [3k]\}$, where each b_j corresponds to subset S_j, and each c_i corresponds to an element in $[3k]$.

Suppose $\epsilon > 0$ is a constant, and let $\lambda = \frac{\epsilon}{6m \cdot k^2}$. The follower's utility function is given as follows (also see Figure 3.3 for an illustration).

- For all $\ell \in [m]$: $u_f(S_\ell, a) = 1$.
- For all $\ell \in [m]$ and $j \in [m]$:

$$u_f(S_\ell, b_j) = \begin{cases} \max\left\{1 - \frac{\delta}{1-\lambda}, \quad 0\right\}, & \text{if } j \neq \ell \\ \min\left\{1, \quad \frac{1-\delta}{\lambda}\right\}, & \text{if } j = \ell \end{cases} \tag{3.9}$$

  This ensures that $u_f(\boldsymbol{x}, b_j) \leq 1 - \delta$ whenever $x_j \leq \lambda$, and $1 - \delta < u_f(\boldsymbol{x}, b_j) \leq 1$, otherwise.

- For all $\ell \in [m]$ and $i \in [3k]$:

$$u_f(S_\ell, c_i) = \begin{cases} \min\left\{(1-\delta) \cdot \frac{k}{k-1+\lambda \cdot k}, \quad 1\right\}, & \text{if } i \notin S_\ell \\ \max\left\{0, \quad 1 - \delta \cdot \frac{k}{1-\lambda \cdot k}\right\}, & \text{if } i \in S_\ell \end{cases} \tag{3.10}$$

  This ensures that $1 - \delta < u_f(\boldsymbol{x}, c_i) \leq 1$ whenever $\sum_{\ell : i \in S_\ell} x_\ell < \frac{1}{k} - \lambda$, and $u_f(\boldsymbol{x}, c_i) \leq 1 - \delta$ otherwise.

The leader's utility function is given as follows.

- For all $\ell \in [m]$:

$$u_l(S_\ell, a) = \frac{1}{k}, \quad \text{and} \quad u_l(S_\ell, c_i) = 0 \quad \text{for all } i \in [3k] \tag{3.11}$$

- For all $\ell \in [m]$ and $j \in [m]$:

$$u_l(S_\ell, b_j) = \begin{cases} 0 & \text{if } j \neq \ell \\ 1 & \text{if } j = \ell \end{cases} \tag{3.12}$$

We show that if the X3C instance is a yes-instance, then the leader obtains utility $\frac{1}{k}$ in a robust Stackelberg equilibrium; otherwise, the leader obtains at most $\frac{1}{2k} \cdot (1+\epsilon)$. Hence, no $\frac{1}{2k} \cdot (1-\epsilon)$-optimal algorithm exists unless P = NP. Intuitively, to obtain the higher

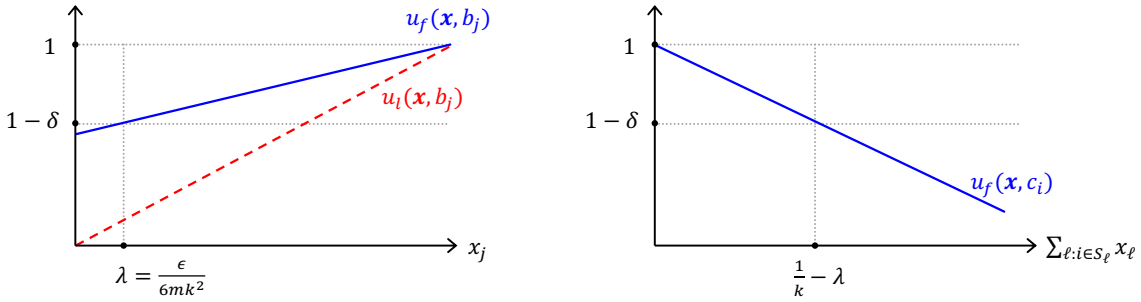


Figure 3.3: Utility Functions of Constructed Instances for the Reduction of Theorem 3.6.

utility of $\frac{1}{k}$, the leader needs to prevent the follower's actions c_i from being a δ-optimal response. According to the utility definition, this requires choosing actions S_ℓ that cover i with a sufficiently high probability close to $1/k$; see Figure 3.3 (right). On the other hand, choosing each S_ℓ comes with a price as it will cause b_ℓ to be a δ-optimal response if the probability reaches λ; see Figure 3.3 (left). Hence, in order to maintain utility $\frac{1}{k}$ for the leader, for each S_ℓ, we should either pick it with probability close to zero (i.e., $< \lambda$), or we pick it with probability at least $\frac{1}{k}$. This is analogous to a discrete choice of S_ℓ as in the X3C. More specifically, the reduction proceeds as follows.

First, suppose that the X3C instance is a yes-instance and J is an exact cover of this instance. Consider the following leader strategy $\boldsymbol{x} = (x_j)_{j\in[m]}$, whereby the leader plays each pure strategy S_j with probability: $x_j = \frac{1}{k}$ if $j \in J$, and $x_j = 0$ if $j \notin J$. The follower's utility for responding to $\boldsymbol{x}$ with each pure strategy is as follows:

- Clearly, $u_f(\boldsymbol{x}, a) = 1$ and $u_l(\boldsymbol{x}, a) = \frac{1}{k}$.
- For each b_j, according to (3.9):

- If $j \in J$, we have $x_j = \frac{1}{k} > \lambda$ and hence, $u_f(\boldsymbol{x}, b_j) > 1-\delta$; Meanwhile, $u_l(\boldsymbol{x}, b_j) = \frac{1}{k}$ according to (3.12).
- If $j \notin J$, we have $x_j = 0 < \lambda$ and hence, $u_f(\boldsymbol{x}, b_j) < 1 - \delta$.

- For each c_i, we have $i \in S_j$ for some $j \in J$ since J is an exact cover. Hence, $\sum_{\ell: i \in S_\ell} x_\ell \geq \frac{1}{k} > \frac{1}{k} - \lambda$, and we have $u_f(\boldsymbol{x}, c_i) < 1 - \delta$.

As a result, $\mathrm{BR}_\delta(\boldsymbol{x}) = \{a\} \cup \{b_j : j \in J\}$, and $\min_{j' \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j') = \frac{1}{k}$.

Conversely, suppose that by playing some strategy $\boldsymbol{x}$, the leader obtains utility at least $\frac{1}{2k} \cdot (1 + \epsilon)$. We show that the instance must be a yes-instance. In this case, we have $\min_{y \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, y) \geq \frac{1}{2k} \cdot (1+\epsilon)$. According to the definition of the leader's utility function, this implies:

- $c_i \notin \mathrm{BR}_\delta(\boldsymbol{x})$ for all $i \in [3k]$. Hence, we have

$$\sum_{\ell: i \in S_\ell} x_\ell \geq \frac{1}{k} - \lambda. \tag{3.13}$$

  Since $|S_\ell| = 3$, we have $\sum_{i \in [3k]} \sum_{\ell: i \in S_\ell} x_\ell = 3 \sum_{\ell \in [m]} x_\ell = 3$, so it holds for all $i \in [3k]$ that:

$$\sum_{\ell: i \in S_\ell} x_\ell = 3 - \sum_{i' \in [3k] \setminus \{i\}} \sum_{\ell: i' \in S_\ell} x_\ell < \frac{1}{k} + 3k \cdot \lambda \leq \frac{1}{k} \cdot (1 + \epsilon/2). \tag{3.14}$$

- If $b_j \in \mathrm{BR}_\delta(\boldsymbol{x})$, then it must be that $x_j \geq \frac{1}{2k} \cdot (1 + \epsilon)$. Hence, for each $j \in [m]$, either $x_j \geq \frac{1}{2k} \cdot (1 + \epsilon)$, or $x_j \leq \lambda$ (as implied by $b_j \notin \mathrm{BR}_\delta(\boldsymbol{x})$). This further implies that for each $i \in [3k]$, there is exactly one ℓ with $i \in S_\ell$ and $x_\ell \geq \frac{1}{2k} \cdot (1 + \epsilon)$: the existence of two or more such ℓ would violate (3.14); on the other hand, if no such ℓ exists, we would have $\sum_{\ell: i \in S_\ell} x_\ell \leq m \cdot \lambda < \frac{1}{k} - \lambda$, which violates (3.13). It follows that, for this

ℓ, we have

$$\begin{aligned} x_\ell &= \sum_{\ell' : i \in S_{\ell'}} x_{\ell'} - \sum_{\ell' : \ell' \neq \ell \text{ and } i \in S_{\ell'}} x_{\ell'} \\ &\geq \sum_{\ell' : i \in S_{\ell'}} x_{\ell'} - (m-1) \cdot \lambda \quad \geq \quad 1/k - m \cdot \lambda \quad > \quad 1/k - \epsilon/k^2, \end{aligned}$$

which implies that the set $J := \left\{ \ell \in [m] : x_\ell \geq \frac{1}{2k} \cdot (1+\epsilon) \right\}$ has at most k element in it: otherwise, $\sum_{\ell \in J} x_\ell > (k+1) \cdot (1/k - \epsilon/k^2) > 1$. Therefore, J is an exact cover, and the X3C instance is a yes-instance.

Therefore, we have shown that no $\frac{1}{2k} \cdot (1-\epsilon)$-optimal algorithm exists unless P = NP. Since $n > 3k$, it is also NP-hard to compute $\frac{1-\epsilon}{n}$-optimal δ-RSE leader strategy for any $\epsilon \in (0, 1]$. This completes the proof. □

Theorem 3.6 shows that it is NP-hard in general to approximate a δ-RSE. However, as a corollary of Property (1) in Theorem 3.4, we show that there exists an efficient algorithm to compute $\frac{\delta}{\Delta}$-optimal δ-RSE leader strategy. This intuitively illustrates that the difficult instances of finding δ-RSE are those with small inducibility gap Δ but large robustness requirement δ. We refer readers to the detailed algorithm of this corollary in Appendix B.2.1

Corollary 3.7. *For Stackelberg games with inducibility gap $\Delta > \delta$, there exists an algorithm that computes $\frac{\delta}{\Delta}$-optimal δ-RSE leader strategy in $O(n)$ time.*

3.4.2 A QPTAS for δ-RSE

We now move to develop a quasi-polynomial time approximation scheme (QPTAS) for computing an approximate δ-RSE for any given δ. This leaves an intriguing open problem to close the gap between the efficiency of this algorithm and the above inapproximability result — specifically, to understand whether a PTAS exists for δ-RSE or whether Theorem 3.6 can

be strengthened to the hardness of obtaining a constant additive approximation (possibly under some assumption like exponential time hypothesis as used by Rubinstein [2016] to rule out PTAS for Nash equilibrium). Our preliminary investigation suggests that either direction seems to require significantly different ideas from our current techniques.

Theorem 3.8. *For any $\epsilon > 0$, we can compute an ϵ-optimal δ-RSE leader strategy in quasi-polynomial time $O\big(m^{\lceil \frac{\log 2n}{2\epsilon^2} \rceil} n \log n\big)$.*

Before presenting the formal proof, we briefly overview the high-level idea. Our algorithm starts with a probabilistic argument similar to [Lipton et al., 2003] for arguing the existence of an approximate Nash equilibrium with a simple format termed the *k-uniform strategy*. Formally, a mixed strategy $\boldsymbol{x} \in \Delta^m$ is called k-uniform for some integer k if every $x_i = k_i/k$ for some integer $k_i \leq k$. Lipton et al. [2003] prove that in any two-player $m \times m$ matrix game there always exists a pair of k-uniform strategies for $k = \frac{12 \ln m}{\epsilon^2}$ that is an ϵ-Nash equilibrium. Consequently, to find an ϵ-Nash, they only need to exhaustively search all possible k uniform mixed strategy pairs. Unfortunately, searching for an ϵ-optimal δ-RSE turns out to require significantly more work due to the bi-level nature of our problem. In fact, the ϵ-optimal δ-RSE is even not a k-uniform strategy in general. This is because while any k-uniform leader strategy $\overline{\boldsymbol{x}}$ and a nearby strategy $\boldsymbol{x}$ will lead to similar leader utilities, they may lead to different set $\mathrm{BR}_\delta(\boldsymbol{x})$ of follower δ-optimal response actions which in turn affects the induced leader utility. Consequently, the major algorithmic part of our proof is to efficiently search, through a carefully crafted binary search procedure, the entire *nearby convex region* of each uniform strategy $\overline{\boldsymbol{x}}$ in order to identify a leader strategy $\boldsymbol{x}$ that induces the most favorable follower response action. Details are presented in the following proof.

Proof of Theorem 3.8. Let $\mathcal{G}_k \subseteq \Delta^m$ denote the set of all k-uniform mixed strategies for the leader (who has m actions). Note that there are $O(m^k)$ many k-uniform strategies[6].

6. The total number can be computed as dividing k items into m parts, while each part can have zero item.

The following lemma is originally from Althöfer [1994] and later used by Lipton et al. [2003] for computing approximate Nash equilibrium. It can be proved via the probabilistic method [Alon and Spencer, 2016].

Lemma 3.9 (Althöfer [1994], Lipton et al. [2003])**.** *Let $A \subseteq [0,1]^{m\times n}$ be the leader's payoff matrix. For any $\epsilon > 0$ and any leader strategy $\boldsymbol{x} \in \Delta^m$, there exists a k-uniform strategy $\overline{\boldsymbol{x}} \in \mathcal{G}_k$ with $k = \lceil \frac{\log 2n}{2\epsilon^2} \rceil$ such that*

$$|u_l(\boldsymbol{x}, j) - u_l(\overline{\boldsymbol{x}}, j)| \le \epsilon \quad \text{for all } j = 1, \cdots, n.$$

We now use Lemma 3.9 to construct subspaces of the leader's strategy space. Specifically, for each $\overline{\boldsymbol{x}} \in \mathcal{G}_k$, we construct $\Delta^{\overline{\boldsymbol{x}}} \subseteq \Delta^m$ such that:

$$\Delta^{\overline{\boldsymbol{x}}} = \left\{ \boldsymbol{x} \middle| \boldsymbol{x} \in \Delta^m \text{ and } |u_l(\boldsymbol{x}, j) - u_l(\overline{\boldsymbol{x}}, j)| \le \epsilon \text{ for all } j = 1, \cdots, n \right\}$$

Note that each $\Delta^{\overline{\boldsymbol{x}}}$ is a convex region since it is defined by a set of linear constraints by writing $|u_l(\boldsymbol{x}, j) - u_l(\overline{\boldsymbol{x}}, j)| \le \epsilon$ as $u_l(\boldsymbol{x}, j) - u_l(\overline{\boldsymbol{x}}, j) \le \epsilon$ and $-u_l(\boldsymbol{x}, j) + u_l(\overline{\boldsymbol{x}}, j) \le \epsilon$. Moreover, $\bigcup_{\overline{\boldsymbol{x}} \in \mathcal{G}_k} \Delta^{\overline{\boldsymbol{x}}} = \Delta^m$, because Lemma 3.9 implies that any $\boldsymbol{x} \in \Delta^m$ belongs to some $\Delta^{\overline{\boldsymbol{x}}}$.

The key to our proof is to compute an approximately optimal δ-RSE leader strategy within each convex region $\Delta^{\overline{\boldsymbol{x}}}$. Note that, fixing any follower response action j, the mixed strategies $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ gives rise to roughly the same leader utility, up to at most ϵ difference by Lemma 3.9. However, this does not imply that they are equally good since different mixed strategies may lead to different sets $\mathrm{BR}_\delta(\boldsymbol{x})$ of δ-optimal follower responses, which in turn induces different leader utilities. So we need to search for the $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ to maximize the worst (over possible follower responses) leader utility, or formally to solve the following

$$\text{optimization problem within } \Delta^{\overline{\boldsymbol{x}}}: \qquad \max_{\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}} \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \tag{3.15}$$

Unfortunately, Problem (3.15) is generally intractable since the feasible region of inside min depends on $\boldsymbol{x}$. We instead solve the following more tractable variant

$$\text{surrogate of Problem (3.15)}: \quad \max_{\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}} \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\overline{\boldsymbol{x}}, j) \tag{3.16}$$

which substitutes $u_l(\boldsymbol{x}, j)$ in Problem (3.15) by $u_l(\overline{\boldsymbol{x}}, j)$. Observe that any optimal solution to Problem (3.16) must be an ϵ-optimal solution to Problem (3.15) because their objective function differs by at most ϵ due to Lemma 3.9 and our restriction of $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$.

What is nice about Problem (3.16) is that its objective function only directly depends on j (whose choice then depends on $\boldsymbol{x}$). This allows us to design the Algorithm 3.1 that can efficiently search for the best $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$. We prove its correctness in the following lemma.

ALGORITHM 3.1: Utility-Verification

Input : Leader strategy $\overline{\boldsymbol{x}} \in \mathcal{G}_k$, its corresponding $\Delta^{\overline{\boldsymbol{x}}}$, and target utility μ.
Output : If $\exists\, \boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ such that $\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\overline{\boldsymbol{x}}, j) \geq \mu$, output `True` and such an $\boldsymbol{x}$; else output `False`.

$Q \leftarrow \varnothing$;
for *every follower action* $j \in [n]$ **do**
  if $u_l(\overline{\boldsymbol{x}}, j) < \mu$ **then**
    $Q \leftarrow Q \cup \{j\}$
for *every follower action* $j \in [n]$ *and* $j \notin Q$ **do**
  Determine if the following Linear Program is feasible:

$$\begin{aligned} \exists \quad & \boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}} \\ \text{s.t.} \quad & u_f(\boldsymbol{x}, j) \geq u_f(\boldsymbol{x}, j'), \forall j' \in [n] \\ & u_f(\boldsymbol{x}, j) \geq u_f(\boldsymbol{x}, j') + \delta, \forall j' \in Q \end{aligned} \tag{3.17}$$

  if *the above linear feasibility problem is feasible for some* j **then**
    return `True` and any feasible solution $\boldsymbol{x}$ of that problem.
return `False`.

Lemma 3.10. *For any* $\mu \in [0, 1]$*, there is a polynomial time algorithm that asserts whether the optimal objective of Problem* (3.16) *is larger than* μ *or not, and in the former case outputs an* $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ *that achieves* $\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\overline{\boldsymbol{x}}, j) \geq \mu$.

Proof. The details of this algorithm are presented in Algorithm 3.1. At a high level, we first identify all "bad" follower actions j's that cannot satisfy our request, i.e., $u_l(\overline{\boldsymbol{x}}, j) < \mu$, and group them into set Q. We then try to see whether there exists an $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ such that its follower δ-optimal response set does not contain any bad actions, i.e., $\mathrm{BR}_\delta(\boldsymbol{x}) \cap Q = \varnothing$. This later question reduces to a series of linear feasibility problems, each for a $j \notin Q$ (the Program (3.17)) deciding whether there exists a $\boldsymbol{x} \in \Delta^{\overline{\boldsymbol{x}}}$ under which the j is the best follower action and the follower's utility from j is at least δ larger than his utility from any $j' \in Q$. □

Armed with Lemma 3.10, we can use binary search to find an $\boldsymbol{x}$ that exactly solves Problem (3.16) after $\log(1/n)$ rounds since we know the problem only has n possible values of μ (i.e. $u_l(\overline{\boldsymbol{x}}, 1), \cdots, u_l(\overline{\boldsymbol{x}}, n)$). This solution will be ϵ-optimal for Problem (3.15). We do this for the $O(m^k)$ possible k-uniform strategies and output the strategy with the largest objective. This is a ϵ-optimal δ-RSE. □

3.5 Statistical Complexity of RSE

In this section, we turn to another complexity study of RSE, i.e., the sample efficiency of learning an ϵ-optimal δ-RSE without initially knowing the leader's or follower's utility matrix. Motivated by the recent work of learning the strong Stackelberg equilibrium (SSE) by Bai et al. [2021], here we extend it to an online learning problem of the δ-RSE. Similar to [Bai et al., 2021], the learner cannot directly observe the mean reward matrix of a Stackelberg game $u_l, u_f \in \mathbb{R}^{m \times n}$, but has to learn to approximate the δ-RSE from the noisy bandit feedback. The motivation of this learning paradigm is from a common multi-agent learning practice today of "centralized training, decentralized execution" [Lowe et al., 2017]. That is, in many robotics and game-playing applications (see, e.g., OpenAI Gym [Brockman et al., 2016]), the learning environments are well-defined such that the game parameters can be learned in a centralized fashion by controlling agents' action profiles. Thus, the agents can learn to estimate game parameters from their noisy feedback, and then deploy the learned

strategies in the decentralized environment to play against unknown opponents.

We describe the learning setting in Definition 4 and start this section by presenting a sample-efficient learning algorithm that can learn an approximate δ-RSE with a utility guarantee. Notably, as a corollary, our sample complexity result strictly strengthens that of [Bai et al., 2021], with both improved *utility guarantee* and better *computational efficiency* for non-degenerate Stackelberg games (i.e., $\Delta > 0$) [7].

Definition 4 (Learning δ-RSE from Bandit Feedback [Bai et al., 2021])**.** *At each round, the learner can query an action pair (i, j) and observe noisy bandit feedback, $r_l(i, j) = u_l(i, j) + \xi, r_f(i, j) = u_f(i, j) + \xi'$, where ξ, ξ' are i.i.d. zero-mean noises with finite variances.*

Theorem 3.11. *There exists a learning algorithm that can learn an approximated δ-RSE of any Stackelberg game $(u_l, u_f) \in \mathbb{R}^{m\times n}$ with leader's utility at least as much as $u_{\text{RSE}}(\delta + 4\epsilon) - 2\epsilon$ using $O(mn \log(mn/\iota)/\epsilon^2)$ samples with probability at least $1 - \iota$.*

Proof of Theorem 3.11. We prove its existence by explicitly constructing the learning algorithm. It starts with the following sampling procedure:

Play each action pair (i, j) for $T = \frac{1}{2\epsilon^2} \log(\frac{2mn}{\iota})$ rounds to get the mean reward estimation $\widetilde{u}_l(i, j) = \frac{1}{T} \sum_{t=1}^{T} r_l^t(i, j), \widetilde{u}_f(i, j) = \frac{1}{T} \sum_{t=1}^{T} r_f^t(i, j)$. According to the concentration inequality, both the estimation $\widetilde{u}_l(i, j), \widetilde{u}_f(i, j)$ have the error bound of ϵ with probability $1 - \frac{\iota}{mn}$. Thus, by union bound, with probability $1 - \iota$, the utility estimation $\widetilde{u}_l, \widetilde{u}_f$ satisfies $\|\widetilde{u}_l - u_l\|_\infty \leq \epsilon, \left\|\widetilde{u}_f - u_f\right\|_\infty \leq \epsilon$. The sample complexity in total is $\frac{mn}{2\epsilon^2} \log(\frac{2mn}{\iota}) = O(mn \log(mn/\iota)/\epsilon^2)$.

Notice that, for some leader strategy $\boldsymbol{x}$, the δ-best response set under u_f can be different from that under $\widetilde{u}_f$, which could result in constant the utility gap. Hence, we set the benchmark to $(\delta + 2\epsilon)$-RSE in Lemma 3.12.

7. Interestingly, the algorithm for learning the mixed strategy SSE in [Bai et al., 2021] happens to be solving an approximate δ-RSE. Quoting the authors' own words in their paper, "*it is unclear whether this program (the δ-RSE problem) can be reformulated to be solved efficiently in polynomial time*". Our Theorem 3.6 confirms that their program indeed is NP-hard.

Lemma 3.12. *For any* $\left\|\widetilde{u}_f - u_f\right\|_\infty \leq \epsilon$*, for any* $\boldsymbol{x} \in \Delta^m$*,* $V(\boldsymbol{x}; u_l, u_f, \delta) \geq V(\boldsymbol{x}; u_l, \widetilde{u}_f, \delta + 2\epsilon)$ *and thus,* $V^*(u_l, u_f, \delta) \geq V^*(u_l, \widetilde{u}_f, \delta + 2\epsilon)$

Here, we denote $V(\boldsymbol{x}; u_l, u_f, \delta)$ as the leader utility of strategy $\boldsymbol{x}$ against δ-rational follower (who takes action from δ-optima response set) in Stackelberg game u_l, u_f. Let $V^*(u_l, u_f, \delta)$ be the δ-RSE of Stackelberg game u_l, u_f. Let (x, j^*) be the $(\delta + 2\epsilon)$-RSE of the Stackelberg game $\widetilde{u}_l, \widetilde{u}_f$. We claim the learning algorithm could simply output (x, j^*) as the approximated δ-RSE of Stackelberg game (u_l, u_f). The remaining proof is to prove the following series of inequalities:

$$\begin{aligned} &V(\boldsymbol{x}; u_l, u_f, \delta) \geq V(\boldsymbol{x}; u_l, \widetilde{u}_f, \delta + 2\epsilon) \geq V(\boldsymbol{x}; \widetilde{u}_l, \widetilde{u}_f, \delta + 2\epsilon) - \epsilon \\ =& V^*(\widetilde{u}_l, \widetilde{u}_f, \delta + 2\epsilon) - \epsilon \geq V^*(\widetilde{u}_l, u_f, \delta + 4\epsilon) - \epsilon \geq V^*(u_l, u_f, \delta + 4\epsilon) - 2\epsilon. \end{aligned}$$

There are four inequalities in the above arguments. The first and third inequality is by the following Lemma 3.12. The equality is by the construction of $\boldsymbol{x}$. The second and last inequality is based on the fact that, for any $\|\widetilde{u} - u\|_\infty \leq \epsilon$, for any $\boldsymbol{x} \in \Delta^m$, $\widetilde{u}(\boldsymbol{x}, j) - u(\boldsymbol{x}, j) = \boldsymbol{x}(\widetilde{u} - u)e_j \in [-\epsilon, \epsilon]$. This concludes our main proof. We defer the proof of Lemma 3.12 to Appendix B.3.1.

□

We make a few remarks on Theorem 3.11. First, we note that the above learning algorithm is sample-efficient but not computationally efficient, since it requires solving the exact $(\delta + 2\epsilon)$-RSE of a Stackelberg game, which we already know is NP-hard to compute in general. However, it is possible to employ the QPTAS Algorithm 3.1 to find an ϵ-optimal δ-RSE according to Theorem 3.8, and this would not change the order of our learning algorithm's approximation ratio.

Second, Theorem 3.11 combined with Theorem 3.4 implies the following corollary about the efficient learning of an approximated SSE. Specifically, leveraging the convergence prop-

erty of δ-RSE, the following Corollary 3.13 strengthens the SSE learning results in the recent work by Bai et al. [2021], in terms of providing better utility guarantee and computationally efficient learning algorithms. Specifically, [Bai et al., 2021] states that under the same sample complexity, a learning algorithm can only learn the u_{SSE} up to $u_{\text{RSE}}(\delta)$ while the gap between them can be arbitrarily large. But our result suggests, as long as the game instance is not degenerated (e.g., with inducibility $\Delta > 0$), the gap can be bounded so that SSE can be efficiently learned up to ϵ utility loss. Moreover, as highlighted by [Bai et al., 2021], their learning algorithm does not have a computational efficiency guarantee and may take exponential time in general, since there is no efficient algorithm to compute the approximated SSE with pessimistic follower tie-breaking. But our result implies that, as long as the game instance is non-degenerate with inducibility $\Delta > \epsilon$, we can efficiently compute ϵ-RSE according to Corollary 3.7.

Corollary 3.13 (Efficient Learning of SSE)**.** *For any Stackelberg game with inducibility $\Delta > 0$, an ϵ-optimal SSE can be learned from $O(\frac{1}{\epsilon^2})$ samples in polynomial time for any $\epsilon < \Delta$.*

Thirdly, we point out that this learning result is almost tight due to the Proposition 3.14 below, where we present the hard instances that fundamentally limit the learnability of δ-RSE. Specifically, in the case when $\Delta < \delta$, the $u_{\text{RSE}}(\delta + 4\epsilon)$ can be arbitrarily worse than $u_{\text{RSE}}(\delta)$ due to the discontinuity of the $u_{\text{RSE}}(\delta)$ function. This is an intrinsic barrier and is further explained by the $\Omega(1)$ utility gap below. Meanwhile, by Theorem 3.4, when $\Delta \geq \delta + 1/L$ for some constant L, the Lipschitz continuity of $u_{\text{RSE}}(\delta)$ implies that $u_{\text{RSE}}(\delta + 4\epsilon)$ can be substituted by $u_{\text{RSE}}(\delta) - O(\epsilon)$ so that the sample complexity dependence on ϵ matches with the $\Omega(1/\sqrt{T})$ of the lower bound instance. See Appendix B.3.2 for the proof of Proposition 3.14.

Proposition 3.14. *For any sample size T, there exists a Stackelberg game instance with inducibility gap Δ such that any algorithm with T samples in learning δ-RSE is $\Omega(1/\sqrt{T})$*

suboptimal if $\delta < \Delta$*, or* $\Omega(1)$ *suboptimal if* $\delta \geq \Delta$*. More specifically, any output leader strategy* $\widehat{\boldsymbol{x}}$ *satisfies the following with probability at least* $\frac{1}{3}$*:*

$$\min_{j \in \mathrm{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) \leq \begin{cases} u_{\mathrm{RSE}}(\delta) - \frac{1/\sqrt{T}}{\Delta - \delta + 1/\sqrt{T}} & \delta < \Delta \\ u_{\mathrm{RSE}}(\delta) - 1/2 & \delta \geq \Delta \end{cases}. \tag{3.18}$$

3.6 Final Remarks

In practice, there are many situations where a follower fails to make the optimal decision in the Stackelberg game. However, the classic solution concept of SSE is not robust for suboptimal follower responses and leads to poor performance of the leader. In this paper, we systematically study a robust variant of Stackelberg equilibrium to account for suboptimal responses from the follower. We propose a well-defined definition of robust Stackelberg equilibrium, δ-RSE, and show some nice properties of the leader's utility under a δ-RSE. We identify the computational complexity for computing or approximating the δ-RSE. We show there does not exist an efficient algorithm for finding an approximate δ-RSE, unless P=NP, while we also propose a QPTAS to compute the ϵ-optimal δ-RSE for any $\epsilon > 0$. Lastly, we provide sample complexity results for learning the δ-RSE when the follower utility is not known initially.

Our results open up possibilities for many other interesting questions. For example, our positive and negative computational results in Section 3.4 have a small gap due to the logarithmic exponent term in the computational complexity of the QPTAS. An immediate direction for future research is to close this gap by either showing a PTAS algorithm or strengthening the hardness result of inapproximability to a constant factor. In addition, our study concerns the most basic normal-form game setups. It is worth understanding the applicability of this robust solution concept to various applications of leader-follower or principal-agent game models, including pricing games [Myerson, 1981, Devanur et al., 2014],

security games [Tambe, 2011], Bayesian persuasion [Kamenica and Gentzkow, 2011b] and contract design [Grossman and Hart, 1992]. Future work can also study analogous robust solution concepts for more general Stackelberg game models such as these with Bayesian follower types [Conitzer and Sandholm, 2006] or with constrained follower strategy sets [Wald, 1945, Goktas and Greenwald, 2021]. Lastly, we would also like to study the learnability of RSEs in different feedback models, e.g., when the learner cannot observe the follower payoffs [Blum et al., 2014, Balcan et al., 2014, Letchford et al., 2009, Peng et al., 2019].

Part III

Feedback Alignment

CHAPTER 4

ISOTONIC MECHANISM FOR PEER REVIEW

4.1 Introduction

In recent years, major AI and machine learning conferences such as NeurIPS, ICML and ICLR have faced a concerning decline in the quality of peer review, posing a significant challenge to the global machine learning community [Langford and Guzdial, 2015, Brezis and Birukou, 2020, Tomkins et al., 2017, Lipton and Steinhardt, 2019]. In particular, the NeurIPS 2014 experiment showed that 49.5% of the papers accepted by one committee would be rejected by another [Lawrence and Cortes, 2014, Langford and Guzdial, 2015]. This inconsistency probability was 50.6% for NeurIPS 2021 [Cortes and Lawrence, 2021]. This troubling trend is, in part, due to the unbalance between the surge in submission volumes and lagged growth of the number of qualified reviewers [Shah, 2022]. As demonstrated by Figure 4.1, many conferences in recent years have been handling submissions at a scale of 10,000 full papers, an almost ten-fold growth in a decade, whereas a large portion of the reviewers are graduate students, demanded to review around 5 papers in a month for each conference [Shah et al., 2018, Stelmakh et al., 2021, Russo, 2021]. Similar trends are seen across various research communities [McCook, 2006, Lajtha and Baveye, 2010, Gropp et al., 2017, Checco et al., 2021], highlighting a systemic issue in the today's peer review systems.

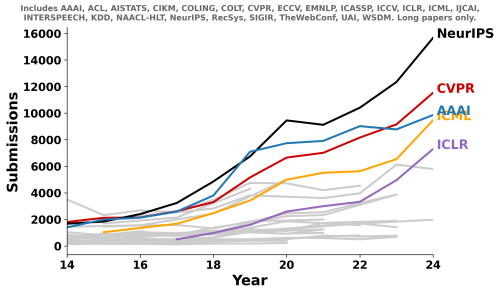


Figure 4.1: The number of paper submissions in major AI and machine learning conferences from 2014 to 2024.

To mitigate this issue, there have been progressive research efforts to improve peer review processes. Some approaches employ machine learning and optimization techniques to improve reviewer assignments, reduce bias, and automate review procedures, while others adopt the economic approach to model the incentives of participants to encourage high-quality reviews (see, e.g., a survey by Shah [2022]). Combining both the statistical and economic perspectives, we propose a mechanism that incentivizes the authors to share evaluations of their own papers to calibrate peer review scores. The conceptual idea is natural: if the reviewer pool cannot supply sufficient information for all the submissions, conference organizers could resort to other sources — in our case, the authors themselves. However, while self-criticism is a merit in science, self-evaluation has been largely overlooked due to the obvious conflicts of interest. That is, authors may be reluctant to provide honest assessments, if revealing a mediocre rating of their own work puts them at disadvantage. Consequently, estimations from manipulated data may be hardly useful. This presents a fundamental design challenge:

> *How can statistical estimation processes effectively elicit data from strategic actors? In particular, is it possible to balance statistical efficiency and incentive compatibility?*

This paper examines the above question in the critical domain of peer review and embarks on a path to bring the mechanism design perspective to statistical estimation and proposes an approach, termed "Isotonic Mechanism", that demonstrates a definitive solution to this design challenge.

A key challenge to begin with, even putting incentive issues aside, is to determine what kind of data could be elicited from authors with reasonable accuracy. One approach might be to ask for highly fine-grained data, such as the authors' own scores for their papers. However, even if authors are completely honest, they may lack the precise knowledge needed to provide reliable scores. Alternatively, one could request more general information (e.g.,

asking authors to identify their favorite paper), but this may not offer sufficient value for improving review scores. The challenge lies in finding an effective middle ground—eliciting data that is both accurate and useful. To this end, we propose focusing on the *ranking* of an author's papers. Such relative information tends to be less noisy than absolute measures and has been successfully applied in various fields, such as learning from human preferences [Yue and Joachims, 2009, Bai et al., 2022].

Next comes our core scientific question: is it possible to truthfully elicit ranking data from authors to improve the statistical efficiency of review scores? Our starting point is an earlier work [Su, 2021] which showcased the possibility in a simplified "research world" with only a single researcher, say, Alice. To formally describe this method, suppose Alice as the only author of the system submits n papers to the conference peer review. Right after submission, Alice provides a ranking π in the form of a permutation of $1, 2, \ldots, n$ that sorts her papers in descending order of quality. Letting $y_1, \ldots, y_n$ denote the (average) review scores of the n papers, the mechanism yields adjusted review scores that are the solution to the following convex optimization program:

$$\begin{aligned} \min_{\boldsymbol{r} \in \mathbb{R}^n} \ & \sum_{i=1}^{n} (y_i - r_i)^2 \\ \text{s.t.} \ & r_{\pi(1)} \geq r_{\pi(2)} \geq \cdots \geq r_{\pi(n)}. \end{aligned}$$

This optimization yields the well-known isotonic regression [Barlow and Brunk, 1972], and its solution ensures consistency with the author-provided ranking while remains as close as possible to the original review scores in the least square sense. Under natural statistical modeling of the review process, this approach provably improve the score estimation. More interestingly, it also guarantees truthful elicitation of Alice's ranking, under standard utility and prior knowledge assumptions. Notably, the later is a non-trivial property. Specifically, it crucially hinges on geometric properties of isotonic regressions, and would not hold for

other estimation methods (e.g., switching from l_2-norm above to l_1-norm). This illustrates the importance of designing appropriate statistical estimation methods that can align with incentives.

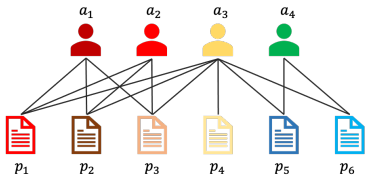


Figure 4.2: An example of an author-paper ownership relation shown as a bipartite graph. An edge between an individual and a paper indicates that this individual is the author of the paper.

This work seeks to address the more realistic situation in academic conferences where many authors submit many papers, with overlapping authorship. This strict generalization exhibits multiple new challenges. First, since Alice's co-authors will also submit rankings for (some of) Alice's papers, Alice would have to account for the affect of her co-authors' data on the utility of Alice. This intricacy is due to the fundamental difference between single-agent and multi-agent decision making setup, hence also lead to our different solution concepts, generalizing from Bayesian optimal decision to Bayes Nash equilibrium. Second, an even more challenging issue is that different papers can *partially* overlap in authorship, as illustrated in Figure 4.2. This is an especially common situation in today's popular machine learning conferences. In such cases, Alice's utility may be affected by other authors' rankings on papers Alice did not contribute. Naive adaptation of the previous Isotonic Mechanism designed for single-author situations can lead to serious incentive issues. For instance, one natural approach is to apply the Isotonic Mechanism for every author's submissions and output final paper scores by averaging their adjusted review scores. Unfortunately, it is easy to find undesirable examples with untruthful behaviors under this mechanism (see Section 4.6.2) because some coauthors would misreport their papers' ranking information in order to improve their own utilities. Therefore, a universal "truth serum" is needed to effectively incorporate self-evaluations from a complex network of overlapping authorship

into the reviewing mechanism.

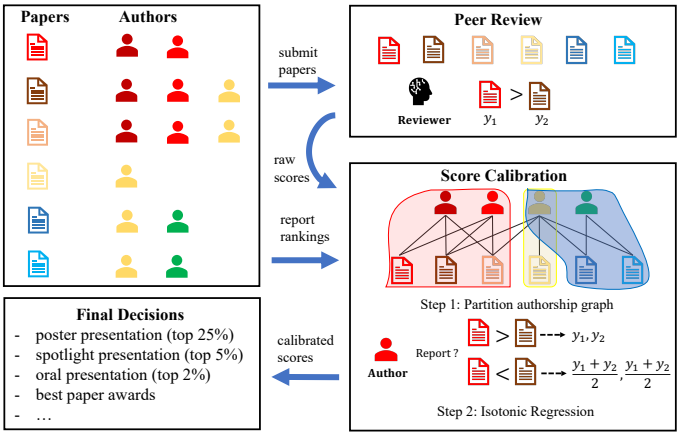


Figure 4.3: An illustration of the conference reviewing procedure assisted by the proposed Isotonic Mechanism.

To address these challenges, this paper develops an approach that advances the earlier design with insights from both algorithm design and mechanism design. In Figure 4.3, we illustrate key procedures of the proposed mechanism. Specifically, the mechanism first resorts to an algorithmic preprocessing step that partitions the author-paper ownership sets into blocks of papers such that each block shares a common set of authors. It then uses the ranking information elicited from the common authors in each block to calibrate the review scores of their papers. Under certain conditions, this proposed mechanism is marked by the following key characteristic:

Main Result 1. *Truthfulness forms a Nash equilibrium under the proposed mechanism.*

Along with other results, the main result of truthfulness property is formally presented in Section 4.6. This result significantly extends the previous truthfulness guarantee of the earlier work [Su, 2021] to cases with overlapping ownership and suggests that the optimal strategy for each author to maximize her utility is to report the ground-truth ranking, given that all the other authors do likewise. In particular, our proof of this new result involves a novel technique concerning majorization in the ordering of paper scores with respect to

both the reviewers' noises and other authors' reports. Moreover, we show the mechanism encourages truthfulness behaviors in several ways. From a game-theoretical perspective, this Nash equilibrium of truthfulness is the most favored, as any other potential Nash equilibrium results in no better utility for any of the authors. From empirical studies, the mechanism induces a common-interest game and the human subjects are found to always prefer the truthful behaviors for the payoff-dominant outcome.

A potential criticism to the above partition-based approach is that we have to give up the authors' ranking information across blocks. One may wonder whether the Isotonic Mechanism can be carefully tailored to elicit additional comparison information among items across partitioned blocks as such information will help to increase the efficiency of isotonic regression — so long as it can be truthfully elicited. Unfortunately, our second main result gives a negative answer.

Main Result 2. *Within a significantly more general class of isotonic-regression-based mechanisms, any truthful calibration mechanism has to be partition-based (for some choice of item partitions), thus cannot use any ordering information about items across the partitioned blocks.*

This result is formally presented in Section 4.7.1. It illustrates the fundamental *trade-off* between the statistical efficiency and incentive requirements in designing truthful mechanisms for peer review systems. Hence the only room left for us to optimize the statistical efficiency of our designed mechanism, while retaining its truthfulness, is to optimize the choice of the item partition; this leads to our third main result. As is conceivable, the choice of partition affects the statistical performance of our mechanism. For example, a partition for Figure 4.2 can be simply $\{p_1, p_2, p_3, p_4, p_5\}$, which share a single common author $\{a_3\}$, or can be $\{p_1, p_2, p_3\}, \{p_4\}$, and $\{p_5, p_6\}$, which share $\{a_1, a_2, a_3\}, \{a_3\}$, and $\{a_3, a_4\}$, respectively. To find high-quality partitions of the author-paper ownership relation, we define a natural *class* of criteria to assess the quality of a partition. Within this class of criteria, we obtain

the following result on this proposed mechanism:

Main Result 3. *A simple nearly linear-time greedy algorithm can obtain a partition for the proposed mechanism that is near-optimal simultaneously for* every *criterion in the class.*

This result is formally stated in Section 4.7.3. We also show that it is NP-hard to find the exact optimal partition even for some simple criterion. In contrast, our greedy algorithm is computationally efficient while achieving constant-ratio approximation *simultaneously* for all the criteria in our considered class. Such simultaneous approximation is a surprisingly nice property of our problem and is a rare phenomenon generally. [1]

Finally, we design a series of experiments to investigate the empirical performance of our mechanism on both ICLR and ICML review data in Section 4.8. Our results show that the mechanism consistently enhances the estimation of review scores across various experimental settings, even beyond the technical assumptions of our theoretical framework. We conclude with a detailed discussion on the real-world applicability of our proposed mechanism in Section 4.9.

4.2 Problem Formulation

This paper considers the review score calibration problem under a network of overlapping ownership.[2] There are m owners and n items. The j-th owner owns a set of items $\mathcal{I}^j \subseteq [n] = \{1, 2, \ldots, n\}$ and let $|\mathcal{I}^j| = n_j$. [3] For each item $i \in [n]$, let R_i be its *ground-truth score*

1. We are aware of only one other situation in approximated majorization for minimizing symmetric convex objectives [Goel and Meyerson, 2006], where there is a simultaneous approximation guarantee but their optimization problem is very different from ours and their ratio is logarithmic in general. This powerful result is recently employed by Banerjee et al. [2023] for welfare guarantee in information design.

2. Besides the motivating example of peer review, calibration problems of this kind exist in general crowdsourcing problems. Hence, for the rest of this paper on the general model, we will use the generic terms, "owner/item", instead of "author/paper".

3. The ownership relation forms a bipartite graph and we formalize their connections in our proofs in Appendix C.5.

and

$$y_i = R_i + z_i$$

be the *raw review score* given by its reviewers with noise z_i. As a friendly reminder of our notation, i and $\mathcal{I}$ are used to denote *item* index and *item* set, respectively, whereas j is used to index the *owner*. Denote by $\boldsymbol{z} = (z_1, z_2, \ldots, z_n), \boldsymbol{R} = (R_1, R_2, \ldots, R_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, respectively, the vector of noise, the ground-truth scores, and the raw review scores.

In applications such as conference paper reviews, the *ground-truth scores* $\boldsymbol{R}$ can be interpreted as the mean evaluation scores of the papers perceived across all experts in the community. This item evaluation score is mathematically well defined and always exists,[4] but it is very difficult to be accurately observed during the review phase for various reasons including insufficient resources to have enough reviews and insufficient time to truly observe the items' impact. This is essentially the motivation of peer review, during which this ground-truth score will be estimated by a few experts sampled from the community. The averaged review scores of these experts form our modeling of the *raw review scores* $\boldsymbol{y}$ — a noisy estimation of ground-truth scores $\boldsymbol{R}$. Therefore, the high-level objective of a calibration mechanism $\mathcal{M}$ is to output the *adjusted review scores* $\widehat{\boldsymbol{R}}$ that provide a more accurate estimation of the ground truth scores $\boldsymbol{R}$. Notably, such a calibration mechanism is naturally compatible with existing peer review systems, making it easier for practitioners to adopt.

To tackle the review score calibration problem with the limited supply of qualified reviewers, we resorts to the design of *owner-assisted calibration mechanisms* that elicit and utilize self-evaluation data from the item owners. We abstract this mechanism design prob-

4. We remark that paper evaluation score should be carefully distinguished from "true merit" of the paper, which is a vague concept and whose existence might be arguable. Unlike merits, a paper's true evaluation score is well-defined, and is what peer review procedure as well as statistical methods like ours is trying to estimate.

lem into the construction of an estimator $\widehat{\mathrm{R}}_{\mathcal{M}}$ for $\boldsymbol{R}$ that combines the raw scores $\boldsymbol{y}$ with the information elicited from owners $\{\pi^j\}_{j=1}^m$ to improve the estimates of $\boldsymbol{R}$. In this paper, we focus on eliciting each owner's ranking of their items, where π^j is a permutation of $1, 2, \ldots, n_j$ that specifies an ordering of the j-th owner's papers in $\mathcal{I}^j$. The benefits of this design choice and the possible relaxations are discussed in Section 4.9. The owner-assisted calibration mechanism then proceeds in the following stages.

1. *Before Submission:* Each owner is notified of $\mathcal{M}$ and observes some information of her items.

2. *During Submission:* Each owner j is asked to report a ranking π^j of her items $\mathcal{I}^j$. [5]

3. *After Submission:* The mechanism outputs adjusted review scores as $\widehat{\boldsymbol{R}} = \widehat{\mathrm{R}}_{\mathcal{M}}(\{\pi^j\}_{j=1}^m; \boldsymbol{y})$.

The key challenge of this design problem is to balance the tension between the statistical efficiency and incentive compatibility. To formalize this problem, we adopt the mechanism design approach in modeling the strategic incentives of item owners on the calibration results. That is, we assume each owner has a utility function on their items' adjusted review scores. Let owner j's utility function be $\mathrm{U}^j : \mathbb{R}^{n_j} \to \mathbb{R}$ such that each owner j derives utility $\mathrm{U}^j\big([\widehat{R}_i]_{i\in\mathcal{I}^j}\big)$ from adjusted review scores $\widehat{\boldsymbol{R}}$, output by the calibration mechanism. We employ the Bayesian game framework [Harsanyi, 1967] to study the owners' strategic decision-making at the information elicitation stage during submission, where the realization of the raw review scores $\boldsymbol{y}$ is uncertain to the owners (e.g., due to the unknown reviewer assignments and noise). We say a profile of owners' report $\{\pi^j\}_{j=1}^m$ forms a Bayes-Nash Equilibrium (BNE) [6] under mechanism $\mathcal{M}$ if for any owner $j \in [m]$, given others' report

5. If an owner does not report a ranking, a uniformly random ranking will be used. As will be shown later, this design ensures that rational owners participate in the mechanism. It is also without loss of generality to assume that owners report the full ranking of their items, and we defer the discussion to Remark 4.2.1.

6. It suffices to consider pure-strategy NEs in most part of this paper and we defer the additional notations for the more general definition of mixed-strategy NEs to Appendix 10.

$\pi^{-j} = \{\pi^{j'}\}_{j' \neq j}$, j's expected utility by reporting π^j is no worse than reporting any other possible ranking $\widetilde{\pi}^j$ or, mathematically,

$$\mathop{\mathbf{E}}_{\boldsymbol{y}}\left[\mathrm{U}^j\left(\widehat{\mathrm{R}}_{\mathcal{M}}(\pi^j, \pi^{-j}; \boldsymbol{y})\right)\right] \geq \mathop{\mathbf{E}}_{\boldsymbol{y}}\left[\mathrm{U}^j\left(\widehat{\mathrm{R}}_{\mathcal{M}}(\widetilde{\pi}^j, \pi^{-j}; \boldsymbol{y})\right)\right], \quad \forall j \in [m]. \tag{Equilibrium Condition}$$

We say a mechanism $\mathcal{M}$ is *truthful* if every owner j reporting the true ranking π^j of her items forms a Bayes-Nash equilibrium under $\mathcal{M}$. Through a standard revelation principle argument [Nisan and Ronen, 1999], it is without loss of generality to restrict the design space to truthful mechanisms. To demonstrate the theoretical guarantees of our mechanism, we make the following assumptions that are natural in peer review applications. We also conduct empirical studies that demonstrate the robust performance of our mechanism for settings even beyond these assumptions in Section 4.8.

Assumption 4.3 (Informed Owners)**.** *For each j, the j-th owner has sufficient knowledge of the ground-truth score $\boldsymbol{R}$ to determine the true ranking of her own items $\mathcal{I}^j$.*

Assumption 4.4 (Exchangeable Noise Distribution)**.** *The review noise vector $\boldsymbol{z} = (z_1, \ldots, z_n)$ follows an exchangeable distribution in the sense that $(z_1, \ldots, z_n)$ has the same probability distribution in $\mathbb{R}^n$ as in $\rho \circ \boldsymbol{z} := (z_{\rho(1)}, \ldots, z_{\rho(n)})$ for any permutation ρ of $1, 2, \ldots, n$.*

Assumption 4.5 (Convex Utility)**.** *For each j, the j-th owner's utility function takes the form of $\mathrm{U}^j(\widehat{\boldsymbol{R}}) = \sum_{i \in \mathcal{I}^j} U^j(\widehat{R}_i)$, where $U^j : \mathbb{R} \to \mathbb{R}$ is a non-decreasing convex function.*

Assumption 4.3 is standard in mechanism design and reflects each owner's private knowledge about their items.[7] Assumption 4.4 imposes symmetry on review noises, and Assumption 4.5 captures a utility structure that naturally arises in "high-risk-high-reward" tourna-

7. For instance, in basic models of auction design, bidders are assumed to perfectly know their value about the item [Nisan and Ronen, 1999]. Our assumption of authors knowing the ranking information is also inspired by many machine learning applications that learn parameters from humans' ranking/comparison data which are posited to be more accurate than their knowledge of absolute values [Yue and Joachims, 2009, Bai et al., 2022].

ment settings — we will discuss the root of this assumption, its adoption in similar economic problems, and empirical evidences from ICLR data in Section 4.9.

4.6 An Isotonic Mechanism for Completely Overlapping Ownership

We begin with a rank-calibrated score estimator $\widehat{\mathrm{R}}(\pi; \boldsymbol{y})$ that employs isotonic regression on raw review scores $\boldsymbol{y}$ and any reported ranking π from the owner, via the following convex program[8],

$$\widehat{\mathrm{R}}(\pi; \boldsymbol{y}) = \underset{\boldsymbol{r} \in \mathbb{R}^n}{\operatorname{argmin}} \ \|\boldsymbol{y} - \boldsymbol{r}\|^2$$
$$\text{s.t. } r_{\pi(1)} \geq r_{\pi(2)} \geq \cdots \geq r_{\pi(n)}.$$

In the general setup with overlapping ownership, we would like to design an mechanism that aggregates information from as many owners as possible, while preserving the desirable properties of truthfulness and statistical efficiency. The most natural design is perhaps to use the (weighted) averaged scores from the estimates based on each owner's reported ranking. The mechanism takes the input of a problem instance, specified by review scores $\boldsymbol{y} \in \mathbb{R}^n$ and reviewer credentials $\{\alpha^j\}_{j=1}^m \in [0,1]^n$, and outputs the adjusted review score $\widehat{\boldsymbol{R}}$ using rankings elicited from all owners. Notably, the reviewer credential is a set of weights $\{\alpha^j\}_{j=1}^m$ that pre-specifies the different levels of influence of the owners; we view them as part of the problem instance that reflects each reviewer's expert level, reputation scores and track records in the given instance. In practice, one could simply set $\alpha^j = 1/m$ to evenly weigh on each owner's reported ranking, or set personalized α^j for each owner j with $\sum_j \alpha^j = 1$ to account for different reviewer's expert level — we include it as a part of the input so

8. $\|\cdot\|$ denotes the ℓ_2 norm throughout this paper.

that it is a choice to the practitioner's discretion. We formally describe these procedures in Mechanism 4.1. The rest of this section is to analyze the situations when it does or does not work.

Mechanism 4.1: Isotonic Mechanism under Completely Overlapping Ownership

Input: Review scores $\boldsymbol{y} \in \mathbb{R}^n$, reviewer credentials $\{\alpha^j\}_{j=1}^m$.

1 **for** *every* $j \in [m] = \{1, 2, \ldots, m\}$ **do**

2 Elicit ranking π^j from owner j.

3 Solve for ranking-calibrated scores $\widehat{\boldsymbol{R}}^j \leftarrow \widehat{\mathrm{R}}(\pi^j; \boldsymbol{y})$.

4 **return** $\widehat{\boldsymbol{R}} = \sum_{j=1}^m \alpha^j \widehat{\boldsymbol{R}}^j$.

4.6.1 Truthfulness under Completely Overlapping Ownership

We start by considering a useful special case in which every submission is owned by every owner, referred to the *completely overlapping ownership* situation. As it turns out, truth-telling forms a Nash equilibrium in this setting under Mechanism 4.1. [9] Moreover, this equilibrium is *payoff dominant* [Harsanyi et al., 1988], i.e., one that is Pareto superior to all other Nash equilibria in the game. Put simply, all owners would prefer this equilibrium, because it simultaneously gives every owner the highest equilibrium utility among all possible equilibrium outcomes. This nice property makes it more plausible to expect agents' truthful behaviors, despite the potential existence of multiple equilibria in the game; more discussions about the truthfulness of this mechanism from the perspective of behavioral theory can be found in Remark 4.1.1.

Theorem 4.1. *Under Assumptions 4.3, 4.4, and 4.5, if the ownership is completely overlapping,*

1. *It forms a Bayes-Nash equilibrium for each owner to truthfully report the ranking in Mechanism 4.1.*

9. This result can be related to the ex-post incentive compatibility — a common goal in mechanism design.

2. This equilibrium of truthful report is payoff dominant.

We defer the proof of Theorem 4.1 to Appendix C.1 and conclude this subsection with additional evidence on the truthfulness of the Isotonic Mechanism according to the behavioral game theory.

Remark 4.1.1 (Truthfulness of Isotonic Mechanism from a behavioral angle)**.** *Mechanism 4.1 induces the strategic game with special structures known as the* common interest games *[Harsanyi et al., 1988, Bacharach, 2018]. We briefly highlight the connections here and discuss evidence from both empirical human preferences and behavioral theory on why truthful behavior should be expected in such games. In its simplest form, the payoff matrix of a common interest game typically takes one of the three structures illustrated in Table 4.1 (only orders of the payoff values matter). Namely, there is an action profile (the upper left cell of each game) that represents the common interest of both players; this corresponds to the truthful action profile* $(\pi^\star, \pi^\star)$ *under the Isotonic Mechanism which forms a payoff-dominant Nash equilibrium, according to Theorem 4.1. Meanwhile, the three variants of common interest games capture the different possible orders of payoff values in the lower right cell and those in the off-diagonal cells. Any variant of these payoff matrices may be realized under the Isotonic Mechanism, since our proof using Jensen's inequality only offers upper bounds on the utilities of these non-truthful action profiles (see e.g., Equation* (C.2)*).*

	$\pi^\star$	π°
$\pi^\star$	$10, 11$	$4, 5$
π°	$4, 5$	$3, 4$

	$\pi^\star$	π°
$\pi^\star$	$10, 11$	$1, 1$
π°	$1, 1$	$3, 4$

	$\pi^\star$	π°
$\pi^\star$	$10, 11$	$1, 5$
π°	$5, 1$	$3, 4$

Table 4.1: An illustration of three typical common interest game payoff matrices.

It turns out that in experimental games and also in real life, people tend to always choose the upper left cell — the payoff dominant *outcomes — in all the three possible game structures above [Bacharach, 2018]. This behavior is expected under the first payoff matrix since* $(\pi^\star, \pi^\star)$ *is the unique Nash equilibrium there. It is somewhat reasonable under the second*

payoff matrix, as argued by Harsanyi et al. [1988] for its "self-reinforcing property" — everyone thinks the other has no reason to prefer the payoff-dominated equilibrium outcomes. However, realizing the payoff dominant outcome becomes less clear under the third payoff matrix, because (π°, π°) *forms a "risk-dominant" Nash equilibrium — i.e., compared to* $\pi^\star$*,* π° *is the less risky action under uncertainty of the opponent's choice [Harsanyi et al., 1988] (specifically, under* (π°, π°) *equilibrium, row player gets* 5 *instead of* 3 *if the opponent deviates from the equilibrium action* π° *to* $\pi^\star$*). Interestingly, standard game theory cannot account for this phenomenon of coordination on the payoff dominant outcome under the second and third payoff matrices, i.e., the Hi-Lo Paradox. One major theory that explains this phenomenon is the* team reasoning *by Bacharach [2018] as a non-game-theoretic rationale, i.e., "What do we want? And what should I do to play my part in achieving this?". If there is common knowledge that both players adopt the team-reasoning mode of choosing their strategies, then both would choose the payoff dominant equilibrium. These also serve as behavioral evidences of agents' truthfulness in Isotonic Mechanism.*

4.6.2 Non-truthfulness Beyond Completely Overlapping Ownership

Despite the nice properties of the Nash equilibrium shown in Theorem 4.1, the condition of completely overlapping ownership is usually not satisfied in many application scenarios — different papers are often written by different sets of authors. To extend Mechanism 4.1 to the case of incompletely overlapping ownership, one natural choice is to average the scores according to the ownership, $\widehat{R}_i \leftarrow \sum_{j=1}^{m} e_i^j \widehat{R}_i^j / \sum_{j'=1}^{m} e_i^{j'}$, where the binary ownership indicator $e_i^j = 1$ if and only if owner j owns item i. The question is whether truthful reporting still forms a Nash equilibrium. Unfortunately, the answer turns out to be "No". Specifically, the owners may gain from the misreporting strategy and one possible manipulation is to use the reputation of good items to promote a bad item — it may only hurt the good item a little, but help the bad item a lot. We demonstrate such kind of strategic manipulations via

a concrete example below.

Example 3 (Non-truthfulness under Partially Overlapping Ownership)**.** *Consider a case of m owners and $n = 3$ items. The items have ground-truth scores $R_1 = 9, R_2 = 8, R_3 = 4$ — one weak and two strong. The ownership is not completely overlapping. As illustrated in Table 4.2, the first $m-1$ owners work together on the two strong items $1, 2$, while the m-th owner works on the item $2, 3$. For each j, let the j-th owner's utility be $\mathrm{U}^j(\widehat{\boldsymbol{R}}) = \sum_{i \in \mathcal{I}^j} \max\{R_i - 5, 0\}$.*

Item \ Owner	1	2	$\cdots$	$m-1$	m	True Scores
1	1	1	$\cdots$	1	0	$R_1 = 9$
2	1	1	$\cdots$	1	1	$R_2 = 8$
3	0	0	$\cdots$	0	1	$R_3 = 4$

Table 4.2: An illustration of the ownership matrix $E = (e_i^j)_{m \times n}$ and ground-truth scores in Example 3.

For simplicity, suppose the reviews are noiseless so that, if all owners report their ranking truthfully, the calibration is perfect, i.e., $\widehat{\boldsymbol{R}} = \boldsymbol{y} = \boldsymbol{R}$. However, we can observe that the m-th owner does not have the incentive to report the ranking of items $1, 2$ truthfully, given that the first $m-1$ owners report the true ranking. That is, under the flipped ranking that $\widetilde{R}_2^m \geq \widetilde{R}_3^m$, Mechanism 4.1 would have $\widetilde{R}_2^m = \widetilde{R}_3^m = 6$ and the adjusted scores $\widetilde{R}_1 = 9, \widetilde{R}_2 = \frac{1}{m}(\widetilde{R}_2^m + \sum_{j=1}^{m-1} R_2^j) = 8 - \frac{2}{m}, \widetilde{R}_3 = \widetilde{R}_3^m = 6$. We can see that the utility of m-th owner under truthful ranking is, $3 < 3 - \frac{2}{m} + 1$, strictly worse than that under the non-truthful ranking for any $m \geq 3$.

At a high level, such strategic behavior is due to the owner's uneven influence on the average scores of its different items — we will formalize this intuition in Theorem 4.3. One may speculate that we can potentially resolve this kind of situation through a careful reweighing process of each owner's influence. The answer also turns out to be "No". Specifically, in the example above, m-th owner, as the sole owner of item 1, would always fully determine the

score of item 1; yet, this owner can never fully determine the score of item 2 unless we choose to ignore other owners' opinion or assign some weight to the raw review scores, which more or less defeats the purpose of this community effort of calibrating review scores. Hence, we need to seek a different approach to help us truthfully aggregate the owners' information in practice.

4.7 Restoring Truthfulness via Partitioning

To ensure the truthfulness beyond complete ownership, this section studies a partition-based approach. Our main idea is to partition the ownership sets into multiple blocks, ensuring complete overlap within each block by some owners. We then individually apply Mechanism 4.1 to each block, eliciting truthful rankings only from the owners who completely own that block, as illustrated in Figure 4.4. Finally, we use the elicited rankings from these blocks to estimate the ground-truth scores. We formally described this procedure in the following Mechanism 4.2.

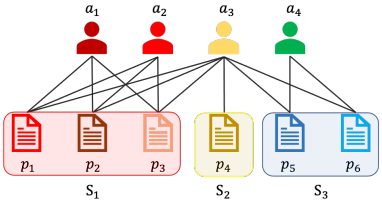


Figure 4.4: A partition of partially overlapping ownership.

Mechanism 4.2: Partition-based Isotonic Mechanism for Partially Overlapping Ownership

Parameters : Item set partition $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ based on ownership relations $\{\mathcal{I}^j\}_{j=1}^m$.
Input: Review score $\boldsymbol{y} \in \mathbb{R}^n$, reviewer credentials $\{\alpha^j\}_{j=1}^m$.

for *every* $\mathcal{S}_k \in \mathcal{S}$ **do**
  Find all owners with complete ownership of $\mathcal{S}_k$, i.e., $\mathcal{T}_k \leftarrow \{j \in [m] : \mathcal{S}_k \subseteq \mathcal{I}^j\}$.
  if $|\mathcal{T}_k| \neq 0$ *and* $|\mathcal{S}_k| > 1$ **then**
    Apply Mechanism 4.1 to item set $\mathcal{S}_k$ with owner set $\mathcal{T}_k$ and weights $\{\alpha^j\}_{j \in \mathcal{T}_k}$ to estimate the ground-truth score of items in $\mathcal{S}_k$, denoted by $\widehat{\boldsymbol{R}}[\mathcal{S}_k]$.
return $\widehat{\boldsymbol{R}} = \{\widehat{\boldsymbol{R}}[\mathcal{S}_k]\}_{k=1}^K$.

Compared to Mechanism 4.1, Mechanism 4.2 additionally determines a set of parameters which include a partition of item set $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ that satisfies $\bigcup_{k=1}^{K} \mathcal{S}_k = [n], \mathcal{S}_k \cap \mathcal{S}_{k'} = \varnothing$. We ask the parameters to only depend on the ownership relations for practical considerations: since the mechanism may be truthful only under some proper choice of parameters, the owners should be informed of the parameters during submission, when the ownership relations are formed but the review scores are not realized. One can observe that in Mechanism 4.2, each owner $j \in \mathcal{T}_k$ has the complete ownership of the items in $\mathcal{S}_k$ by construction. As a corollary of Theorem 4.1, it is easy to show that Mechanism 4.2 is truthful under any partition $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ and any weights $\{\{\beta_k^j\}_{k=1}^K\}_{j=1}^m$, because we only elicit rankings of items within each block from those who completely own the block. Thus, all owners $j \in \mathcal{T}_k$ will report truthful ranking over items in $\mathcal{S}_k \subseteq \mathcal{I}^j$. Owing to the independence of score estimates for items between different partition blocks, the overall mechanism is truthful; we formalize it in the following corollary.

Corollary 4.2. *Mechanism 4.2 is truthful for any input instance and parameter choice in the following sense: it forms a Bayes-Nash equilibrium for each owner to truthfully report the ranking of their items within each block specified by the partition.*

Remark 4.2.1 (Practical Implementation of Mechanism 4.2)**.** *While Mechanism 4.2 only elicits the ranking of items within each block $\mathcal{S}_k$, in practical implementations, the mechanism designer may simply ask owners to rank all their items but commit to only use part of the ranking of each owner as specified by the mechanism. Under such a committed mechanism, any owner will be indifferent about the order of any two items in any two different partition sets $\mathcal{S}_k$ and $\mathcal{S}_{k'}$ since their order will never be used by the mechanism. Thus, the owner can be assumed to break ties in favor of the designer and reveal their full ranking truthfully.*[10]

10. A formal argument for such tie-breaking is by adding a negligible amount of randomness for picking an arbitrarily different partition, which consequently creates a negligible amount of incentive for any owner to report other rankings truthfully — known as equilibrium refinement via randomization in mechanism design [Nisan and Ronen, 1999].

Notably, however, the designer's commitment to not using any order information beyond what the mechanism specified (despite such information being elicited) is important, since otherwise, truthfulness will not hold. In reality, the mechanism designer as a trusted authority (e.g., the organizers of a large ML conference) usually has such commitment power.

4.7.1 The Necessity of Partition-based Isotonic Mechanisms

At this point, one may wonder whether restricting ranking elicitation to partitioned item sets is necessary, as it inevitably forces the mechanism to ignore some of the authors' revealed information across the partition blocks. For example, consider an instance with $m = n = 3, \mathcal{I}^1 = \{1, 2\}, \mathcal{I}^2 = \{2, 3\}, \mathcal{I}^3 = \{1, 3\}$ — i.e., each owner owns two items, and each item has two co-owners (see the illustration in Table 4.3). We can see that, with any partition of the item set, Mechanism 4.2 can elicit ranking information from at most one owner. However, a more aggressive design could be eliciting the ranking information from every owner j about her items in $\mathcal{I}^j$ and then apply a generalized version of Mechanism 4.1 by using the elicited (partial) ranking of j's items. This is certainly a statistically more efficient design, but the key question is whether a mechanism of such kind can still be guaranteed to be truthful. Our study next shows that the answer is unfortunately "No". We show that, within a much broader class of isotonic-regression-based mechanisms, the partition-based mechanism as prescribed in Mechanism 4.2 is essentially the only candidate that can guarantee truthful owner behaviors — that is, one may have to give up eliciting the comparison information between the partitioned blocks in order to trade for incentive properties. This reveals an intrinsic tradeoff between statistical efficiency and incentive guarantee within the general class of isotonic mechanisms. We leave it as an intriguing open direction to explore alternative estimation methods other than isotonic regression that can balance statistical efficiency and incentive compatibility — no such method is known so far even for the single-owner case.

We start by generalizing Mechanism 4.2 to a broader class of isotonic mechanisms, as

Item \ Owner	1	2	3
1	1	0	1
2	1	1	0
3	0	1	1

Item \ Owner	1	2	3
1	1	0	1
2	1	1	0
3	0	1	1

Table 4.3: An illustration of partition-based mechanism (left) and a more general mechanism (right). In the partition-based mechanism, each color denotes a partitioned item set block. In the more general mechanism, different owners can have different item set partitions, as marked by different colors.

described in Mechanism 4.3. In this generalized isotonic mechanism, the calibrated score of any item $i \in [n]$ is now allowed to depend on the information elicited from all its owners, denoted by $\mathcal{J}^i \subseteq [m]$. Concretely, the input to Mechanism 4.2 and Mechanism 4.3 is the same, but the parameter choices in Mechanism 4.3 are strictly more general: (1) the item partition $\mathfrak{S}^j = \{\mathcal{S}_1^j, \mathcal{S}_2^j, \dots \mathcal{S}_{K^j}^j\}$, satisfying $\bigcup_{k=1}^{K^j} \mathcal{S}_k^j = \mathcal{I}^j, \mathcal{S}_k^j \cap \mathcal{S}_{k'}^j = \varnothing$, is now allowed to be different across owners; (2) additionally, a weight vector $\boldsymbol{\beta}^j = [\beta_1^j, \beta_2^j, \dots, \beta_n^j] \in \mathbb{R}_{\geq 0}^n$ is introduced to allow owner j's more fine-grained influence on her items with itemized weights. This is much more powerful than j's influence in Mechanism 4.2 that is restricted to be the same among items. The parameters are also determined from the input problem instance, specifically the ownership relation $\{\mathcal{I}^j\}_{j=1}^m$. The adjusted score for each item is then similarly determined by a weighted linear combination of the rank-calibrated scores from each owner. Notably, parameter β_i^j for any $i \notin \mathcal{I}^j$ is never used by Mechanism 4.3 hence can be arbitrary. We nevertheless defined $\boldsymbol{\beta}^j$ as a vector in $\mathbb{R}_{\geq 0}^n$ mainly for notational convenience.

Mechanism 4.3: Generalized Isotonic Mechanism

Parameters : Personalized item partition $\mathfrak{S}^j = \{\mathcal{S}_1^j, \dots, \mathcal{S}_{K^j}^j\}$ and itemized weights $\boldsymbol{\beta}^j$ for each owner $j \in [m]$ based on ownership relations $\{\mathcal{I}^j\}_{j=1}^m$.

Input: Review scores $\boldsymbol{y} \in \mathbb{R}^n$, reviewer credentials $\{\alpha^j\}_{j=1}^m$.

for *every* $j \in [m]$ **do**

  Apply Mechanism 4.1 to owner j with partition $\mathfrak{S}^j$ to estimate the ground-truth score, denoted by $\widehat{R}_i^j$, for each item i in $\mathcal{I}^j$.

return $\left\{\widehat{R}_i = \sum_{j \in \mathcal{J}^i} \alpha^j \beta_i^j \widehat{R}_i^j / \sum_{j' \in \mathcal{J}^i} \alpha^{j'} \beta_i^{j'}\right\}_{i=1}^n$.

It is easy to see that the class of Mechanism 4.3 strictly contains the class of Mechanism

4.2. On the one hand, it can be immediately verified that Mechanism 4.2 with any item set partition $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ can be formulated as Mechanism 4.3 with some parameter $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ (see the proof of Theorem 4.3 for the details). On the other hand, Mechanism 4.3 does have strictly more expressivity than Mechanism 4.2. For example, the non-partition-based mechanism illustrated above in Table 4.3 can be captured by the following elicitation rule: $\mathfrak{S}^1 = \{\{1,2\}\}, \boldsymbol{\beta}^1 = [1,1,0], \mathfrak{S}^2 = \{\{2,3\}\}, \boldsymbol{\beta}^2 = [0,1,1]$ and so on. One may now wonder whether this strictly more general Mechanism 4.3 is also strictly more powerful. Our next result shows that the answer is "NO" if the mechanism is truthful.

Theorem 4.3 (The Necessity of Partition)**.** *Under Assumptions 4.3, 4.4, and 4.5, for any $\mathcal{M}$ in the format of Mechanism 4.3 that is truthful for every input, there exists a $\mathcal{M}'$ in the format of Mechanism 4.2 that elicits no less ranking information. Formally, if the order of any two items i, i' is truthfully elicited by $\mathcal{M}$, then their order is also truthfully elicited by $\mathcal{M}'$.*

The formal proof is somewhat involved hence deferred to Appendix C.4. At a high level, our proof is structured with two main steps. First, we characterize the necessary conditions under which Mechanism 4.3 is truthful, which is base on the following key Lemma 4.4. Second, we show that for any Mechanism 4.3 under the necessary condition, a Mechanism 4.2 can be constructed to elicit as least as much ranking information.

Lemma 4.4. *Under Assumptions 4.3, 4.4, and 4.5, if a Mechanism 4.3 with parameters $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ is truthful, then the following two conditions must both hold:*

(I) *Each owner has balanced influence on the items within each partition blocks for any input, i.e.,*

$$\text{for any } j \in [m], \mathcal{S} \in \mathfrak{S}^j, i, i' \in \mathcal{S}, \qquad \text{we have } \omega_i^j = \omega_{i'}^j,$$

where $\omega_i^j = \alpha^j \beta_i^j / \sum_{j' \in \mathcal{J}^i} \alpha^{j'} \beta_i^{j'}$ denotes owner j's relative influence on the score of

item i.

(II) *The parameters $\{\mathfrak{S}^j, \boldsymbol{\beta}^j\}_{j=1}^m$ has a valid partition structure in the following sense,*

for any $j, j' \in [m], \mathcal{S} \in \mathfrak{S}^j, \mathcal{S}' \in \mathfrak{S}^{j'}$, if $\mathcal{S} \cap \mathcal{S}' \neq \varnothing$, then $\beta_i^j = \beta_{i'}^j$, for all $i, i' \in \mathcal{S} \cup \mathcal{S}'$.

Moreover, these two conditions are equivalent to each other.

4.7.2 Partition Optimization and its Hardness

As illustrated above, any partition will make Mechanism 4.2 truthful and that partition-based isotonic mechanisms are the only truthful mechanisms within a much broader class of isotonic mechanisms. These together point to the last key piece of our design — identifying the item partition that can elicit the most information from owners to refine our score estimation. Generally, smaller the size of a block is, the less ranking information is contained in the block.[11] However, on the other hand, longer blocks usually have less joint owners since it is less usual for multiple owners to jointly own many items. Thus the pursuit of longer blocks will prevent us from eliciting rankings from multiple authors. To address this tradeoff, we formalize two different metrics for identifying partitions which, respectively, determine the performance and robustness of the resulting mechanism. We study their intrinsic tradeoffs and implications to the computational complexity of partition optimization.

Informativeness (as optimization objective). In regard to the performance, we aim to identify partitions that are most informative for the review score calibration. On one hand, larger-sized blocks generally contains more ranking information, whereas in the extreme case if a block contains only one item, the mechanism cannot elicit any ranking information within the block. On the other hand, the selection of a large block may sometimes render

11. For instance, the ranking of a size-10 item block has $\frac{10\times 9}{2} = 45$ pairs of comparison, whereas the rankings of two size-5 item blocks have $2 \times \frac{5\times 4}{2} = 20$ pairs of comparison.

other blocks small, leading to complex tradeoff about partition choices. To systemically evaluate the coverage quality of a partition $\mathfrak{S}$, we introduce an *informativeness* objective function, $\text{obj}(\mathfrak{S}) = \sum_{k=1}^{K} w(|\mathcal{S}_k|)$, determined by some *wellness function* w on the size of each partition block. For examples, consider the following two choices of the wellness function $w(\cdot)$. One is $w(x) = x^2$ as a simple quadratic function of block size, which relates to the number of pairwise comparisons within a block:

$$\textbf{comparison-focused objective:} \quad \text{obj}(\mathfrak{S}) = \sum_{k=1}^{K} |\mathcal{S}_k|^2. \tag{4.1}$$

Another choice is $w(x) = \max\{x-1, 0\}$, which is called *size-focused* objective, defined below:

$$\textbf{size-focused objective:} \qquad \text{obj}(\mathfrak{S}) = \sum_{k=1}^{K} \max\{|\mathcal{S}_k| - 1, 0\} = \sum_{k=1}^{K} [|\mathcal{S}_k| - 1] = n - K, \tag{4.2}$$

where the second equation is because any block in the partition has at least 1 item (i.e., $|\mathcal{S}_k| \geq 1$). Since n is a constant, maximizing the size-focused objective is equivalent to minimizing the size of the partition K (thus the name "size-focused").

Generally, it is reasonable to expect that function $w(\cdot)$ should be monotonically *increasing* and *convex* — to see convexity, if a block of length l is broken into two blocks with length l_1, l_2, then the original longer block must be preferred, i.e., $w(l) \geq w(l_1) + w(l_2)$ for any $l = l_1 + l_2$, which is precisely the definition of convexity. Another natural requirement is $w(0) = 0$ in order to prevent empty blocks to contribute any value to the partition wellness. Both functions above satisfy these properties, and so does any monomial $w(x) = x^p$ for $p \geq 1$. Beyond these properties, however, it appears impossible to exactly know the format of $w(\cdot)$. To overcome this challenge, we resort to robust algorithms in subsequent subsection 4.7.3 and look to design algorithms that can *simultaneously* perform well for *every* wellness function satisfying these properties.

Robustness (as optimization constraint). If the informativeness metric above is regarded as the objective we want the partition to maximize, then the robustness metric can be viewed as a constraint we wish the partition to satisfy. In regard to robustness, aggregating information from multiple owners can help to mitigate each individual owner's noisy perception of their items in situations where Assumption 4.3 does not hold. Thus, the minimal number of owners for each partition block matters. Formally, we say a partition $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ is L-strong if there are at least L owners who share all items in the same partition block, i.e., $|\mathcal{T}_k| \geq L, \forall \mathcal{S}_k \in \mathfrak{S}$ with $|\mathcal{S}_k| > 1$. The larger L is, the less prone Isotonic Mechanism is to the noise in each owner's perceived ranking. Notably, $L \geq 1$ is the bare minimum requirement for Mechanism 4.2 to work. Next we show that generating a feasible L-strong partition reduces to generating a feasible 1-strong partition for a different ownership instance. This reduction helps us to convert any L-strongness constraint to a certain 1-strongness constraint during our design of the partition optimization algorithm.

Proposition 4.5 (Reduction to 1-Strong Partition)**.** *For any ownership instance* $\mathcal{O} = \{\mathcal{I}^j\}_{j\in[m]}$*, we can construct in* $O\left(m^L n\right)$ *time a different ownership instance* $\mathcal{O}' = \{\overline{\mathcal{I}}^j\}_{j\in[m]}$ *with the same owner and item set such that any partition* $\mathcal{S}$ *is* L*-strong in* $\mathcal{O}$ *if and only if* $\mathcal{S}$ *is* 1*-strong in* $\mathcal{O}'$*.*

Proposition 4.5 shows that to optimize the informativeness objective, it is without loss of generality to design algorithms to maximize the objective under 1-strongness, for a transformed ownership set instance. We defer its formal proof to Appendix C.2 but provide a simple example below to illustrate the main idea of the the reduction: any 2-strong partition in the ownership instance $\mathcal{O} = \{\{1,2\},\{1,2,3\},\{2,3\}\}$ is a 1-strong partition in the ownership instance $\mathcal{O}' = \{\{1,2\},\{2\},\{2,3\}\}$, and vice versa. Essentially, the instance $\mathcal{O}'$ is constructed by useing one owner to represent each subset of L owners in the original instance $\mathcal{O}$, whose common items are owned by this owner. This is a polynomial-time reduction for small constant L, which is often the case in practice due to the small number of co-authors

on a paper. Notably, this reduction does not fundamentally simplify the problem due to the hardness of maximizing the informativeness objective, as shown in the following proposition; see Appendix C.3 for its formal proof.

Proposition 4.6 (Hardness of Partition Optimization)**.** *It is NP-hard to find the optimal partition* $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ *that maximizes the size-focused objective in Eq.* (4.2) *subject to* 1*-strongness.*

4.7.3 Fast Greedy Partition with Robust Approximation Guarantees

Despite the computational hardness for the optimal partition, we show that a natural greedy algorithm presented in Algorithm 4.4 can efficiently find 1-strong partitions with a provably good approximation guarantee. In words, this greedy algorithm just iteratively selects the largest residual paper set owned by some owner. This algorithm can be implemented in time almost linear in the input size $\sum_{j\in[m]} |\mathcal{I}^j|$, up to a $\log(m)$ factor, thus is essentially the fastest algorithm one could hope to design.

ALGORITHM 4.4: A Greedy Algorithm for 1-Strong Partition

Input: Ownership sets $\{\mathcal{I}^j\}_{j=1}^m$.

Initialize the partition as $\mathfrak{S} = \{\}$ and set of selected items $\overline{\mathcal{I}} = \varnothing$.
while $\overline{\mathcal{I}} \subset [n]$ **do**
    Determine the largest residual paper set $\mathcal{S}^* = \mathcal{I}^{j^*} \setminus \overline{\mathcal{I}}$ where $j^* = \operatorname{argmax}_{j\in[m]} |\mathcal{I}^j \setminus \overline{\mathcal{I}}|$.
    Update $\overline{\mathcal{I}} \leftarrow \overline{\mathcal{I}} \cup \mathcal{S}^*$ and $\mathfrak{S} \leftarrow \mathfrak{S} \cup \{\mathcal{S}^*\}$.
return $\mathfrak{S}$.

The highlight of this greedy algorithm is that, though the algorithm itself is fully agnostic to any partition objectives, its output is a robustly high-quality partition for *every* natural wellness function, i.e., monotone and convex wellness functions, as formalized in the following Theorem 4.7. In particular, consider the following hypothesis class $\mathcal{W}$ of wellness functions

$$\mathcal{W} := \{w : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} | w(0) = 0, w \text{ is convex, non-decreasing}\}.$$

Algorithm 4.4 enjoys the following strong approximation guarantees for every w within $\mathcal{W}$, whose proof is deferred to Appendix C.5. In Remark 4.7.1, we give the exact approximation ratio for some optimization objectives of special interests and show its tightness. It is also worth mentioning that while Algorithm 4.4 only finds a partition satisfying 1-strongness, the algorithm as well as its constant approximation guarantees can be readily extended to find the optimal L-strong partition in $O\left(m^L n\right)$ time, using the reduction in Proposition 4.5.

Theorem 4.7 (Robust Approximation of Greedy)**.** *For any ownership instance $\{\mathcal{I}^j\}_{j\in[m]}$ with $N = \sum_{j\in[m]} |\mathcal{I}^j|$ total number of author-paper pairs, Algorithm 4.4 runs in $O(N \log m)$ time and outputs a 1-strong partition that is simultaneously a $c(w)$-approximation of the optimal objective for every $w \in \mathcal{W}$, where*

$$c(w) = \inf\left\{\frac{w(x)}{w'_-(x)x}\middle| w'_-(x) > 0, x \geq 2\right\}.^{12}$$

Remark 4.7.1. *Theorem 4.7 readily implies that the partition found by the greedy algorithm* simultaneously *approximates at least $1/2$ of both the optimal comparison-focused objective and size-focused objective. More generally, for α-th degree polynomial functions $w(x) = x^\alpha$, the greedy algorithm simultaneously guarantees $(1/\alpha)$-fraction of the optimal objective since $w(x)/[w'(a)x] = 1/\alpha$ for every x. Such objective-agnostic property of Algorithm 4.4 is especially desirable in our application scenarios where the concrete form of the partition objective is difficult (if not impossible) to know. Moreover, the approximation ratio in Theorem 4.7 for the greedy algorithm is provably tight under every monomial objective $f(\mathfrak{S}) = \sum_{k=1}^K |\mathcal{S}_k|^\alpha$, $\alpha > 1$, and the construction of lower bound instances can be found in Appendix C.5.1.*

12. Above, $w'_-(a) = \lim_{x\to a^-} \frac{w(x)-w(a)}{x-a}$ denotes the left derivative of w at a.

4.8 Experiments

In this section, we study the empirical performance of our proposed mechanism in the realistic settings of academic conference peer review. One major challenge for our evaluation is that its performance measures rely on the underlying ground-truth scores, and such information is unattainable in most applications — indeed, if we already know the ground-truth scores, peer review would not be needed anymore. Fortunately, from both ICLR and ICML, we are able to collect some parts of their data and synthesize some of the unobservable parts. Below, we describe our experiment setups and results.

4.8.1 Experiment Setups: Datasets, Baselines and Metrics

We start by elaborating on the preprocessing procedures and characteristics of the two real-world datasets we experimented on.

1. **ICLR 2021-2023:** We collect ownership relations and review scores from ICLR 2021-2023 [ICL, 2021, 2022, 2023] that are made publicly available from OpenReview.net. Based on the recorded review scores, we simulate the ground-truth scores $\boldsymbol{R} = \mathbf{y} - \boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2)$ is a zero-mean Gaussian random variable with standard deviation $\sigma \in \{1, 2, 3, 4\}$.

2. **ICML 2023:** We collect ownership relations, review scores as well as the authors' reported ranking on their papers from ICML 2023, based on surveys from OpenRank.cc. We only keep the authors (and their papers) who have participated the survey and provided a ranking of their submissions. For each paper, we split its review scores by their reviewers' confidence level. We set the review score with the least confidence level as the paper's raw review score $\boldsymbol{y}$, and the average score of remaining reviews as the paper's ground-truth score $\boldsymbol{R}$.

For comparison, we consider the following alternative calibration approaches, evaluated under the mean squared error (MSE), $\frac{1}{n}\left\|\boldsymbol{R}-\widehat{\boldsymbol{R}}\right\|^2$: *baseline*, which directly uses the raw review scores; *random*, which uses a randomly generated partition for the Isotonic Mechanism.

Table 4.4 summarizes the basic statistics of ICLR 2021-2023 that we collected as well as the partitions determined from their ownership sets. There is a clear growing trend in conference size in terms of submission and author number. Moreover, from both partition objective functions, the greedy partition is a clearly better choice than the randomly generated partition, though we are unable to empirically examine the approximation ratio of Greedy (Algorithm 4.4), since it is intractable to find exactly optimal partition for such large instances given its NP-hardness.

	#authors	#papers	$\text{obj}(\mathcal{S}_{\text{greedy}})$	$\text{obj}(\mathcal{S}_{\text{random}})$	$\text{obj}'(\mathcal{S}_{\text{greedy}})$	$\text{obj}'(\mathcal{S}_{\text{random}})$
ICLR 2021	8875	2964	5242	2515	1365	1132
ICLR 2022	10583	3328	6123	2966	1574	1324
ICLR 2023	15372	4881	9764	4600	2489	2045

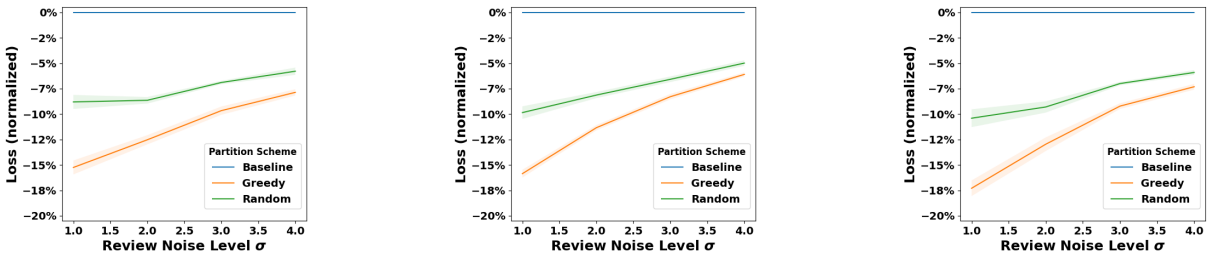
Table 4.4: Statistics of ICLR 2021-2023, where $\mathcal{S}_{\text{greedy}}, \mathcal{S}_{\text{random}}$ are greedy and random partitions generated in the ownership graph of each year's conference, obj, obj$'$ are respectively the comparison-based and size-based objective, defined in Equations (4.1) and (4.2).

4.8.2 Experiment Results

Given the two rich sets of real world peer review data, we empirically study the following three questions about the performance of our proposed mechanism.

Q1: How does the mechanism's calibration quality change under different levels of noise? Here we use the ICLR dataset where we can control the noise level in our simulation of ground-truth score. Figures 4.5 compares the empirical performance of Isotonic Mechanisms with different baselines at different review noise levels. We can see that the proposed Isotonic Mechanism can mitigate a substantial amount of noise from the review

score and improve the precision metric over the baseline approach. The percentage MSE loss in Figure 4.5 suggests our proposed method is able to mitigate about $10-20\%$ amount of review noise in the past three years of ICLR. In addition, the performance of Isotonic Mechanisms steadily improves as the conference size grows — an encouraging sign for the applicability of the Isotonic Mechanism. This performance is also reflected to determining accepted paper (i.e., top 30%) as illustrated in the lower panel of Figure 4.5. It suggests that $5\% \sim 15\%$ of the mistakenly rejected papers can be rectified to acceptance through the proposed score calibration method.[13]

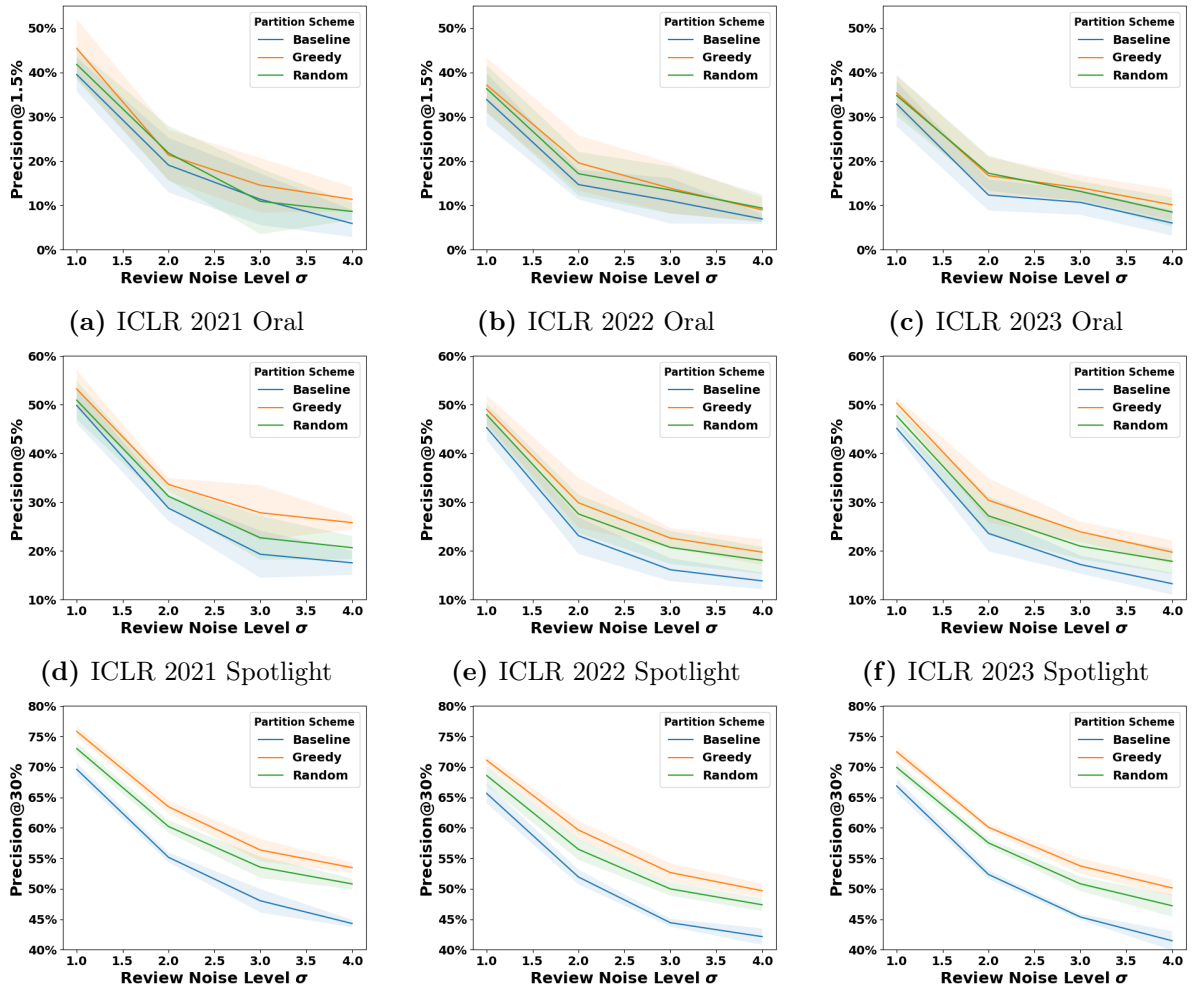


(a) ICLR 2021 **(b)** ICLR 2022 **(c)** ICLR 2023

Figure 4.5: The mean square error (MSE) loss of scores calibrated by different models (normalized as the percentage change, $\frac{\text{model}-\text{baseline}}{\text{baseline}}$) under varied noise level σ.

Q2: The informativeness and robustness tradeoff of the Isotonic Mechanism when there is authors' perception noise. Our isotonic mechanism offers a flexibility for practioners to decided the number authors from who they would like to elicit ranking information from (i.e., the L-strongness of the mechanism). Eliciting from a single author can lead to larger blocks in the partition hence more efficiency, but suffers more risk of miscalibration when the authors have noise perception of the ranking. Our second experiment tests such informativeness vs robustness tradefoff. This requires us to generate problem instances with carefully structured ownership relations in order for a diverse set of partitions to emerge. We thus use pure simulation data here. Specifically, in our simulation, the authors

13. We also plotted how the MSE improvement influence the accuracy of more stringent paper selections, in particular spotlight (top 5%) and oral (top 1.5%) presentations. The performance has similar trend of improvement, hence is omitted here to avoid repetition.



(a) ICLR 2021 Oral **(b)** ICLR 2022 Oral **(c)** ICLR 2023 Oral

(d) ICLR 2021 Spotlight **(e)** ICLR 2022 Spotlight **(f)** ICLR 2023 Spotlight

(g) ICLR 2021 Poster **(h)** ICLR 2022 Poster **(i)** ICLR 2023 Poster

Figure 4.6: The percentile precision based on scores calibrated by different models under varied review noise σ.

perceive score as $R_i + \zeta_i^j$ and control the variance of authors' perception noise variable ζ_i^j in a range from $\{0.1, 0.5, 1, 2\}$ while we vary the partition structure. We then construct a class of instances whose connections can be characterized by a 7-level ternary tree with 3^7 items and $(3^7 - 1)/2$ authors. Intuitively, the partition at one extreme is to have only a single block of all items owned by the root node (author), and there is only one author we can elicit. At the other extreme, the partition is to have one block for the set of items owned by each leaf node, and we can elicit from the authors at this leaf node and all its ancestor nodes. Hence there exist L-strong partitions with L from 1 to 7. In Figure 4.7, we can

observe a clear trend of tradeoff: in the case of low noise (blue curve), the large block size is preferred to have more pairwise information, while in the case of high noise (red curve), the small block size is preferred to have more authors per block (larger L). In cases of medium noise (green or orange curve), the optimal block size is a careful balance of the block size and L-strongness. The trend showcases the importance of eliciting from multiple authors, e.g., to enforce the L-strong partition constraint, though the parameters shall be tailored carefully for the noise level in real-world applications.

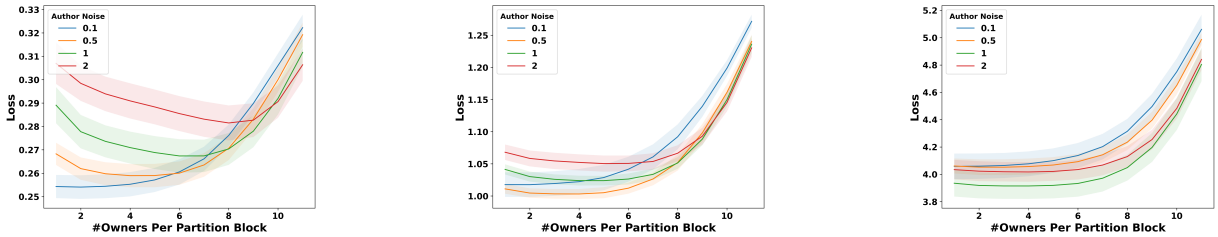


Figure 4.7: The (MSE) loss of Isotonic Mechanisms under different levels of authors' perception noise with varying number of authors per partition block and different review noise $\sigma = 1$ (left), 2 (middle), 4 (right).

Q3: Evaluations on the real ICML 2023 data. Our last set of experiments focuses on the ICML dataset which is richer than the ICLR data since this data contain reported ranking from the real-world survey, hence allowing us to conduct an almost all real experiments, except that need a surrogate of ground-truth scores – for this, we use the average of two high confidence review scores as the surrogate of "ground-truth scores", whereas the average of the two lower confidence review scores as the "review score". The performance of Isotonic Mechanisms is presented in Table 4.5. From these results, we are able to state with high confidence that Isotonic Mechanisms under both partition schemes are capable of reducing the noise in the review process.

	Baseline	Greedy	Random
MSE	1.4267	1.3012	1.3210
F-Statistics	-	15.828	9.911
P-Value	-	7.032×10^{-5}	1.652×10^{-3}

Table 4.5: The loss of Isotonic Mechanisms under different partition scheme in the ICML dataset, along with the results of F-test and P-value on whether the loss indeed decreases.

4.9 Final Remarks

In this work, we design and analyze the partition-based Isotonic Mechanism to assist peer review in machine learning conferences, where multiple authors can contribute to various papers. We establish its formal guarantees of truthfulness, statistical and computational efficiency within a theoretical framework that may serve as a foundation for future research. However, our mechanism is not without limitations. Below we discuss the applicability of the technical assumptions in our theoretical analysis and point out the potential directions for further investigation.

Owner's Private Information. Assumption 4.3 is a standard private knowledge assumption, commonly adopted in the mechanism design literature (e.g., in auction design it is often assumed that bidders know their values or private signals). Here we assume owners have accurate knowledge on the information we elicit, i.e., their items' ranking. This assumption is not to ask owners to have precise knowledge on ground-truth scores of their items; agreement on the (much coarser) ranking information suffices. We remark the intrinsic tradeoff here when eliciting information from owners. On one hand, eliciting very fine-grained score information can be risky since owners may not have accurate information. On the other hand, eliciting overly coarse information may not be useful any more. As a compromise, we believe the elicitation of ranking information is a reasonable "middle ground" to begin with, though it is also an interesting future direction to explore other possibilities. Additionally, in Section **??** where we evaluate our algorithm on situations with authors' noisy perception

of the true paper ranking, we still observe improvements on score estimation by the isotonic mechanisms, indicating that this approach remains statistically effective when owners can supply additional (even somewhat noisy) information not yet included in review scores.

That said, our current mechanism cannot guarantee a truthful Nash equilibrium if some co-owners would "truthfully" report a misleading ranking according to his noisy knowledge of his paper. The mechanism also ceases to be truthful if an owner can to some degree predict the realization of review noise, e.g., due to ex-ante insider information on the reviewers. Meanwhile, it remains a fundamental design challenge for future work to ensure the authors' incentive compatibility when they hold outside information or conflicting information with each other.

Noise Structures in Review Scores. The exchangeable noise structure in Assumption 4.4 is also a natural choice and has been widely adopted in mechanism design literature. The typical rationales behind this assumption are as follows: an owner determines the ranking of his paper *ex-ante*, i.e., before any submission and reviewer assignment have been made, a moment that he does not know how his papers will be reviewed. Thus, the best this owner could assume is the symmetry of review noises for all his papers, which is precisely described by our exchangeable noise assumption. Similar justification on such exchangeable noise assumption is also given in a recent work by [Maskin, 2023], though for a completely different mechanism design problem. We point out that our Assumption 4.4 is a strict relaxation of the i.i.d. noise assumption, which is perhaps more often adopted in the peer review literature (See e.g., [Baba and Kashima, 2013, MacKay et al., 2017, Tan et al., 2021]). Notice that the noise satisfying Assumption 4.4 does not even need to have a zero mean. In addition, if the review score of each paper is averaged over several reviews and the number of reviewers assigned to each paper is treated as an i.i.d. random variable, then the exchangeable noises allow the papers to have different variances.

In addition, an advantage of the exchangeable noise assumption is that it is non-parametric,

without the need of modeling the distribution family. If one is willing to assume specific format of the noise distributions, then the exchangeable noise assumption may be further relaxed (e.g., see recent work by Yan et al. [2023] using exponential family to model review score distributions). However, without such structural distribution assumption, we conjecture it will be quite difficult (if not impossible) to relax this exchangeable noise assumption in our setup.

Owners' Utility Structures in Peer Review. The convex utility in Assumption 4.5 may not appear as common as the two assumptions above, hence worth some more discussions here. Utility assumptions are often subject to its application context, especially in mechanism design without money [Schummer and Vohra, 2007]. The convex utility reflects the nature that items receiving higher rating tend to generate significantly more reward. We believe such convex utility is particularly suitable in the context of evaluating research (e.g., conference review), fundamentally due to the "high-risk-high-reward" nature of research, recognized by various funding agencies [Wagner and Alexander, 2013, Cao and Zhang, 2022, Franzoni, 2023]. A spiritually similar context is in the tournament design [Lazear and Rosen, 1981, Rosen, 1985], where participants are incentivized to exert more effort or take on more risk to compete for the best as the potential rewards are significantly larger at the top; participants' utilities are commonly assumed to be convex in their ratings (akin to the scores of papers in our setting). Notably, while consumers' utilities in consumer theory are often assumed to be concave in the *quantity* of products due to the law of diminishing return, the utility function U^j in our problem is rather to capture *quality* or *scarcity* of a product (e.g., accepted papers). This assumption of convexity in scarcity is consistent with the classic assumption of concavity in quantity. That is, if a product is hardly accessible (e.g., papers with high scores), or an award is conferred to only a few, we tend to derive more utility — a phenomenon widely observed in behavioral economics and social psychology [Veblen, 2017, Verhallen, 1982].

Moreover, most publication venues have been implementing various metrics to promote

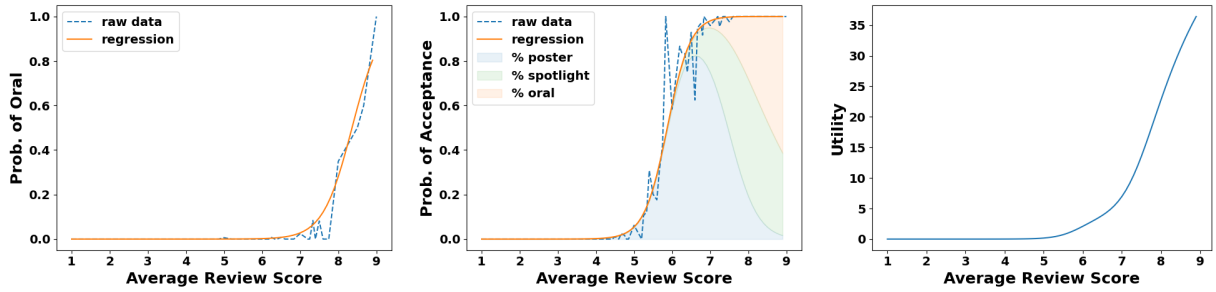


Figure 4.8: The first two plots base on the ICLR 2022 dataset illustrate the probability of an paper getting accepted as an "oral", "spotlight" "poster" w.r.t. its average review score. The dashed line denotes the estimated probability from raw data, the smooth line denotes the probability predicted by logistic regression. The last plot illustrates the expected utility curve based on a paper's chance of receiving different acceptance labels.

high-quality research. For instance, given the large number of accepted papers at AI/ML conferences today, these venues start to distinguish accepted papers with labels such as "poster", "spotlight" and "oral"; oral papers are promoted with extra resources such as longer presentations and more official media highlights, which naturally lead to much higher utility than an accepted poster paper. Here, we are able to observe evidence from real-world conference data that supports our convex utility assumption. As illustrated in the left panel of Figure 4.8, we can see that the probability of a paper receiving oral presentation at ICLR 2022 is a convex function of its averaged review score. The middle panel of Figure 4.8 plots the acceptance probability — summing up over acceptances with different labels $\sigma \in \Sigma =$ {"poster", "spotlight" and "oral"} — as a function of the review score. While the entire curve is not convex, we observe increasing fraction of acceptance with higher labels as the score increases. Suppose an author derives utility $u(\sigma)$ for label-σ acceptance for $\sigma \in \Sigma$. The right panel of Figure 4.8 plots the expected utility as a function of score using $u(\sigma) \propto 1/\mathbf{Pr}[\sigma]$, where $\mathbf{Pr}[\sigma]$ is the ratio of accepted label-σ papers among all papers with a given score (i.e., the utility $u(\sigma)$ of a label-σ paper is assumed to be propositional to its scarcity). We can see that this expected utility curve $U(\widehat{y})$ is convex according to the empirical data. In practice, authors with more papers are expected to derive higher utility $u(\sigma)$ on those more scarce paper labels, in which case the convex utility assumption is even easier to satisfy. As

a neat coincidence, these authors with more papers happen to also be those who will supply more ranking information to our mechanism hence will be incentivized to do so truthfully. In Remark 4.7.2, we describe how the design of the rating metrics could encourage stronger convexity of author utility.

Remark 4.7.2 (A case of the "convexifying" owner utilities in peer review)**.** *The strong convexity of utility could result from not only the high-risk-high-reward nature of research but also the artful design of peer review systems. Existing measures include the selection of best papers and selective labels of "oral" and "spotlight". We observe that a potential redesign of the review rating metrics can even further convexify owners' utility. Specifically, the method is to set less level choices for excellent papers (e.g., merging traditional "7:very good" and "8:excellent" to just a single score of "8:excellent") while use more fine-grained rating for regular papers. Assuming the exchangeable review noises, such kind of designs will make the differences between regular papers wider while the differences between excellent papers narrower. An explicit goal of this mechanism is to better distinguish papers of different qualities at the acceptance borderline. Meanwhile, despite having narrower score differences, the papers of high ratings typically require much less deliberation (e.g., most of them are accepted with probability* $> 95\%$ *at ICLR 2022). More importantly, this mechanism can implicitly encourage the convexity of acceptance curve by, intuitively, "stretching" the middle borderline region of the curve whereas "condensing" its high-score region. Such redesign is already taken place at ICLR 2024, where reviewers can only choose ratings from a list of* non-linear *scores ("1","3","5","6","8","10").*

For future work to relax the current utility assumption, one direction is to study utility functions that are monotone in the adjusted scores with the design of sequential review, despite the downside of prolonged peer review procedures [Zhang et al., 2024]. In addition, an author's value of a paper could depend on factors such as authorship order, the number of co-authors, and her relative contributions to the paper [Demaine and Demaine, 2023].

This modeling perspective could more accurately reflect the reality that authors may assign different values to their co-authored submissions. Recognizing the potential heterogeneity of the utility function poses a significant, yet crucial challenge to resolve. Another direction is to understand the potential group manipulations or collusion under our mechanism. This is generally a challenging problem in mechanism design, as it requires incentive analysis beyond the unilateral deviation in Nash equilibrium.

Lastly, it is important to realize that theory might only take us so far in this field, due to the complex nature of human subjects. We view our work as a theoretical foundation for the initiatives aimed at eliciting self-evaluations and managing incentives in scientific reviews. The empirical results based on the *OpenRank* survey in ICML 2023 are promising, but still has some important gaps from a full test deployment — the authors have no strategic incentives as their reports would not influence the peer review process. As an initial step, Isotonic Mechanism can function as an independent layer atop existing peer review processes, with its calibrated scores serving as reference points to assist expert human decision-making (e.g., helping area chairs to detect anomalies and request additional reviewers). We look forward to more real-world experiments to better understand the potential limitations of Isotonic Mechanisms in practice and drive iterative improvements towards solving the owner-assisted calibration problem.

CHAPTER 5

UNCOUPLED LEARNING TOWARDS RATIONALIZABILITY

5.1 Introduction

Two seminal results in uncoupled [1] multi-agent learning are that agents using no regret learning algorithms will converge to a coarse correlated equilibrium (CCE) whereas the stronger no-swap regret learning algorithms will bring agents to a correlated equilibrium (CE) [Foster and Vohra, 1999, Blum and Mansour, 2005]. In both results, however, the converging sequence is the *average* of agents' historical plays; the analysis of the *last-iterate* convergence, a strictly stronger convergence guarantee, has been a well-known challenge in the study of uncoupled learning dynamics, despite significant research efforts devoted to its study even until today. [2] For modern machine learning applications, the last-iterate convergence is often more desirable due to the difficulty of averaging agent's actions, typically represented by neural networks. In addition, it is also more realistic as an approach to predict equilibrium outcomes of economic systems, e.g., by assuming uncoupled learning agents in the system — after all, it is the most recent state (i.e., the "last iterate") of the system that matters and evolves, while the average history of the system may not have any real-world meaning. Given these challenges of obtaining last iterate convergence towards equilibria in general game as well as its strong real-world motivations, we seek to answer the following question in this paper:

> *Are there natural equilibrium concepts, other than the often studied CE or CCE, for which the last-iterate convergence guarantee of uncoupled learning dynamics*

1. In the uncoupled learning setup, the learning rule of each agent must not rely on any opponent's historical actions or payoffs [Pradelski and Young, 2012, Daskalakis et al., 2011, Cai et al., 2023]

2. Even in the zero-sum games, it is only recently shown by Cai et al. [2023] that a finite last-iterate convergence rate can be provably achieved by uncoupled learning dynamics under bandit feedback. See also a recent paper by Anagnostides et al. [2022c] for detailed discussions about this challenge, relevant works and a few generalizations beyond zero-sum games in which last-iterate convergence could be possible.

can be established for general games? If so, what is the solution concept, and what algorithm is guaranteed to converge?

To answer the above question, we initiate the study of last-iterate convergence of uncoupled multi-agent bandit learning towards a fundamental game-theoretic solution concept, known as *rationalizability*, developed through a series of seminal economic works [Bernheim, 1984, Pearce, 1984, Brandenburger and Dekel, 1987, Milgrom and Roberts, 1990]. In particular, this paper pursues *rationalizability* as a natural multi-agent learning objective for several reasons. First, without converging to the rationalizability, it is impossible to reach the Nash equilibrium (NE) or CE (see Corollary 5.3). At a high level, rationalizability is a more permissive and robust solution concept that relaxes the stringent belief assumptions in both Nash [Bernheim, 1984, Pearce, 1984] and CE [Aumann, 1987, Brandenburger and Dekel, 1987]: while both NE and CE requires each player to respond optimally to the *accurate* belief about her opponents' strategy profile, players under rationalizability may take any actions that are best responses to some *erroneous but rational* belief about her opponents' strategy profile (see Definition 6). This also leads to the second point that rationalizability is often deemed a more realistic outcome to expect in games with uncertainty [Dekel and Siniscalchi, 2015]. The notion originates from the epistemic approach to understand agents' rational behavior in a non-cooperative environment without perfect knowledge of others' strategy profile — it characterizes the outcomes arise from the only common knowledge of rationality. We thus can think of rationalizability as a proxy for us to understand how much "rationality" learning algorithms can obtain under the uncoupled learning setup, compared to human agents.

Rationalizability is also related to the basic strategic concept of *dominance elimination* studied since the early days of the game theory field [Gale, 1953, Raiffa and Luce, 1957] — we say, an action a of some agent is *dominated* by another action a' in a strategic game if the agent's payoff of action a is always smaller than her payoff of a', regardless of what

actions other agents play. While it is debatable that whether regular humans would ever play an equilibrium in a strategic game [Rabin, 1993, Wright and Leyton-Brown, 2010], it is widely observed and well accepted that rational human players generally would avoid playing dominated actions [Fudenberg and Liang, 2019]. Therefore, an intriguing question is whether there are learning algorithms that can efficiently approach such kind of human wisdom. Notice that, after eliminating some dominated actions, other actions may then start to become dominated and thus require an additional iteration of dominance elimination; the *iterated dominance elimination* turns out to be a highly non-trivial task in the uncoupled learning setup and many of the existing algorithms are provably inefficient. As we will explain in Section 5.4, the strategy profiles surviving this process of *iterated dominance elimination* coincides with the set of rationalizable outcomes [Bernheim, 1984, Aumann, 1987], and it thus serves another key motivation to design algorithms with efficient convergence guarantee towards rationalizability.

Notably, iterated dominance elimination in games is not as rare as they may first sound. For example, Alon et al. [2021a] recently show that for randomly generated two-player $m \times n$ games with $m = o(\log(n))$, the fraction of actions that survive iterated dominance elimination tends to 0 as $n \to \infty$. The concept also has a variety of applications, including voting [Moulin, 1979], auctions [Azrieli and Levin, 2011], market design [Abreu and Matsushima, 1992], supermodular game [Milgrom and Roberts, 1990], oligopolistic competition [Börgers and Janssen, 1995] and global games [Carlsson and Van Damme, 1993]. One celebrated example is Akerlof's "market for lemons" [Akerlof, 1978]. Each seller in this market is looking to sell used cars which are equally likely to have quality `H`/high, `M`/medium or `L`/low (low quality cars are also known as "lemons" in America). Prospective buyers value `H`-cars at \$1000, `M`-cars at \$500 and `L`-cars at \$0, whereas sellers value keeping a `H`-car at \$800, `M`-car at \$400 and `L`-car at \$0 (these values are only privately known to sellers). Akerlof studies the situation that sellers precisely know their car's type whereas buyers cannot distinguish

the good cars from lemons. Suppose that the car types are uniformly distributed on the market in the beginning, due to the inability to distinguish car quality, any buyer will immediately eliminate any price above her average value $\$500 = (1000 + 500 + 0)/3$. After this elimination, the buyer's price becomes lower than `H`-car's reservation value, and thus drive `H`-car sellers out of the market. Consequently, the buyer will *gradually learn* that the market has no `H`-cars under price \$500 and thus will *iteratively* eliminate any price above $\$250 = (500 + 0)/2$, which then further drives `M`-car sellers out of the market. Ultimately, Akerlof observes that this iterated dominance elimination procedure will drive all good cars out of the market, and only lemons are ever traded. In this paper, we shall examine the more realistic situation when buyers and sellers do not know the exact average value of the car qualities in advance but only have noisy bandit feedback about each sold car. We seek to understand *how fast the market collapse observed by Akerlof may happen when players have such noisy bandit information feedback.* In Appendix 5.7, we will revisit this example as an empirical illustration of our theoretical results.

At this point, an immediate thought one might have is whether any standard no-regret learning algorithm would already suffice to eliminate the (obviously bad) dominated actions. The answer is indeed Yes, but with a crucial limitation that they may necessarily take *exponentially* many rounds, as we will prove later in Section 5.5. A surprising insight revealed from our formal results is that the classic notion of *regret* in multi-agent settings is not fully aligned with the performance of iterated dominance elimination. First, the history that standard no-regret algorithms exploit could become the inertia that impedes the iterative process of dominance elimination. This claim shall be self-evident in the proof of Theorem 5.7. Second, the notion of "regret", designed for either stochastic or adversarial settings, fails to encourage the coordination that facilitates the iterative learning process of the learning agents. These observations echoes with the findings of Viossat and Zapechelnyuk [2013] that the Hannan sets [Hannan, 1957] may contain highly non-rationalizable outcomes.

Motivated by the aforementioned fundamentality and intricacies, this paper studies how agents in a uncoupled multi-agent system can learn to *rationalize* — or equivalently to iteratively eliminate all dominated actions — under *noisy bandit information* feedback. This is also a natural generalization of the well-known *action elimination* problem [Even-Dar et al., 2006] to multi-agent setups. Our study reveals interesting new challenges of learning in game-theoretical settings that its algorithm design may require different ideas from the classical online learning under either adversarial or stochastic environment assumptions. Notably, an interesting recent follow-up work by Wang et al. [2023] considered the same problem setup as us and proposed an iterative best response approach that improves on our convergence rate by some linear factors. However, their proposed algorithm is designed in a coordinated fashion: when an agent is estimating the mean reward of his actions, the remaining agents must keep playing the same action profile. In such a design, each agent can infer the strategies of all other agents at any time. While they obtained better learning efficiency than our algorithm, such coordinated design of learning algorithm violates the uncoupled learning setup and is not our goal in this paper (also see additional discussions in Section 5.3). In contrast, our work predicates the efficient convergence even when agents only have a loose agreement on the type of learning algorithms without any further intention or capability to coordinate (e.g., the sellers on the Akerlof's market for lemons). Moreover, experiments in Appendix 5.7 demonstrate that the performance of our algorithm could remain robust facing potentially adversarial opponents.

Contributions At the conceptual level, our key contribution is to identify the solution concept of *rationalizability* as a fundamental multi-agent learning objective. On the one hand, it is the best possible outcome under the uncertainty of others' strategy profile in uncoupled multi-agent learning problems; on the other hand, reaching rationalizability has important economic implications in various classes of games. Our technical contributions are twofold. First, we provide formal barriers of reaching *rationalizability* under noisy bandit

feedback. To do so, we identify an interesting benchmark class of dominance solvable game instances, coined *diamond-in-the-rough game* (DIR), and show that a broad class of no-regret online learning algorithms, including the Dual Averaging algorithm [Nesterov, 2009, Xiao, 2010], has to run exponentially many rounds to eliminate all dominated actions with non-increasing learning rates. Moreover, we prove that the algorithms with the stronger no *swap* regret suffers similar exponentially slow convergence. Second, we propose a new variant of the Exp3 algorithm with a carefully designed diminishing history mechanism to overcome the barriers in such learning tasks and prove its efficiency of eliminating all dominated actions within polynomially many rounds in the sense of last-iterate convergence. Our experiments demonstrate the effectiveness of our algorithm not only in the synthetic DIR games but also in other real-world games.

5.2 Background and Related Work

Rationalizability and Epistemic Game Theory While the classical game theory takes a top-down perspective, specifying the outcomes of a game by different solution concepts, the epistemic approach to game theory takes a bottom-up perspective, asking under what epistemic conditions will players behave with respect to particular solution concept [Dekel and Siniscalchi, 2015]. In particular, using the mathematical tools developed in seminal works by Nobel laureates Harsanyi [1967] and Aumann [1976], it concerns decision problems under uncertainty such as the choices and knowledge of other players. The notion of *rationalizability* is key to the development of epistemic game theory. Bernheim [1984], Pearce [1984] proposed the concept of rationalizability as the logical consequence of assuming that the only common knowledge is game structure and the rationality of the players, drawing an important connection to the iterated elimination of strictly dominated strategies. Brandenburger and Dekel [1987] introduced the solution concept of correlated rationalizability that allows players to have correlated conjectures over others' actions. Notably, there has been

increasing usage of the correlated version of rationalizability, in part based on the influential argument of Aumann [1987]: in games with more than two players, correlation may express the fact that what player 3 thinks that player 1 will do may depend on what he thinks player 2 will do, which has no connection with any overt or even covert collusion between player 1 and 2. Our paper also adopts this notion of correlated rationalizability. In addition, Aumann [1987] derives the CE from the additional assumption that the players are "Bayesian rational" with a common prior, and Aumann and Brandenburger [1995] provides an epistemic characterization of the Nash equilibrium in terms of mutual knowledge of strategy choices. More recently, the solution concepts of interim independent rationalizability [Ely and Pęski, 2006], interim correlated rationalizability [Dekel et al., 2007] are developed for the incomplete information games.

Multi-agent Learning in Games Multi-agent learning in games has been of interest since the early days of artificial intelligence and economics [Von Neumann and Morgenstern, 2007, Brown, 1951, Hart and Mas-Colell, 2000]. In recent years, there is a growing body of work on decentralized no-regret dynamics and their equilibrium convergence properties in various special classes of games including zero-sum game [Daskalakis et al., 2011, Rakhlin and Sridharan, 2013, Syrgkanis et al., 2015, Daskalakis et al., 2018, Daskalakis and Panageas, 2018, Mertikopoulos et al., 2018a], concave game [Mertikopoulos and Sandholm, 2016, Bravo et al., 2018, Mertikopoulos and Zhou, 2019], potential games [Cohen et al., 2017b], monotone games [Cai et al., 2022], and auctions [Feng et al., 2021]. Different from the goal of these works on convergence to CCE or NEs in special classes of games, we target convergence in arbitrary multi-player games but to a relaxed equilibrium notion, i.e., learn to *rationalize* by removing dominated actions. Indeed, economists [Viossat, 2015] framed this learning goal broadly into the question *whether evolutionary processes lead economic or biological agents to behave as if they were rational.* Our study focuses on the convergence properties of various no-regret learning algorithms specifically in the multi-agent *bandit* learning setting (a.k.a.

the "radically uncoupled" setup [Foster and Young, 2006]). Interestingly, our proposed mechanism of diminishing history resonates with the well-established studies of both behavioral economy [Fudenberg and Peysakhovich, 2016] and political science [Axelrod and Hamilton, 1981]. It is also seen in one form or another (such as increasing learning rate or recency bias) of many learning algorithms for different purposes [Rakhlin and Sridharan, 2013, Syrgkanis et al., 2015, Bubeck et al., 2017, Agarwal et al., 2017, Lee et al., 2020], some of which even beyond the domain of online learning [Jin et al., 2018, Brown and Sandholm, 2019]. In the experimental section, we will compare our algorithm with some of them from these previous works.

Another line of work studied the fast convergence to approximate efficiency and CCEs by regularized learning algorithm with properties such as Variation in Utilities (RVU) [Syrgkanis et al., 2015], or low approximate regret [Foster et al., 2016]. Different from the goal of these works on convergence to CCEs or NEs (for special game classes), we focus on a different but arguably equally fundamental goal, i.e., learning to *rationalize* by removing iteratively dominated actions. Similar objective is also examined by [Viossat and Zapechelnyuk, 2013], who showed that continuous fictitious play (CFP) eliminates all strictly dominated strategies and therefore converges to the unique Nash equilibrium in strictly dominance solvable games. However, CFP needs the computation of agents' best response, which requires expert knowledge of the utility function. Meanwhile, [Cohen et al., 2017a] focused on the dominance elimination property of the no-regret learning algorithm, Hedge, on generic games. This algorithm requires full information feedback and is thus not applicable to our setting where each agent can only observe the utility of actions she took and may not even know the existence of her opponents in the uncoupled learning setup.

Optimism and Diminishing History The mechanism to focus on the recent experience is also seen in one form or another (such as increasing learning rate or recency bias) of many learning algorithms for different proposes. [Brown and Sandholm, 2019] introduced the

variant of counterfactual regret minimization that discounts the prior iterations to prevent earlier costly mistake from hampering the convergence. [Jin et al., 2018] used exponential discount for Q-learning but with respect to episode length instead of time. For online learning algorithms, [Rakhlin and Sridharan, 2013] introduced an "optimistic" variant of Mirror Descent with a minor but effective modification that counts the last reward observation twice. [Syrgkanis et al., 2015] showed that natural classes of regularized learning algorithms with a property of recency bias provably achieves faster convergence rates in normal form games. [Chen and Peng, 2020, Daskalakis et al., 2021, Anagnostides et al., 2022a] shows that the optimistic variant of Hedge enjoys the $O(\text{ploy}(\log T))$ regret, as well as the swap regret under the black-box transformation by [Blum and Mansour, 2005, Stoltz and Lugosi, 2005]. In the online bandit learning setting, a helpful technique to introduce historical bias is to apply increasing learning rate [Bubeck et al., 2017]. The increasing learning rate turns out to be powerful in several recent works: [Agarwal et al., 2017] used it to maintain a more delicate balance between exploiting and exploring so a master algorithm could perform almost as well as its best base algorithm, and [Lee et al., 2020] employed it to effectively cancel the potentially large variance of the unbiased estimators in high-probability regret bound analysis. However, as we will demonstrate in the experiment, none of these techniques can efficiently overcome the barrier of eliminating iteratively dominated actions, which necessitates our attempt of designing new algorithms.

Interestingly, our mechanism of diminishing history resonates with the well-established studies of both behavioral economy and political science. [Fudenberg and Peysakhovich, 2016] pointed out that recency bias is a behavioral pattern commonly observed in game-theoretic environments. In the renowned book, *The Evolution of Cooperation*, [Axelrod and Hamilton, 1981] empirically demonstrated that the tit-for-tat strategy (simply copying the last move of the opponent) is the most successful strategy in the repeated prisoner's dilemma game. They accordingly pointed out an important insight for game strategy design

– to be *provocable* to both retaliation and forgiveness, as tit-for-tat ignores all the good or bad experience from the opponent's previous silence or betray more than one round ahead. Nevertheless, such aggressive strategy would not work in online learning, as memory is critical for algorithm to optimize its decision from the past observation. Hence, our proposed algorithm generalizes such philosophy with a carefully designed discounting mechanism to balance the influence of history in online decision making while maintain the "provocability" crucial in certain game theoretical environment.

5.3 Preliminaries

In this section, we introduce the problem setup of this paper, starting with some basic notations and definitions from game theory. An N-player game in normal form consists of a (finite) set of agents $\mathcal{N} = \{1, \dots, N\}$, where the n-th agent have a finite set of actions (or pure strategies) $\mathcal{A}_n$. Let $\mathcal{A} := \prod_{n \in \mathcal{N}} \mathcal{A}_n$ denote the action space, $\boldsymbol{a} := (a_1, \dots a_N) \in \mathcal{A}$ denote the action profile, and $a_{-n} \in \mathcal{A}_{-n}$ as the action profile excluding agent-n's action. Without loss of generality, we assume every agent has K actions, i.e., $|\mathcal{A}_n| = K$.[3] Each agent n has a payoff function $u_n : \mathcal{A} \to [-1, 1]$ that maps the action profile (a_n, a_{-n}) of all agents' actions to the nth agent's payoff $u_n(a_n, a_{-n})$.[4] We denote such game instance as $\mathcal{G} := \mathcal{G}(\mathcal{N}, \mathcal{A}, u)$. In addition, each agent n may randomize her action by playing a *mixed strategy*, $x_n \in \Delta_{\mathcal{A}_n}$, from the simplex over $\mathcal{A}_n$. Denote $\mathcal{X}_n := \Delta_{\mathcal{A}_n}$ as the mixed strategy space of agent n, and $\mathcal{X} := \prod_{n \in \mathcal{N}} \mathcal{X}_n$ as the space of all mixed strategy profiles $x := (x_1, \dots, x_N)$ aggregating over all agents. Denote $x_n(a_n)$ as the probability of playing action a_n under mixed strategy x_n. Let $u_n(x_n, x_{-n}) := \sum_{a_1 \in \mathcal{A}_1} \cdots \sum_{a_N \in \mathcal{A}_N} u_n(a_1, \dots, a_N) \prod_{n \in \mathcal{N}} x_n(a_n)$ be the expected payoff of agent-n under the strategy profile x.

3. As long as each player's number of actions is upper bounded by the constant K, our main results hold. In fact, our results only depend on the total number of actions across all players.

4. Bounded utility is assumed for convenience but not essential, since it can always be re-scaled.

Dominated Actions and Iterated Dominance Elimination We say action a_n is strictly *dominated* by a *mixed* strategy $x_n \in \Delta_{\mathcal{A}_n}$, if $u_n(x_n, a_{-n}) > u_n(a_n, a_{-n})$, $\forall a_{-n} \in \mathcal{A}_{-n}$. Dominated actions are perhaps the simplest generalization of sub-optimal actions in single-agent decision making. The procedure for an agent to remove all her dominated actions is called *dominance elimination*. In the single-agent setting, dominance elimination degenerates to the widely studied best arm identification since all other actions are dominated by the optimal arm [Bubeck et al., 2009]. However, in multi-agent setups, eliminating all iteratively dominated actions may require many *iterations* of dominance elimination, as illustrated in the introduction. Moreover, an action dominated by a mixed strategy is not necessarily dominated by any pure strategy. So it is important to consider dominance elimination by mixed strategies. The process of iteratively applying such procedure to remove iteratively dominated actions is called *iterated elimination of strictly dominated strategies* (IESDS). This motivates our following natural definition of *elimination length*.

Definition 5. *For any finite game $\mathcal{G}$, we define the* ***elimination length*** L_0 *as the minimum number of iterations that IESDS needs to eliminate all iteratively dominated actions in $\mathcal{G}$. For any successful execution of IESDS with elimination length L_0, the corresponding* ***elimination path*** *is a sequence of* ***eliminated sets*** $(E_l)_{l=1}^{L_0}$ *where E_l contains all eliminated actions until iteration $l \in \{1, \cdots, L_0\}$.*

By definition, we have $E_1 \subset E_2 \subset \cdots \subset E_{L_0}$ and $|E_{L_0}| < \sum_{n=1}^{N} |\mathcal{A}_n| = KN$. Since the elimination path $(E_l)_{l=1}^{L_0}$ of a game $\mathcal{G}$ might not be unique, when we refer to an eliminated set E_l, we consider E_l from **any** possible elimination path. Let Δ be the smallest utility gap between any iteratively dominated action and the correspondingly dominant strategy during IESDS over all possible elimination paths. [5]

5. This notation is to draw an analogy to the stochastic bandit setting, where Δ typically denotes the gap of the means to different reward distributions, which largely determines the intrinsic difficulty of the problem. With slight abuse of notation, we also use Δ to represent the simplex by convention; the two use cases should be easily distinguishable.

Rationalizability and Rationalizable Actions Let $x_{-n} \in \mathcal{X}_{\mathcal{A}_{-n}}$ be a *belief* of the n-th agent on the (possibly correlated) strategy profile of the other players. [6] We say an action $a_n \in \mathcal{A}_n$ is the best response to a *belief* x_{-n}, if $a_n \in \operatorname{argmax}_{a \in \mathcal{A}_n} u_n(a, x_{-n})$. The notion of rationalizable action is defined in a recursive way: an action is rationalizable if it is the best response to a "rational" belief supported on other agents' rationalizable actions. This situation happens under the common knowledge of rationality (despite incomplete knowledge of other agents' strategy) — that is, every agent is rational, every agent thinks that every agent is rational, every agent thinks that every agent thinks that every agent is rational, and so on in any higher order beliefs. We formalize it in the following definition.

Definition 6 (Rationalizable Actions [Osborne and Rubinstein, 1994])**.** *Suppose that there exists* $\{\mathcal{Z}_n \subseteq \mathcal{A}_n\}_{n=1}^{|\mathcal{N}|}$ *such that for any* $n' \in \mathcal{N}, a_{n'} \in \mathcal{Z}_{n'}$, $a_{n'}$ *is a best response to some belief supported only on* $\mathcal{Z}_{-n'}$*. In this case, we say an action* $a_n \in \mathcal{A}_n$ *is rationalizable for agent* n*. The solution concept of* rationalizability *is formed by* any *strategy profile that supports only on the rationalizable actions.*

Problem Setup We study how multiple agents can learn to rationalize under noisy bandit payoff feedback in an radically uncoupled fashion. At each round $t \in [T]$, each agent n takes an action $a_n(t)$, which together forms the action profile $\boldsymbol{a}(t)$. Then, agent n individually observes from the environment a noisy, bandit feedback, $u_n(\boldsymbol{a}(t)) + \xi_{n,t}$, that is, its payoff under the action profile $\boldsymbol{a}(t)$ perturbed by a noise $\xi_{n,t}$. With a slight abuse of notation, we denote $u_n(t) = u_n(\boldsymbol{a}(t)) + \xi_{n,t}$ hereinafter. We assume $\{\xi_{n,t}, \mathcal{F}_t\}_0^{+\infty}$ is a Martingale Difference Sequence (MDS) with finite variance. Specifically, $\{\xi_{n,t}, \mathcal{F}_t\}_0^{+\infty}$ satisfies: 1) Zero-mean: $\mathbb{E}[\xi_{n,t}|\mathcal{F}_{t-1}] = 0$ for all $t = 1, 2, \ldots (a.s.)$; 2) Finite variance: $\exists \sigma > 0$ s.t.

6. The original definition of (independent) rationalizability [Bernheim, 1984, Pearce, 1984] requires the *belief* to be a product of independent probability measures on each of the action sets $\mathcal{A}_{n'}$ for $n' \in \mathcal{N} \setminus \{n\}$. Under this more restricted notion of rationalizability, the Theorem 5.1 in the next section holds only in two-player games. However, throughout this paper we will instead focus on the more commonly adopted definition of (correlated) rationalizability, [Brandenburger and Dekel, 1987, Milgrom and Roberts, 1990].

$\mathbb{E}[\|\xi_{n,t}\|^2|\mathcal{F}_{t-1}] \leq \sigma^2$ for all $t = 1, 2, \ldots (a.s.)$.

Learning Goals Given the above problem setting, we focus on designing an online learning algorithm that when all agents use such an algorithm, they will learn to play rationalized strategy with high probability in a last-iterate convergence manner. We require that each agent's learning rule is radically (or strongly) uncoupled [Hart and Mas-Colell, 2003, Foster and Young, 2006] in the sense that an agent's strategy has to be adaptive to her own historical observations and meanwhile cannot depend on all other agents' historical actions and payoffs. Intuitively, the motivation of uncoupled learning is to insure that any agent's strategy cannot be inferred by other agents. This crucially requires each agent's strategy to depend on his own information, even conditioned on other agents' information. Note that the coordination-based algorithm of [Wang et al., 2023] is *not* uncoupled because an agent's strategy can be inferred from others' strategy profile at any time according the coordination rule specified by their algorithms. Conceptually similar requirement is imposed in many works on multi-agent learning [7] to avoid the degeneracy to tailored learning rules of computing approximated equilibria from estimated game payoffs. We also remark that many of the prior works [8] in multi-agent learning either focus on full information or gradient feedback, or consider the time-average convergence, our learning objective of last-iterate convergence under noisy bandit feedback provides arguably the strongest and most stable guarantee.

5.4 On the Pursuit of Rationalizability

The notion of rationalizability and dominance elimination has been less popular to the machine learning community than other game-theoretical solution concepts such as NE and CE. Thus to motivate our seemingly "unconventional" pursuit of the rationalizability, we

7. See [Daskalakis et al., 2011, Farina et al., 2022, Anagnostides et al., 2022b, Cai et al., 2023].

8. See [Hannan, 1957, Freund and Schapire, 1999, Daskalakis et al., 2011, Cherukuri et al., 2017, Cohen et al., 2017a, Syrgkanis et al., 2015, Mertikopoulos and Zhou, 2019, Mazumdar et al., 2020].

devote this section to highlight its theoretical and conceptual importance, as well as its potential real-world applications.

5.4.1 Key Properties of Rationalizability

Rationalizability turns out to be equivalent to iterated dominance elimination due to the elegant minimax theorem, as formally described by the following theorem.

Theorem 5.1. *[Osborne and Rubinstein, 1994] The set of any agent n's rationalizable actions are precisely the set of agent n's actions that survive the process of IESDS.*

The proof of the above theorem hinges on the following observation. That is, any agent n's action $\widehat{a}_n$ is a strictly dominated action if and only if there exists *no* belief μ_n over all other agent's actions that makes $\widehat{a}_n$ a best response of agent n (and thus $\widehat{a}_n$ cannot be rationalizable). To see this, note that $\widehat{a}_n$ is a strictly dominated action implies there exists some mixed strategy x_n such that $u_n(x_n, a_{-n}) - u_n(\widehat{a}_n, a_{-n}) > 0$ for any a_{-n}. Equivalently, $\max_{x_n \in \Delta_{\mathcal{A}_n}} \min_{a_{-n}} [u_n(x_n, a_{-n}) - u_n(\widehat{a}_n, a_{-n})] > 0$. By the linearity of $u_n(x_n, a_{-n})$ in x_n and strong duality, we know $\min_{\mu_n \in \Delta_{\mathcal{A}_{-n}}} \max_{a_n \in \mathcal{A}_n} [u_n(a_n, \mu_n) - u_n(\widehat{a}_n, \mu_n)] > 0$. That is, for any agent n's belief $\mu_n \in \Delta_{\mathcal{A}_{-n}}$, there is an action a_n that is strictly better than $\widehat{a}_n$ and thus $\widehat{a}_n$ can never be a best response to any agent n's belief.

From the perspective of epistemic game theory, rationalizability is a solution concept that only assumes the common knowledge of rationality, while CE additionally assumes the common prior, i.e., a correct belief of other players' strategy profile [Brandenburger and Dekel, 1987, Aumann, 1987]. These observations lead to an important fact that the set of rationalizability is a superset of correlated equilibria.

Corollary 5.2. *[Osborne and Rubinstein, 1994] For any finite game, every action used with positive probability in a CE is rationalizable.*

This is also implied by Theorem 5.1, as no CE should put non-zero probability on any iteratively dominated action profile and thus must be rationalizable.

In addition, if the process of IESDS terminates with a single action profile in the remaining action space, this action profile must be the unique NE and also the unique CE of the game [Viossat, 2008]. In this case, game $\mathcal{G}$ is called *mixed-strategy solvable* [Alon et al., 2021a] — a more general notion than the classic dominance solvable game that is defined on dominance elimination by pure strategy.

Corollary 5.3. *For any mixed-strategy solvable game, the only rationalizable action profile coincides with the unique NE (thus the unique CE) of the game.*

Rationalizability also has nice predictions for a large class of economic games, known as the supermodular games [Topkis, 1979, Milgrom and Roberts, 1990], which encompass many well-known games such as Cournot duopoly [Cournot, 1838] and Bertrand competition [Bertrand, 1883]. Generally speaking, a finite, normal-form game is supermodular if each agent n's utility function $u_n(a_n, a_{-n})$ has *increasing difference* in a_n and a_{-n}, i.e., for all $a'_n \geq a_n$ and $a'_{-n} \geq a_{-n}$, $u_n(a'_n, a'_{-n}) - u_n(a_n, a'_{-n}) \geq u_n(a'_n, a_{-n}) - u_n(a_n, a_{-n})$. This require the set $\mathcal{A}_n, \mathcal{A}_{-n}$ to be partially ordered, e.g., as the price or effort level. [9] The formal definition of supermodular games is deferred to Appendix D.1. One example of supermodular game is the bank run model [Diamond and Dybvig, 1983], when more depositors withdraw their funds from a bank, it is better for other depositors to do the same (see Section 5.4.2 for other examples).

Theorem 5.4. *[Milgrom and Roberts, 1990] For any supermodular game, the largest (resp. smallest) rationalizable strategy profile coincides with the largest (resp. smallest) NE of the game.*

9. More generally, when players have multi-dimensional strategy spaces, $\mathcal{A}_n$ must be a complete lattice and u_n is supermodular in a_n for any fixed a_{-n}. When u_n is twicely differentiable, it is equivalent to have $\partial^2 u_n / \partial a_n \partial a_m \geq 0, \forall m \neq n$.

The proof of the above theorem uses a fixed point argument: there exists a function $f : \mathcal{A} \to \mathcal{A}$ such that, from the largest action profile $\boldsymbol{a}^0$ in $\mathcal{A}$, iteratively setting $\boldsymbol{a}^{t+1} = f(\boldsymbol{a}^t)$ leads to a fix point, $\boldsymbol{a}^* = \lim_{i\to\infty} \boldsymbol{a}^t$, which is also the largest NE of the game. Specifically, we can construct $f(a_1, a_2, \cdots, a_{-N}) = (\overline{\beta_i}(a_{-1}), \overline{\beta_i}(a_{-2}), \cdots, \overline{\beta_i}(a_N))$, where $\overline{\beta_i}(a_{-n})$ denotes the largest element in agent n's best response set, $\text{argmax}_{a\in\mathcal{A}}\, u_n(a, a_{-n})$. The sequence $\{\boldsymbol{a}^t\}$ is non-increasing in t by induction: Since $\boldsymbol{a}^0$ is the largest action profile, $\boldsymbol{a}^1 \leq \boldsymbol{a}^0$. Given that $\boldsymbol{a}^t \leq \boldsymbol{a}^{t-1}$, we have for any $n \in \mathcal{N}$, $a_n^{t+1} = \overline{\beta_i}(a_{-n}^t) \leq \overline{\beta_i}(a_{-n}^{t-1}) = a_n^t$, since $\forall a_n \in \mathcal{A}_n, u_n(a_n, a_{-n}^t) - u_n(a_n^{t+1}, a_{-n}^t) \leq u_n(a_n, a_{-n}^{t-1}) - u_n(a_n^t, a_{-n}^{t-1})$ by supermodularity. The fixed point $\boldsymbol{a}^* = f(\boldsymbol{a}^*)$ is an NE by definition, as $a_n^* = \overline{\beta_i}(a_{-n}^*), \forall n \in \mathcal{N}$.

Meanwhile, $\boldsymbol{a}^*$ is also the largest rationalizable action profile, as any action $a_n > a_n^*$ is iteratively dominated. That is, for any agent n, any of its action $a_n > a_n^1$ is dominated, since $u_n(a_n, a_{-n}) - u_n(a_n^1, a_{-n}) \leq u_n(a_n, a_{-n}^0) - u_n(a_n^0, a_{-n}^0) < 0$. By induction, given that any $a_{-n} > a_{-n}^{t-1}$ is (iteratively) dominated, any action $a_n > a_n^t = \overline{\beta_i}(a_{-n}^{t-1})$ is iteratively dominated: by supermodularity, $\forall a_{-n} \leq a_{-n}^{t-1}, u_n(a_n, a_{-n}) - u_n(a_n^t, a_{-n}) \leq u_n(a_n, a_{-n}^{t-1}) - u_n(a_n^t, a_{-n}^{t-1}) < 0$, where the last inequality is strict as $a_n^t \notin \text{argmax}_{a\in\mathcal{A}}\, u_n(a, a_{-n}^{t-1})$.

Similarly, we can show that starting from the smallest action profile and iteratively taking the smallest best response mapping $\underline{\beta_i}(\cdot)$ would lead to the smallest NE. This proof at its core is built upon two seminal results: Tarski's fixed point theorem [Tarski, 1955] that for any monotone function on a complete lattice, the set of all its fix points forms a sublattice; Topkis' monotonicity theorem [Topkis, 1979] that for any supermodular game, each agent's best response function is monotone.

5.4.2 Notable Examples

Here we list several well-known game instances where rationalizability are desirable outcomes with important economics implications. The first two classes of games are dominance solvable games, and the other two are supermodular games.

The Market for "Lemons" [Akerlof, 1978] Consider a market of used cars with a buyer and N sellers. Each seller i has a car of quality q_i, and two actions $a_i \in \{1, 0\}$, respectively, to list or not to list his car. Without loss of generality, let $q_N > q_{N-1} > \cdots > q_1$. The buyer has his action $p \in \mathcal{P}$ from a set of prices to buy a car from sellers. Suppose the buyer and sellers move simultaneously. As assumed by Akerlof, the buyer has no information about each seller's car quality before posting the price. The seller also decides whether or not to list his car (i.e., $a_i = 1$ or 0) without knowing the price. In our experiments, we assume seller i has a reservation value $\widetilde{q}_i = q_i + \epsilon$ which is a noisy perception of his car quality with zero-mean noise ϵ. For those who did choose to list their cars, if the buyer's price are below their reservation value, they would refuse to sell, but still suffers a small and fixed opportunity cost c_1. We denote $b_i = \mathbb{1}[\widetilde{q}_i \leq p]$ as whether the car of seller i gets sold. In contrast, the buyer is uninformed of the car quality he could buy, though the trade would generate welfare so that her revenue is a multiplier c_2 of the average quality $\overline{q} = \frac{\sum_{i\in[N]} b_i \cdot q_i}{\sum_{i\in[N]} b_i}$. Hence, each seller i receives payoff $u_i = a_i(b_i(p - q_i) - c_1)$, and the buyer receives payoff $u_0 = c_2 \cdot \overline{q} - p$, where $c_1 > 0, c_2 > 1$ are game parameters.

Our Proposition D.4 in Appendix D.5 shows that this game has elimination length at least $2N - 1$ for small c_1 and its NEs are that the buyer sets a price no higher than q_1, and all sellers refuse to list. This outcome is certainly not a surprise and was observed by Akerlof as a *market collapse*, due to asymmetric information among sellers and buyers which gradually drives the high quality car sellers out of the market in order. The additional insight of our Proposition D.4 is that this market collapse may follow a long procedure of iterated dominance elimination. We remark that the opportunity cost $c_1 > 0$ is only assumed to capture how much sellers prefer not selling if the price just matches their values. Our results hold for $c_1 = 0$ as well, but will need to work with weakly dominance elimination with a tie breaking in favor of not listing.

Decentralized Matching under Aligned Preference [Niederle and Yariv, 2009] In a decentralized matching market game, each firm (on one side) simultaneously make an offer to a single worker (on the other side), and the workers accordingly decides whether to accept, reject or defer the offers they receive. Ferdowsian et al. [2020] show that if the firms and workers' preferences are aligned (i.e., firms ranks the workers identically and vice versa), then the outcome of stable matching in this market is rationalizable, and moreover, the unique NE surviving iterated elimination of weakly dominated strategies. This rationalizable strategy profile coincides with the celebrated deferred acceptance algorithm by Gale and Shapley [1962].

Bertrand Competitions [Bertrand, 1883] This is a fundamental economic model of competition . It considers a market of N firms that produces identical goods (i.e., substitutes). Each firm n produces with constant unit costs c_n and sets a price a_n from some (possibly discretized) interval $[0, a_{\max}]$. Each firm has a utility function $u_n(a_n, a_{-n}) = (a_n - c_n) \cdot D_n(a_n, a_{-n})$, where $D_n(a_n, a_{-n})$ is the demand function with its elasticity being a non-increasing function of the other firms' prices a_{-n}. Milgrom and Roberts [1990] shows that many common forms of demand such as linear function, logit function satisfies this property and the corresponding Bertrand competitions are therefore supermodular games.

Arms Races [Milgrom and Roberts, 1990] In this game, the players are two countries engaged in an arms race. Each player chooses a level of arms a_n from some (possibly discretized) interval $[0, a_{\max}]$ and receives as its payoff $u_n(a_n, a_{-n}) = -C(a_n) + B(a_n - a_{-n})$, where C is the cost function based on the arm level and B is the welfare function which is concave w.r.t. the difference of arm level. That is, the marginal return to additional arms is an increasing function of the foe's armament level. This is a supermodular game, since for all $a'_n \geq a_n$ and $a'_{-n} \geq a_{-n}$, $u_n(a'_n, a'_{-n}) - u_n(a_n, a'_{-n}) \geq u_n(a'_n, a_{-n}) - u_n(a_n, a_{-n}) \iff B(a'_n - a'_{-n}) - B(a_n - a'_{-n}) \geq B(a'_n - a_{-n}) - B(a_n - a_{-n})$, due to the concavity of B.

5.5 Formal Barriers of Multi-Agent Learning towards Rationalizability

Perhaps surprisingly, we show that even for dominance solvable games under full information feedback, standard bandit learning algorithms will necessarily take *exponentially* many rounds to eliminate all dominated actions. Since this barrier is already significant in two-player games, in this section we shall focus on the two-player case with a finite action set $[K]$. We refer to the row player as agent A, column player B, and use indices $i, j \in [K]$ to denote their pure actions.

5.5.1 "Diamond in the Rough" – A Benchmark Game for Multi-Agent Learning

We introduce an interesting class of two-player *dominance solvable* games, which serves as a challenging benchmark for multi-agent learning. For reference convenience, we call it "*Diamond in the Rough*", whose meaning should become clear in the following formal definition.

Definition 7 (The Diamond-In-the-Rough (DIR) Games)**.** *A diamond-in-the-rough (DIR) game is a two-player game parameterized by (K, c). Each agent have K actions and utility function*

$$u_1(i,j) = \begin{cases} i/\rho & i \le j+1 \\ -c/\rho & i > j+1 \end{cases}, \qquad u_2(i,j) = \begin{cases} j/\rho & j \le i \\ -c/\rho & j > i \end{cases}. \tag{5.1}$$

where $c > 0$ and $\rho = \max\{K, c\}$ is for normalization purpose. Hence, the payoff matrix of

the DIR(K, c) game is given by

$$\frac{1}{\max\{K,c\}}\begin{bmatrix} (1,1) & (1,-c) & (1,-c) & \cdots & (1,-c) \\ (2,1) & (2,2) & (2,-c) & \cdots & (2,-c) \\ (-c,1) & (3,2) & (3,3) & \cdots & (3,-c) \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & (K-1,K-1) & (K-1,-c) \\ (-c,1) & (-c,2) & \cdots & (K,K-1) & (K,K) \end{bmatrix}. \tag{5.2}$$

The DIR game exhibits a "nested" dominance structure, which makes it challenging to play rationally. Specifically, observe that A's action 2 dominates action 1. However, this is not the case for B since if A were to play action 1, B's utility $-c$ of action 2 is significantly worse than utility 1 of action 1. Nevertheless, after A eliminates action 1, B's action 2 starts to dominate B's action 1. This property holds in general for DIR game — B's action $j+1$ dominates her action j *only when* A eliminates his actions $\{1, \cdots, j\}$ and similarly A's action $i+1$ dominates action i *only when* B eliminates her actions $\{1, \cdots, i-1\}$. In the end, the real "diamond" is hidden at the action pair (K, K), which is the best for both agents. However, to find this "diamond", both agents have to sequentially remove all the "rough" actions $1, 2, \cdots, K-1$. This is thus the name "*Diamond in the Rough*".

As we will demonstrate both theoretically and empirically, *the DIR game highlights a fundamental challenge in multi-agent learning* — i.e., whether an agent's action is good or bad depends on what actions her opponents take, and such inter-agent externality makes it challenging to learn the optimal decisions. We end this subsection by summarizing a few useful properties of any DIR(K, c) game.

Proposition 5.5. *The following properties hold for any DIR(K, c) game:*

1. *The game is dominance solvable by alternatively eliminating A and B's action in order $1, 2, ...,$ until reaching the last strategy profile (K, K). The elimination length $L_0 =$*

$2K - 2$. [10]

2. *The strategy profile (K, K) is the unique CE (and thus the unique NE as well). Moreover, both agents achieve the maximum possible utility at this equilibrium.*

5.5.2 No-swap Regret $\not\Rightarrow$ Efficient Iterated Dominance Elimination

Our following theorem shows that for *any* small ϵ, there exist DIR($\log(\frac{1}{\epsilon}), c$) games for some constant c, in which an ϵ-CE may *never* play the unique CE (K, K) and, moreover, will lead to a much smaller utility than the agent's equilibrium utility.[11]

Theorem 5.6. *For any $\epsilon > 0$ and any DIR(K, c) game satisfying $\log(1/\epsilon) \leq (2K-2)\log(c)$, the game always has an ϵ-CE which plays the (unique) CE strategy (K, K) with probability 0. Moreover, the welfare of this ϵ-CE is at most $\frac{1+\lceil \log(1/\epsilon)/\log(c) \rceil}{2K}$ fraction of the equilibrium welfare.*

Specifically, by picking $K = \log(1/\epsilon)$ and $\log(c) = 1$, we obtain a DIR($\log(\frac{1}{\epsilon}), c$) game which admits an ϵ-EC that will put 0 probability at the unique CE (K, K) and has utility at most half of the players' equilibrium utilities. This may appear quite counter-intuitive at the first glance, since how come an ϵ-EC be so such "far away" from the real and unique CE. This turns out to be due to a subtle difference — that is, the ϵ in "ϵ-CE" is measuring the ϵ difference in *agent utilities*,[12] whereas the "far away" conclusion reflected in Theorem 5.6 is measuring the true distance between agent's *action probabilities*. Though in the limit as $\epsilon \to 0$, the ϵ-CE will tend to an exact CE, Theorem 5.6 suggests that agent's strategies

10. While the equilibrium is obvious in DIR, we remark that we can easily swap the rows and column making it difficult to determine even the elimination path under bandit feedback and strategic learning setting (without communication).

11. The key property here is that the game payoffs depend on ϵ *logarithmically*. Such an instance with payoffs that depends on ϵ linearly would be less surprising and also easier to construct.

12. A distribution $\pi \in \Delta_{\mathcal{A}}$ over action profiles is a $\epsilon-$CE if $\sum_{a_{-n}} \pi(a_n, a_{-n}) [u_n(a'_n, a_{-n}) - u_n(a_n, a_{-n})] \leq \epsilon$ for any two actions $a_n, a'_n \in \mathcal{A}_n$ and for any player n. When $\epsilon = 0$, this definitions degenerate to the standard CE.

and equilibrium utilities can be far from the exact CE even when $\epsilon > 0$ is extremely small compared to the game payoff. Therefore, the fact that an agent does not have much incentive to deviate when at an ϵ-CE does not imply that the action it plays is close to the (exact) CE action profile, neither implies his utility is close to the CE utility. *This insight also explains why classic no regret learning algorithm designed based on maximizing accumulated rewards may perform poorly for the task for iterated dominance elimination*, which will be formally proved in our next theoretical result as well as further justified in our numerical results in Section 5.7.

To concretize the message in Theorem 5.6, consider a very small diamond-in-the-rough game with $c = K = 10$. With the state-of-the-art $O(\log^4 T/T)$ swap regret algorithm by Anagnostides et al. [2022a], the empirical distribution of the no-regret learning agents is guaranteed to be a $10^{-9}-$CE after $T = 10^{13}$ rounds. However, since $\log(1/\epsilon) = \log(10^9) < (2K-2)\log(c)$, Theorem 5.6 implies that this $10^{-9}-$CE may still never play the equilibrium strategy (K, K) and its welfare is at most $\frac{1+\lceil \log(1/\epsilon)/\log(c)\rceil}{2K} = \frac{1}{2}$ of the equilibrium welfare. Notably, 10^{13} rounds of sequential interactions could take several days for a modern CPU to simulate. Our experimental results in Appendix 5.7 further confirm the mathematical analysis in Theorem 5.6 — our simulation shows that, surprisingly, the player actions generated by no-swap regret algorithms will get stuck on the first few actions for both players and remain far from the unique CE (K, K) even after 10^8 rounds.

5.5.3 Exponential-Time Convergence of Merit-based Algorithms

We now show that a broad and natural class of online learning algorithms, dubbed *merit-based* algorithms, fails to eliminate all iteratively dominated actions efficiently. Roughly speaking, a merit-based algorithm has the following property: at each time step t, if the accumulated reward from action i exceeds that from action j, the learning algorithm will be more likely to play action i than action j. We call online learning algorithms with such

property "merit-based" learning algorithms. Perhaps unsurprisingly, many celebrated learning algorithms are merit-based, e.g., Follow-the-Perturbed-Leader (FTPL), and the entire class of Dual Averaging (DA) algorithms with symmetric mirror maps including Exponential Weight (EW), lazy gradient descent (LGD) and fictitious play (see our detailed discussion in Appendix D.3). Formally, we define merit-based algorithms as follows.

Definition 8. *Let $\boldsymbol{y}_t = (y_1(t), \cdots, y_K(t))$ be the vector that stores the (possibly weighted) accumulated rewards for all the actions before round t, and $\boldsymbol{p}_t = (p_1(t), \cdots, p_K(t))$ be the probability distribution of the algorithm taking each action at round t. We call an algorithm merit-based if $y_i(t) > y_j(t)$ implies $p_i(t) \geq p_j(t)$ for any $i, j \in [K], t > 0$.*

In other words, a merit-based algorithm maps the accumulated score vector $\boldsymbol{y}_t$ at each round to a distribution $\boldsymbol{p}_t = (p_1(t), \cdots, p_K(t)) \in \Delta_K$ with some "order-preserving" function F (whose formal definition can be found in Appendix D.3) and then randomly samples an action from $\boldsymbol{p}_t$ as the next move. The algorithm is also allowed to specify a learning rate sequence $\{\eta_t\}$ to accumulate the total rewards from each round. For convenience of our arguments, we formulate this general class of online learning algorithms into the following Algorithm 5.1.

ALGORITHM 5.1: The Merit-based Algorithm Framework

Input: An order-preserving function $F : \mathcal{Y} \to \Delta_K$, learning rate sequence $\{\eta_t > 0\}$.
$\boldsymbol{y}_1 \leftarrow (0, \cdots, 0)$
for $t = 1 \dots T$ **do**
  Compute $\boldsymbol{p}_t = F(\boldsymbol{y}_t)$
  Draw an action i_t from the distribution $\boldsymbol{p}_t$.
  Receive the expected payoff $\widetilde{\boldsymbol{u}}_t = (\widetilde{u}_1(t), \cdots, \widetilde{u}_K(t))$ for each action i from the first-order oracle.
  Update $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t \widetilde{\boldsymbol{u}}_t$.

Notably, the notion of "merit-based" applies to both the full-information setting (i.e., the rewards of all actions are revealed after taking an action) and the bandit setting (i.e., only the reward of the taken action is revealed). In the following analysis, we show that even with

access to full-information feedback, any merit-based algorithm needs at least exponentially many rounds to eliminate all iteratively dominated actions.

We consider the situation where all agents are running merit-based algorithms with typically adopted non-increasing learning rates.[13] That is, at any round t, each agent will do the standard update for any action i using estimated reward $\widetilde{u}_i(t)$ with learning rate η_t that is non-increasing in t. Perhaps surprisingly, even with perfect gradient feedback (as oppose to the noisy gradient from bandit feedback)[14], agents following the merit-based algorithms will take provably exponential rounds to converge to the unique NE in DIR games.

Theorem 5.7. *Consider the DIR(K, c) game with any $K \geq 3$ and $c \geq 3K^2$. Suppose two agents both follow a merit-based algorithm 5.1 equipped with a non-negative, bounded,* non-increasing *learning rate sequence $\{\eta_t\}_{t=1}^{\infty}$. Then agent B will place at most $1/2$ probability on the (unique) pure NE strategy at any round $t \leq 3^{K-2}$.*[15]

Theorem 5.7 proves the inefficiency of the family of merit-based algorithms with a non-increasing learning rate sequence in terms of eliminating iteratively dominated strategies. We remark that the requirement $c \geq 3K^2$ is only necessary to the proof technique and does not imply the DIR game is easy to solve when $c < 3K^2$. As we will demonstrate later in the experiments, the choice of $c = O(K)$ already results in an extremely slow empirical convergence rate for Exp3 algorithm and its variants.

Additional Discussions on Connection to Related Works. Cohen et al. [2017a] showed that under full information feedback, the EW algorithm with learning rate $\eta_t = \frac{1}{t^b}$

13. Variants of EW (a.k.a., multiplicative weight updates) have been proved to converge to equilibria in other games such as potential games [Cohen et al., 2017b] and concave games [Bravo et al., 2018].

14. The perfect gradient of each agent n in the game is a vector with each entry, $\widetilde{u}_{a_n}(t) = \sum_{a_{-n}} p_{a_{-n}}(t) u_n(a_n, a_{-n})$, equivalent to the expected payoff of each action a_n given the opponents' strategy profile, since the loss is a linear function of the payoff. The noisy gradient estimated from bandit feedback can be found in our newly designed algorithm in the following section.

15. This rules out even the (weaker) average-sequence convergence (in distribution sense) to the unique NE of the game.

for $0 < b < 1$ eliminates dominated actions exponentially fast. They remarked that their proof can be extended by induction to argue that the sequence of play induced by EW will ultimately eliminate all iteratively dominated strategies of the game. Our Theorem 5.7 shows that the number of rounds needed by their algorithm will, unfortunately, be exponential in the elimination length. That is, it is easy to eliminate dominated actions for one iteration, but difficult to iteratively eliminate them all. Cohen et al. [2017a], Mertikopoulos and Zhou [2019] showed the EW algorithm with proper learning rate converges to a pure NE exponentially fast with high probability if that equilibrium satisfies a global variational stability. Our negative result does not contradict their results because the global variational stability condition does not hold in DIR games. The verification of this claim can be found in Appendix D.2.3. [Laraki and Mertikopoulos, 2013] considered the replicator dynamics, the continuous version of EW, and showed the probability of playing any iteratively dominated action shrinks to 0 over time. This result matches ours that all iteratively dominated actions will be removed as $t \to \infty$. However, the convergence rates in continuous-time dynamics are not necessarily meaningful, as t can be reparametrized arbitrarily; it is unclear how a time-dependent rate can be translated to its corresponding rate in a discrete-time framework as a function of iterations, because infinitely many iterations are required to simulate a continuous dynamics. Therefore, our exponential time convergence result is a more explicit characterization of the learning barrier in iterated dominance elimination.

5.5.4 Proof of Theorem 5.7

Overview of the Proof. The proof of Theorem 5.7 is involved and requires new ideas since we are not aware of any result of similar spirit in the literature. We start by highlighting the key ideas here before diving into the technical arguments. Without loss of generality, we consider player B and let the action set be $\{b_1, \cdots, b_K\}$. The key to our proof is to show the existence of a constant $c_0 > 1$ such that if B's action probabilities $p_B(t)$ at time t concentrate

on $\{b_1, \cdots, b_n\}$ for the first T_n rounds, then it must also concentrate on $\{b_1, \cdots, b_{n+1}\}$ for the first $T_{n+1} = c_0 T_n$ rounds. The intuition behind this fact comes from the property of the DIR game: any pure action looks very profitable before it becomes iteratively dominated and thus could have accumulated a very large score before any merit-based algorithm starts to penalize it. As a result, any merit-based algorithm has to suffer additional rounds *proportional to the current time step* to eliminate the next action, which incurs an exponential convergence time. A intricate induction is needed to make this intuition a formal argument, which we formally show next.

Step 1: Preparations for the Proof. We consider the DIR game with $c \geq 3K^2 > K$, so the normalization factor in Equation (5.2) is $1/c$. Both agents follow Algorithm 5.1 with a non-increasing learning rate sequence $\{\eta_t\}_{t=1}^{\infty}$ in a repeated DIR game. Let $u_{A,i}(t), u_{B,i}(t), p_{A,i}(t), p_{B,i}(t)$ denote agent A and agent B's payoffs[16] and probabilities of picking the i-th action at round t, respectively. We further let

$$y_{A,i}(t) = \sum_{s=1}^{t-1} u_{A,i}(s)\eta_s, \qquad y_{B,i}(t) = \sum_{s=1}^{t-1} u_{B,i}(s)\eta_s$$

be the ith arm's accumulated weights of agent A and agent B till round t. Then, according to Algorithm 5.1,

$$(p_{A,1}(t), \cdots, p_{A,K}(t)) = F(y_{A,1}(t), \cdots, y_{A,K}(t)),$$

$$(p_{B,1}(t), \cdots, p_{B,K}(t)) = F(y_{B,1}(t), \cdots, y_{B,K}(t)).$$

Step 2: The Dueling Lemmas and Their Proofs. At a high level, by leveraging the "order-preserving" property from the definition of merit-based algorithms, our proof employs a complex induction argument that if both players have significant probabilities to play the

16. For simplification of notation, we omit the bar "~" above u, but it should still be interpreted as the expected payoff of an action given opponent's strategy profile.

first n actions within time T_n, then they must also have significant amount of probabilities to play the first $n+1$ actions within time $(1+\frac{2c}{3K^2})T_n$. This intuition is formalized by the following two lemmas. The first lemma says that if B has constant probability to play actions from $1, 2, \cdots, n$ within the first $(1+\frac{2c}{3K^2})^{n-1}$ rounds, then any dual averaging algorithm used by agent A must play actions $1, 2, \cdots, n, n+1$ with some constant probability during the same period. The second lemma is an analogous (though subtlely different) conclusion that uses A's behavior to argue B's trajectory properties. It is such dueling between A's and B's behaviors — thus the name of *dueling lemmas* — that make their convergence to rationalizable actions exponentially slow.

We believe these dueling lemmas reveal the intrinsic difficulty of using merit-based algorithms for iterated dominance elimination, and thus provide a formal proofs for them below. The key challenge in their proofs is to manage the inter-agent utility externalities by setting the right parameter scales at which the two agents' strategies duel.

Lemma 5.8. *For any $1 \leq n \leq K-2$, if there exists $T_n \geq 1$ such that $\sum_{j=1}^{n} p_{B,j}(t) \geq \frac{1}{K-n+1}$ for any $t \leq T_n$, then we must have $\sum_{i=1}^{n+1} p_{A,i}(t) \geq \frac{1}{K-n}, \forall t \leq T_{n+1}$ for $T_{n+1} = (1+\frac{2c}{3K^2})T_n$.*

Lemma 5.9. *For any $1 \leq n \leq K-2$, if there exists $T_{n+1} \geq 1$ such that $\sum_{i=1}^{n+1} p_{A,i}(t) \geq \frac{1}{K-n}$ for any $t \leq T_{n+1}$, then we must have $\sum_{j=1}^{n+1} p_{B,j}(t) \geq \frac{1}{K-n}, \forall t \leq T_{n+1}$.*

Proof. **of Lemma 5.8**. At each step $t \leq T_n$, using the assumed condition that $\sum_{j=1}^{n} p_{B,j}(t) \geq$

$\frac{1}{K-n+1}$ and $\frac{K}{c} \leq \frac{1}{3K}$, we have for any $i = n+2, \cdots, K$

$$
\begin{aligned}
u_{A,i}(t) &= \sum_{j=1}^{n} p_{B,j}(t) \cdot (-1) + (1 - \sum_{j=1}^{n} p_{B,j}(t)) \cdot (\frac{i}{c}) && \text{by Def. of DIR games} \\
&\leq \sum_{j=1}^{n} p_{B,j}(t) \cdot (-1) + (1 - \sum_{j=1}^{n} p_{B,j}(t)) \cdot (\frac{K}{c}) \\
&\leq \frac{1}{K-n+1} \cdot (-1) + (1 - \frac{1}{K-n+1}) \cdot (\frac{K}{c}) && \text{by assumed conditions} \\
&\leq -\frac{1}{K-n+1} + \frac{K-n}{K-n+1} \cdot \frac{1}{3K} && \text{since } c \geq 3K^2 \\
&< \frac{-2}{3(K-n+1)} && \text{since } \tfrac{K-n}{K} < 1 \\
&\leq -\frac{2}{3K}. && \tfrac{-2}{3(K-n+1)} \text{ decreases in } n
\end{aligned}
$$

Suppose player A uses any DA algorithm with learning rate $\{\eta_t\}_{t=1}^{\infty}$. Using the above strict upper bound for $u_{A,i}(t) (< -\frac{2}{3K})$, the cumulative rewards of any action $i \in \{n+2, \cdots, K\}$ at time step $t = T_n + 1$ can be upper bounded as

$$
y_{A,i}(T_n + 1) = \sum_{t=1}^{T_n} \eta_t u_{A,i}(t) < -\frac{2\sum_{t=1}^{T_n} \eta_t}{3K}, \quad \forall i = n+2, \cdots, K. \tag{5.3}
$$

Now let $T_{n+1} = (1 + \frac{2c}{3K^2}) \cdot T_n$ and consider any $t \leq T_{n+1}$. First, if $t \leq T_n + 1$, we have $y_{A,i}(t) = \sum_{\tau=1}^{t-1} \eta_\tau u_{A,i}(\tau) < -\frac{2\sum_{\tau=1}^{t-1} \eta_\tau}{3K} \leq 0$. We now consider $T_n + 1 < t \leq T_{n+1}$. For any

$i = n+2, \cdots, K$, we have

$$\begin{aligned}
y_{A,i}(t) &= y_{A,i}(T_n+1) + \sum_{s=T_n+1}^{t-1} \eta_t u_{A,i}(s) \\
&< -\frac{2\sum_{t=1}^{T_n}\eta_t}{3K} + \sum_{s=T_n+1}^{t-1} \eta_{T_n+1} \cdot \frac{K}{c} && \text{by Ineq. (5.3) and } \eta_t \le \eta_{T_{n+1}} \\
&\le -\frac{2\sum_{t=1}^{T_n}\eta_t}{3K} + \frac{K}{c} \cdot \eta_{T_n+1} \cdot (\frac{2c}{3K^2} T_n) && \text{since } t - T_n \le \tfrac{2c}{3K^2} T_n \\
&= -\frac{2\sum_{t=1}^{T_n}(\eta_t - \eta_{T_n+1})}{3K} \le 0, && \text{due to non-increasing learning rate}
\end{aligned}$$

Therefore, $y_{A,i}(t) < 0$ for any $t \le T_{n+1}$. However, note that $y_{A,1}(t) \ge 0$ for any $t > 0$, we thus conclude that $y_{A,i}(t) < y_{A,1}(t)$ for any action $i \ge n+2$ at any time $t \le T_{n+1}$.

Apply the definition of merit-based algorithms, we obtain $p_{A,i}(t) \le p_{A,1}(t)$ and as a result,

$$\begin{aligned}
K - n - 1 &\ge (K-n-1) \cdot [p_{A,1}(t) + \sum_{i=n+2}^{K} p_{A,i}(t)] \\
&\ge \sum_{i=n+2}^{K} p_{A,i}(t) + (K-n-1) \cdot [\sum_{i=n+2}^{K} p_{A,i}(t)] \quad \text{by } p_{A,1}(t) \ge p_{A,i}(t), \forall i \ge n+2 \\
&= (K-n)[\sum_{i=n+2}^{K} p_{A,i}(t)]. \qquad (5.4)
\end{aligned}$$

This implies $\sum_{i=n+2}^{K} p_{A,i}(t) \le \frac{K-n-1}{K-n}, \forall t \le T_{n+1}$ and thus $\sum_{i=1}^{n+1} p_{A,i}(t) \ge \frac{1}{K-n}$, as desired. □

Proof. **of Lemma 5.9**. Using the assumed condition $\sum_{i=1}^{n+1} p_{A,i}(t) \ge \frac{1}{K-n}$ for any $t \le T_{n+1}$,

we can derive an upper bound for $u_{B,j}(t)$ for any $j = n+2, \cdots, K$, as follows

$$\begin{aligned}
u_{B,j}(t) &= \sum_{i=1}^{n+1} p_{A,i}(t) \cdot (-1) + (1 - \sum_{i=1}^{n+1} p_{A,i}(t)) \cdot (\frac{j}{c}) && \text{by Def. of DIR games} \\
&\leq \sum_{i=1}^{n+1} p_{A,i}(t) \cdot (-1) + (1 - \sum_{i=1}^{n+1} p_{A,i}(t)) \cdot (\frac{K}{c}) \\
&\leq \frac{1}{K-n} \cdot (-1) + (1 - \frac{1}{K-n}) \cdot (\frac{K}{c}) \\
&\leq -\frac{1}{K-n} + \frac{K-n-1}{K-n} \cdot \frac{1}{3K} \\
&< -\frac{2}{3(K-n)} \leq -\frac{2}{3K}.
\end{aligned}$$

Hence, we similarly conclude that $y_{B,j}(t) < 0 \leq y_{B,1}(t)$ for any $j \geq n+2$ and $t \leq T_{n+1}$, which yields $p_{B,j}(t) \leq p_{B,1}(t)$. Then we similarly follow Inequality (5.4) to obtain $\sum_{j=n+2}^{K} p_{B,j}(t) \leq \frac{K-n-1}{K-n}$ and thus $\sum_{j=1}^{n+1} p_{B,j}(t) \geq \frac{1}{K-n}, \forall t \leq T_{n+1}$, which completes the proof. □

Step 3: Concluding the Proof. Armed with the dueling lemmas 5.8 and 5.9, we are now ready to conclude the proof. Specifically, the following is a direct corollary of the two lemmas.

Corollary 5.10. *For any $1 \leq n \leq K-2$, if there exists $T_n \geq 1$ such that $\sum_{j=1}^{n} p_{B,j}(t) \geq \frac{1}{K-n+1}$ for any $t \leq T_n$, then for $T_{n+1} = (1 + \frac{2c}{3K^2})T_n$ we must have $\sum_{j=1}^{n+1} p_{B,j}(t) \geq \frac{1}{K-n}$ for any $t \leq T_{n+1}$.*

Note that the "if" condition in the above corollary clearly holds for $n = 1$, in which case we can verify that $T_1 = 1$ satisfies $\sum_{j=1}^{n} p_{B,j}(t) = p_{B,1}(1) \geq \frac{1}{K} = \frac{1}{K-n+1}$ for any $t \leq T_n$) since both players start by taking actions uniformly at random. Applying corollary 5.10 inductively until any $n \leq K-2$, we obtain that $\sum_{j=1}^{n+1} p_{B,j}(t) \geq \frac{1}{K-n}, \forall t \leq T_{n+1} = (1 + \frac{2c}{3K^2})^n T_1$. Plugging in $T_1 = 1$ and the upper bound $K-2$ of allowed value for n, we thus have

$$\sum_{j=1}^{K-1} p_{B,j}(t) \geq \frac{1}{2}, \forall t \leq T_{K-1} = (1 + \frac{2c}{3K^2})^{K-2}$$

Since $c \geq 3K^2$, so $T_{K-1} \geq 3^{K-2}$. The above inequality implies that within the first 3^{K-2} time steps, player B will never play the unique equilibrium action K with probability larger than $1/2$. This rules out the convergence of the two agents' strategies — either in the average sense or last-iterate sense — to the NE within 3^{K-2} time steps. Note that since this statement holds deterministically since we are in the simpler setup in which the agents have full feedback, and thus can fully observe the (expected) payoff (i.e., perfect gradient feedback). Since (K, K) is the unique NE of this game, Moreover, a larger value of c will imply a larger $T_{K-1} = (1 + \frac{2c}{3K^2})^{K-2}$ value and would further slow down the convergence rate.

5.6 `Exp3-DH` and its Efficiency towards Rationalizability

5.6.1 The Exp3 with Diminishing History (`Exp3-DH`) Algorithm

Motivated by the barriers in Section 5.5, we now introduce a novel algorithm *Exp3 with Diminishing History* (`Exp3-DH`) described in Algorithm 5.2, which turns out to provably guarantee efficient elimination of all iteratively dominated actions, with high probability.

`Exp3-DH` is an EW-style algorithm but with the following important characteristics:

1. `Exp3-DH` uses an unbiased payoff estimator $\widetilde{u}_{i_t}(t) = \frac{u_{i_t}(t)}{p_{i_t}(t)}$. This turns out to be crucial since our proof has to upper bound the absolute value $\left|\sum_{t=1}^{T} \gamma_t(\widetilde{u}_a(t) - u_a(t))\right|$ whereas standard single-agent bandit problems only need a one-side upper bound for $\sum_{t=1}^{T} \gamma_t(\widetilde{u}_a(t) - u_a(t))$ and thus a biased conservative estimation of $\widetilde{u}_{i_t}$ there can be helpful (actually, is essential for some algorithms). However, similar conservative reward estimators appear difficult to work in our problem because it may let a dominating action

ALGORITHM 5.2: `Exp3` with Diminishing History (`Exp3-DH`)

1 **Input:** Number of actions K, parameter ϵ_t, β .
2 **Initialization:** $y_i(0) = 0, \forall i \in [K]$.
3 **for** $t = 0, 1, \ldots, T$ **do**
4 Set

$$p_i(t) = (1 - \epsilon_t)\frac{\exp(y_i(t))}{\sum_{j\in[K]} \exp(y_j(t))} + \frac{\epsilon_t}{K} \quad \forall i \in [K].$$

5 Draw action i_t randomly accordingly to the probability distribution $(p_1(t), p_2(t), \ldots p_K(t))$.
6 Observe realized payoff $u_{i_t}(t)$.
7 Compute *unbiased* payoff estimator $\widetilde{u}_i(t) = \frac{u_i(t)}{p_i(t)} \cdot \mathbb{I}(i = i_t), \forall i \in [K]$.
8 Update

$$y_i(t+1) = (\frac{t}{t+1})^\beta \cdot y_i(t) + \widetilde{u}_i(t), \quad \forall i \in [K].$$

lose its advantage. This highlights an interesting difference between eliminating iteratively dominated actions in multi-agent setup and learning optimal actions in single-agent setup.

2. During reward updates at Step 8, `Exp3-DH` will always discount *each action*'s previous reward $y_i(t)$ by a factor $(\frac{t-1}{t})^\beta$ for some carefully chosen parameter β, *regardless of whether this action is taken at this round or not*. Therefore, the rate of historical rewards will gradually diminish, which thus leads to the name of our algorithm.

Effective Learning Rates. Note that the update in Step 8 of Algorithm 5.2 only captures the recursive relation between $y_i(t+1)$ and $y_i(t)$. From this recursion, we can easily derive how $y_i(t)$ depends on all previous payoff estimation $\widetilde{u}_i(\tau)$ for $\tau = 0, 1, \cdots, t$, which is the follows,

$$y_i(t+1) = \sum_{\tau=0}^{t} (\frac{\tau}{t})^\beta \widetilde{u}_i(\tau). \tag{5.5}$$

For notational convenience, we call $\gamma_\tau^{(t)} = (\tau/t)^\beta$ the *effective* learning rate. Notably, the learning rate $\gamma_\tau^{(t)}$ for any fixed past payoff estimation $\widetilde{u}_i(\tau)$ *dynamically decreases* as the round t becomes large. This means that the weight of the historical estimation $\widetilde{u}_i(\tau)$ will become smaller and smaller as time goes. In other words, the algorithm exhibits *recency*

bias and gradually "forgets" histories and always relies more on recent payoff estimations.

We end this section by comparing `Exp3-DH` with previous Exp3-style algorithms. In standard Exp3 algorithm [Auer et al., 2002], the *effective* learning rate is exactly its learning rate γ_t, which is set to a constant $\sqrt{\frac{\log K}{KT}}$ or $O(\sqrt{\frac{\log K}{Kt}})$ to guarantee a sub-linear regret. Some other variants[17] of Exp3 use non-increasing effective learning rate γ_t. However, `Exp3-DH` differs from these algorithms in at least two key aspects: (1) its effective learning rate $\gamma_\tau^{(t)}$ is increasing in τ, i.e., biased towards recent reward estimations; (2) $\gamma_\tau^{(t)}$ will be re-scaled every time t increases, i.e., a new observation comes. The first deviation is justified by our Theorem 5.7 since DA with any decreasing learning rate necessarily suffer exponential convergence. We note that increasing learning rate has been recently studied for single-agent bandit problems [Bubeck et al., 2017, Lee et al., 2020, Agarwal et al., 2017] and for faster convergence to coarse correlated equilibrium in games [Syrgkanis et al., 2015].[18] Unfortunately, our experimental results in Appendix 5.7 show that these algorithms fail to efficiently eliminate all iteratively dominated actions, which illustrates that careful design of learning rate is necessary for efficient iterative dominance elimination.

5.6.2 Efficient Convergence of `Exp3-DH` under Noisy Bandit Feedback

We now present the theoretical guarantee for `Exp3-DH`. Due to randomness, no algorithm guarantees dominance elimination for certain. Thus, we introduce a natural notion of *essential elimination*:

Definition 9 (*ε-Essential Elimination*)**.** *We say that an action $i \in \mathcal{A}_n$ is* ε-essentially eliminated *at time step T if agent-n's probability of playing i satisfies $p_i(T) \leq \frac{\varepsilon}{4KN}$.*

Note that if all the actions in E_{L_0} are *essentially eliminated* at time step T, the mixed

17. See [Neu, 2015, Mertikopoulos and Zhou, 2019, Cohen et al., 2017b, Bravo et al., 2018]

18. Syrgkanis et al. [2015] use optimistic follow the regularized leader (OFTRL) with recency bias. The obtained algorithm with entropy regularizer can be viewed as EW with dynamically increasing learning rate.

strategy $x(T)$ given by the probability distribution $p(T)$ of all agents satisfies $\|x(T)-x^*\|_1 \leq \frac{\varepsilon}{4KN} \cdot 2(KN-N) < \frac{\varepsilon}{2}$. Therefore, ε-*Essential Elimination* implies last-iterate convergence. We are now ready to give the convergence guarantee for `Exp3-DH` in Theorem 5.11:

Theorem 5.11. *Consider any game $\mathcal{G}(\mathcal{N}, \mathcal{A}, u)$ with elimination length L_0 and elimination set $\{E_l\}_{l=1}^{L_0}$. Suppose all agents in $\mathcal{N}$ run `Exp3-DH` with parameters $\beta > 0$ and $\epsilon_t = t^{-b}$ for some $b \in (0,1)$. Then for any $\varepsilon, \delta \in (0, \frac{1}{2})$, any action in $E_l \setminus E_{l-1}$ will be* ε-essentially eliminated *with probability at least $1 - 2|E_l|(T_l+s)^2\delta, \forall s \geq 0$, from time step $t = T_l$ to $t = T_l + s$, where the sequence $\{T_l\}_{l=1}^{L_0}$ is defined recursively below:*

1. *T_1 is an integer such that for any $t \geq T_1$,*

$$\frac{t^{-b}}{K} + \exp\left(4\left(\sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b}}\right)\log^{\frac{1}{2}}\frac{2K}{\delta}\cdot t^{\frac{1+b}{2}} - \frac{\Delta t}{16(1+\beta)}\right) < \frac{\min\{\varepsilon, \Delta/2\}}{4KN}. \quad (5.6)$$

2. *For any $l \geq 1$,*

$$T_{l+1} \geq \max\left\{(1+\frac{8}{\Delta})^{\frac{1}{1+\beta}}\cdot T_l, T_l + \frac{(1+\beta)^2(4+\Delta)^2(8+\Delta)^2}{4(1+2\beta)\Delta^2}\log\frac{1}{\delta}\right\}. \quad (5.7)$$

Specifically, if all agents run `Exp3-DH` on $\mathcal{G}$, then all iteratively dominated actions will be ϵ-essentially eliminated after T_{L_0} number of rounds in the last iterate convergence sense with probability at least $1 - 2KNT_{L_0}^2\delta$.

To interpret Equation (5.6) and (5.7) in Theorem 5.11, it will be easier if we focus only on its leading terms. It is helpful (though not necessary) to think of $\beta \approx L_0$ and $b \approx 1/2$. When t is large, the exponential term in Equation (5.6) will be dominated by t^{-b}/K term since its exponent decreases exponentially in t. Therefore, we would expect T_1 to have order around $(N\min\{\varepsilon, \Delta\}^{-1})^{1/b}$. In Equation (5.7), the second term in the "max" is usually dominated by the first term, therefore T_{L_0} roughly has order $\Delta^{-L_0/\beta}T_1$. Therefore, Theorem 5.11 indicates: 1. `Exp3-DH` needs polynomially many rounds to ε-essentially eliminate the

first set of dominated actions in E_1; 2. once a set of iteratively dominated actions are ε-essentially eliminated, `Exp3-DH` takes polynomially additional steps to eliminate the next set of iteratively dominated actions. Therefore, `Exp3-DH` guarantees to ε-essentially eliminate all iteratively dominated actions in polynomially many rounds. Specifically, the following corollary gives an explicit and formal upper bound for T_{L_0}. Its proof is deferred to Appendix D.4.

Corollary 5.12. *If all agents run* `Exp3-DH` *on game* $\mathcal{G}(\mathcal{N}, \mathcal{A}, u)$ *with parameters* $\beta > 0$ *and* $\epsilon_t = t^{-\frac{1}{3}}$*, then after* $\widetilde{O}(\max\{N^3, K^{1.5}\} \max\{\varepsilon^{-3}, \Delta^{-3}\}(1+\sigma^3)\beta^{1.5} \log^{1.5} \frac{1}{\delta})$ *number of rounds*[19]*, all the iteratively dominated actions in* $\mathcal{G}$ *will be* ε-essentially eliminated *with probability at least* $1-\delta$.

Parameter Choices. The optimal choice of β depends on Δ and L_0, which reveal the intrinsic difficulty of iterative dominance elimination in $\mathcal{G}$. However, if Δ, L_0 is unknown, a conservative choice of β is the total number of agent actions KN. This always guarantees polynomial round convergence, concretely, in $\widetilde{O}(\max\{N^3, K^{1.5}\} \max\{\varepsilon^{-3}, \Delta^{-3}\}\beta^{1.5} \log^{1.5} \frac{1}{\delta})$ rounds. We also note that Corollary 5.12 instantiated a choice of $b = 1/3$. This is to balance the dependence on (K, N) and (ε, Δ). If $b = 1/2$, the upper bound would be $\widetilde{O}(\max\{N^2, K^2\} \max\{\varepsilon^{-2}, \Delta^{-4}\}\beta^2 \log^2 \frac{1}{\delta})$. When $2 = N \ll K$ and $\Delta \sim O(K^{-1})$ in DIR games, a smaller b is preferred. This will be demonstrated in our experiments. More generally, a good choice of b will depend on the game structures.

Corollary 5.12 also captures the convergence rate of ϵ as a function of time horizon T, given $b = 1/3$. Specifically, $\epsilon = O(1/\sqrt[3]{T})$ with the caveat that omitted constants in the big O depends on the game parameters N, K, Δ, δ. If we choose $b = 1/2$, the convergence rate will be $\epsilon = O(1/\sqrt{T})$, however its dependence on the game parameters will be worse.

19. By convention, $\widetilde{O}$ notation omits logarithmic terms.

5.6.3 Proof of Theorem 5.11

Overview of the Proof The proof of Theorem 5.11 employs induction to prove the convergence of uncoupled learning. We are not aware of similar ideas in previous multi-agent learning setups, and thus present its formal proof in the next section in case it is of independent interest. Our proof deviates from standard single-agent online learning analysis in at least two key aspects. First, since the effective learning rates are dynamically changing in `Exp3-DH`, the concentration of the estimated $y_i(t)$ needs to be re-evaluated at each round. To do so, we need to construct a sub-martingale for each round t and argues its high-probability concentration. The second major distinction is that our proof needs to bound the *difference* of the estimated reward for any iteratively *dominating* action i and *dominated* action j, whereas standard (single-agent) bandit learning only needs to upper bound the reward estimation for each action i. In the latter case, conservative reward estimation (e.g., in Exp3.P [Auer et al., 2002]) is helpful. However, a conservative reward estimation may make i lose its dominance advantage in our problem. Fortunately, due to the unbiased reward estimation in `Exp3-DH`, we can use Azuma's inequality to bound the concentration from both sides.

Our proof employs an induction argument. In the base case, we demonstrate that `Exp3-DH` requires polynomially many rounds to ε-essentially eliminate the first set of dominated actions E_1. Subsequently, in the induction phases, we provide further evidence that `Exp3-DH` takes polynomially additional steps, with high probability, to eliminate the next set of iteratively dominated actions. This leads to a polynomial time last-iteration convergence. The key insight driving each induction phase is that the accumulated scores of the remaining iteratively dominated actions can always be upper-bounded by the number of steps taken so far (which constitutes a polynomial term). As a result, correcting such misconceptions by choosing an appropriate decaying rate to discount the history also takes no more than polynomial time. We present the detailed formal argument next.

Step 1: Preparations for the Proof. Since we will be working mostly on agent's actions, we use notation a and x to denote an agent's pure and mixed strategy in this proof. Our proof relies on the following two technical lemmas, the proofs of which are deferred to Appendix D.4 so that they will not distract the core arguments. The first lemma (Lemma 5.13) is standard and bounds the difference between the estimated $\widetilde{u}_a(t)$ and the realized true $u_a(t)$) via the Azuma's inequality. The second lemma (Lemma 5.14) is more involved and shows that the accumulated weighted rewards for any dominated action a will be far less than the expected accumulated weighted rewards of any mixed strategy $x \in \Delta_{\mathcal{A}_n}$ that dominates a.

Lemma 5.13. *For any $a \in \mathcal{A}, T > 0$ and any sequence $\{\gamma_t > 0\}_{t=1}^T$, with probability $1-\delta$, we have*

$$|\sum_{t=1}^{T} \gamma_t^{(T)}(\widetilde{u}_a(t) - u_a(t))| < 2\left(\sqrt{K(1+\sigma^2)\log\frac{2}{\delta}\cdot\sum_{t=1}^{T}\frac{(\gamma_t^{(T)})^2}{\epsilon_t}}\right). \tag{5.8}$$

Lemma 5.14. *For any fixed $T > 0$, if an action $a \in \mathcal{A}_n$ is strictly dominated by a mixed strategy $x = (x_1, \cdots, x_K) \in \Delta_{\mathcal{A}_n}$, then*

$$y_x(T+1) - y_a(T+1) \geq \sum_{t=1}^{T}\gamma_t^{(T)}[u_x(t) - u_a(t)] - 4\left(\sqrt{\log\frac{2K}{\delta}}\cdot\sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{(\gamma_t^{(T)})^2}{\epsilon_t}}\right) \tag{5.9}$$

holds with probability $1-\delta$. In particular, if $\gamma_t = (t/T)^\beta, \epsilon_t = t^{-b}$ and $T > \beta$, we have

$$y_x(T+1) - y_a(T+1) \geq \frac{\Delta\cdot T}{1+\beta} - 4\left(\sqrt{\log\frac{2K}{\delta}}\cdot\sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b}\cdot T^{1+b}}\right). \tag{5.10}$$

Armed with the above two lemmas, the proof of Theorem 5.11 follows a carefully tailored induction argument.

Step 2: Proof of the Base Case. First, we prove the elimination of all actions in E_1 in

the first iteration. For any dominated action pair $x \succ a$ (i.e., x dominates a), we conclude from Lemma 5.14 that with probability $1-\delta$, the probability of playing a satisfies

$$\begin{aligned}
p_a(t) &= \frac{\epsilon_t}{K} + \frac{\exp(y_a(t))}{\sum_{a\in\mathcal{A}_n}\exp(y_{a'}(t))}(1-\epsilon_t) \\
&\le \frac{t^{-b}}{K} + \frac{\exp(y_a(t))}{\sum_{a'\in\mathcal{A}_n} x_{a'}\exp(y_{a'}(t))} && \text{since } x_{a'}\in[0,1] \\
&\le \frac{t^{-b}}{K} + \frac{\exp(y_a(t))}{\exp(\sum_{a'\in\mathcal{A}_n} x_{a'}y_{a'}(t))} && \text{by convexity of } exp(\cdot) \\
&= \frac{t^{-b}}{K} + \frac{\exp(y_a(t))}{\exp(y_x(t))} \\
&< \frac{t^{-b}}{K} + \exp\left(4\sqrt{\frac{eK(1+\sigma^2)\log\frac{2K}{\delta}}{1+2\beta+b}}\cdot t^{\frac{1+b}{2}} - \frac{\Delta t}{1+\beta}\right) && \text{by Eq (5.10)} \\
&< \frac{t^{-b}}{K} + \exp\left(4\left(\sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b}}\right)\log^{\frac{1}{2}}\frac{2K}{\delta}\cdot t^{\frac{1+b}{2}} - \frac{\Delta t}{16(1+\beta)}\right) && (5.11) \\
&< \frac{\min\{\varepsilon,\Delta/2\}}{4KN}, && (5.12)
\end{aligned}$$

where the last Inequality (5.12) holds for any $t \ge T_1$ by the definition of T_1. Therefore, from the union bound we conclude that with probability at least $1-|E_1|(T_1+s+1)\delta > 1-2|E_1|(T_1+s)^2\delta$, actions in E_1 will be ε'-essentially eliminated at each round in $\{T_1,\cdots,T_1+s\}$, where $\varepsilon' = \min\{\varepsilon,\Delta/2\}$.

Step 3: Proof of the Induction Step. We now move to the more involved induction step. Assume there exists T_k such that for any $s>0$, actions in E_k are ε'-essentially eliminated during $t\in(T_k,T_k+s]$ with probability $1-2|E_k|(T_k+s)^2\delta$, i.e.,

$$\mathbb{P}\Big[p_a(t)\le\frac{\varepsilon'}{4KN},\forall a\in E_k,\forall T_k<t\le T_k+s\Big]\ge 1-2|E_k|(T_k+s)^2\delta,\forall s>0.$$

We investigate how many additional iterations we need to ε'-essentially eliminate actions in

E_{k+1}. In particular, we will show that if T_{k+1} satisfies Eq (5.7), then

$$\mathbb{P}\Big[p_a(t) \le \frac{\varepsilon'}{4KN}, \forall a \in E_{k+1}, \forall T_{k+1} < t \le T_{k+1} + s\Big] \ge 1 - 2|E_{k+1}|(T_{k+1}+s)^2\delta, \forall s > 0,$$

which then complete our proof of the theorem.

By the definition of E_k, if no agents play any action in E_k, then for any action $a \in E_{k+1} \setminus E_k$ of agent i there must exist a mixed strategy x of agent i that dominates a. Assuming `Exp3-DH` has run $T_k + T_0$ steps, we derive a sufficient condition for T_0 such that `Exp3-DH` can ε-essentially eliminate actions in E_{k+1} starting for any $t > T_k + T_0$. In particular, we show that for sufficiently large T_0 and any $s > 0$, with probability at least $1-2|E_{k+1}|(T_k+T_0+s)^2\delta$, actions in E_{k+1} are ε'-essentially eliminated from steps T_k to $T_k + T_0 + s$.

First of all, for any $a \in E_{k+1} \setminus E_k$ that is iteratively dominated by some x and any $t \in (T_k, T_k+T_0+s]$, from Lemma 5.14 we conclude with probability $1-(|E_{k+1}|-|E_k|)(T_0+s)\delta$, it holds that

$$\begin{aligned}
&y_x(t+1) - y_a(t+1) \\
\ge &\sum_{s=1}^{t} \gamma_s^{(t)}[u_x(s) - u_a(s)] - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{s=1}^{t}\frac{(\gamma_s^{(t)})^2}{\epsilon_s}}\right) \\
\ge &\sum_{s=T_k+1}^{t} \gamma_s^{(t)}[u_x(s) - u_a(s)] - 2\sum_{s=1}^{T_k}\gamma_s^{(t)} - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{s=1}^{t}\frac{(\gamma_s^{(t)})^2}{\epsilon_s}}\right),
\end{aligned} \tag{5.13}$$

where the last inequality used the bound $u_x(s) - u_a(s) \ge -2$ for any s due to bounded utilities.

In order to further lower bound the RHS of Eq (5.13), we need the following lemma:

Lemma 5.15. *If T_0 satisfies that*

$$T_0 \geq \max\Big\{\frac{(1+\beta)^2(4+\Delta)^2(8+\Delta)^2}{4(1+2\beta)\Delta^2}\log\frac{1}{\delta}, \Big[(1+\frac{16}{\Delta})^{\frac{1}{1+\beta}}-1\Big]\cdot T_k\Big\}, \tag{5.14}$$

then for any $s \geq 0, T' = T_0 + s$ and $T = T' + T_k$, we have for any $a \in E_{k+1} \setminus E_k$, with probability at least $1 - (|E_{k+1}| - |E_k|)T'\delta$,

$$\sum_{t=T_k+1}^{T_k+T'} \gamma_t^{(T)}[u_x(t) - u_a(t)] > \frac{\Delta}{4}\sum_{t=T_k+1}^{T_k+T'} \gamma_t^{(T)}. \tag{5.15}$$

The intuition behind Eq (5.15) is, since actions in E_k are rarely played during $t \in (T_k, T_k + T_0 + s]$, $u_x(t) - u_a(t) \geq \Delta$ must hold with high probability in this entire period. A standard technique to prove this is to use a union bound. However, this idea does not work here as it requires the undesirable event (i.e., x_{-i} still contains certain essentially eliminated actions at a specific round) to happen with an extremely small probability $O(\frac{1}{T_k+T_0})$, which we cannot afford unless with an exponentially large T_k. To overcome this challenge, we take a different route and construct a sub-martingale regarding the utilities. The detailed proof of Lemma 5.15 is technical and we defer it to Appendix D.4.

Now let $T_{k+1} = T_k + T_0$. The probability of Eq (5.15) to hold for any $T' \in [T_0, T_0 + s]$ is thus at least $(1 - (|E_{k+1}| - |E_k|)\delta\sum_{t=T_0}^{T_0+s} t)$. Substitute Eq (5.15) into Eq (5.13) and apply a union bound, we conclude that with a probability at least

$$\begin{aligned}
&1 - 2|E_k|(T_k + T_0 + s)^2\delta - (|E_{k+1}| - |E_k|)(T_0 + s)\delta - (|E_{k+1}| - |E_k|)\delta\sum_{t=T_0}^{T_0+s} t \\
\geq &1 - 2|E_{k+1}|(T_k + T_0 + s)^2\delta = 1 - 2|E_{k+1}|(T_{k+1} + s)^2\delta,
\end{aligned}$$

the following inequality holds for any $T \in [T_{k+1}, T_{k+1} + s]$ (we omit the superscript (T) on γ_t in the following derivations):

$$
\begin{aligned}
& y_x(T+1) - y_a(T+1) \\
& \geq \sum_{t=T_k+1}^{T} \gamma_t [u_x(t) - u_a(t)] - 2\sum_{t=1}^{T_k} \gamma_t - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{\gamma_t^2}{\epsilon_t}}\right) && (5.16) \\
& \geq \left(\frac{\Delta}{8}\sum_{t=T_k+1}^{T}\gamma_t - 2\sum_{t=1}^{T_k}\gamma_t\right) + \left(\frac{\Delta}{8}\sum_{t=T_k+1}^{T}\gamma_t - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{\gamma_t^2}{\epsilon_t}}\right)\right) \\
& \geq 0 + \left(\frac{\Delta}{16}\sum_{t=T_k+1}^{T}\gamma_t + \frac{1}{2}\sum_{t=1}^{T_k}\gamma_t - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{\gamma_t^2}{\epsilon_t}}\right)\right) && (5.17) \\
& > \frac{\Delta}{16}\sum_{t=1}^{T}\gamma_t - 4\left(\sqrt{\log\frac{2K}{\delta}} \cdot \sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{\gamma_t^2}{\epsilon_t}}\right) && (5.18) \\
& > \frac{\Delta t}{16(1+\beta)} - 4\sqrt{\left(\frac{eK(1+\sigma^2)}{1+2\beta+b}\right)\log\frac{2K}{\delta}} \cdot t^{\frac{1+b}{2}}, && (5.19)
\end{aligned}
$$

where Eq (5.17) holds because it is straightforward to verify $\frac{\Delta}{16}\sum_{t=T_k+1}^{T}\gamma_t^{(T)} - \sum_{t=1}^{T_k}\gamma_t^{(T)} \geq 0$ when $T_0 \geq \left[(1+\frac{16}{\Delta})^{\frac{1}{1+\beta}} - 1\right] \cdot T_k$, Eq (5.18) holds because $1 \geq \Delta > \frac{\Delta}{8}$, and Eq (5.19) holds because of Lemma 5.14. Therefore, from Eq (5.11) and Eq (5.12) and the fact $T_k > T_1$ we conclude that all the actions in E_{k+1} will be ε-essentially eliminated during round $t \in (T_{k+1}, T_{k+1}+s]$ with probability $1 - 2|E_{k+1}|(T_{k+1}+s)^2\delta$ as long as T_0 satisfies Eq (5.14). By induction, we complete the proof.

5.7 Empirical Evaluations

Baselines We compare `Exp3-DH` with a rich collection of basic or state-of-the-art online learning algorithms listed below:

(a) The classical Exp3 algorithm;

(b) Exp3.P which has no regret with high probability [Auer et al., 2002];

(c) OMWU is the optimistic variant of MWU to obtain faster convergence to coarse correlated equilibria in games [Daskalakis et al., 2021];

(d) BM-OMWU applies the black-box reduction technique of Blum and Mansour (BM) [Blum and Mansour, 2005] onto OMWU, which guarantees faster convergence to $\epsilon-$CE [Anagnostides et al., 2022a];[20]

(e) The online mirror descent algorithm with log barrier regularizer, OMD-LB [Foster et al., 2016], which also uses increasing learning rate schedule [Lee et al., 2020, Agarwal et al., 2017, Bubeck et al., 2017].

We let all learning agents follow the same type of learning algorithm, i.e., self-play, in games described below and compare their convergence trend.

Metrics To measure the progress of iterative dominance elimination, we define the notion of *elimination distance* (ED) for any action i, $\Lambda(i)$, as the number of elimination iterations needed before this action start to be eliminated. Formally, $\Lambda(i) \equiv \arg\max_{0\leq l\leq L_0}\{i \notin E_l\}$, where $E_0 = \varnothing$. The elimination distance of any undominated action is L_0, the elimination length (see Definition 5). We let $\sum_{i\in\mathcal{A}_n} x_n(i)\frac{\Lambda(i)}{L_0}$ be the normalized ED of mixed strategy x_n of any agent n. We then introduce the *Progress of Elimination* (PoE) metric, as the normalized elimination distance aggregated over all agents in the game at round t,

$$\text{PoE}(t) = \frac{1}{N}\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{A}_n} p_{n,i}(t)\cdot\frac{\Lambda(i)}{L_0} \in [0,1].$$

Therefore, the larger PoE is, the more actions the learning agents have eliminated. When PoE reaches 1, the learning agents have removed all dominated actions and converged to the desirable set of rationalizable actions.

20. The original algorithms in the paper are designed for full information feedback setting. To ensure fair comparison, we modify the OMWU and BM-OMWU to use the standard bandit estimator in EXP3.

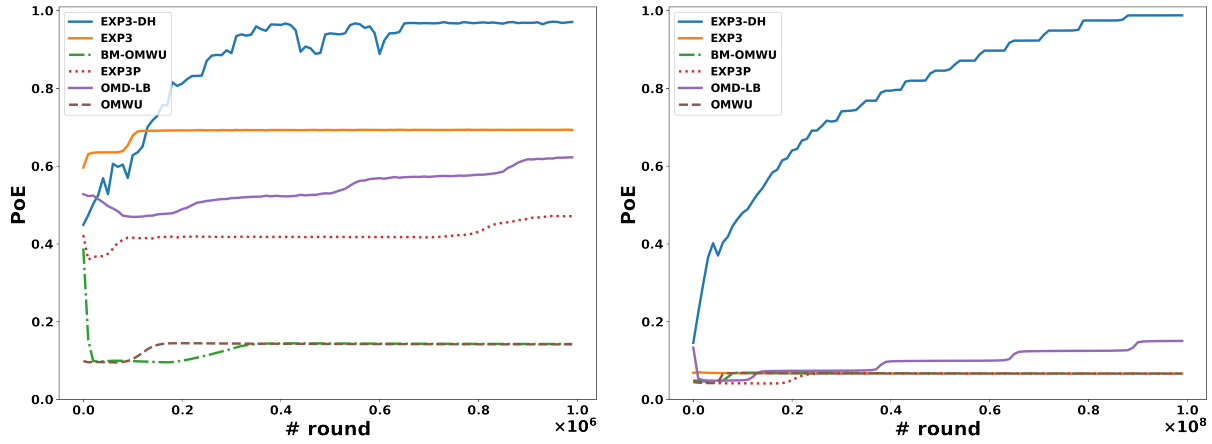


Figure 5.1: Progress of Elimination (PoE) in a smaller **DIR**$(10, 20)$ **game (left)** over $T = 10^6$ rounds and a larger **DIR**$(20, 40)$ **game (right)** over $T = 10^8$ rounds. In both games, i.i.d. Gaussian noise with std. 0.1 is added onto agents' payoffs. The performance of `Exp3-DH` is represented by blue solid line while five baseline algorithms are represented by other notations shown in the legend.

DIR game To illustrate our theoretical results, we conduct a set of experiments in the DIR game, where all agents following the same type of learning algorithm. For DIR game, we use $b = 0.2, \beta = 2K \approx L_0$ as the parameter of `Exp3-DH`. In Figure 5.1, we can clearly observe that `Exp3-DH` exhibits a superior performance in both cases, and enjoys a greater advantage in a larger game instance, where the other baselines even struggle to eliminate the first few dominated actions. In addition, OMD-LB with increasing learning rate also displays relatively good performance especially in the larger and harder instance, compare to other learning algorithms with non-increasing learning rate. But in the harder instance of DIR game with just 20 actions on the right, all the baseline algorithms can only do elimination for 5 or 6 iterations out of $2K - 2 = 38$ iterations.

The Market for "Lemons" (see Section 5.4.2) We also examine the algorithm performance on this famous example of the adverse selection problem by Akerlof [1978]. Our numerical experiments aim at testing how fast the market will collapse, if every seller i have noisy and bandit perception $\widetilde{q}_i$ of their car's true quality. In our experiment instances, we set $c_1 = 3, c_2 = 1.5$, i.i.d. noise $\epsilon \sim \mathcal{N}(0, 5)$. Respectively, we choose $N = K - 1 = 50$ or 200 and

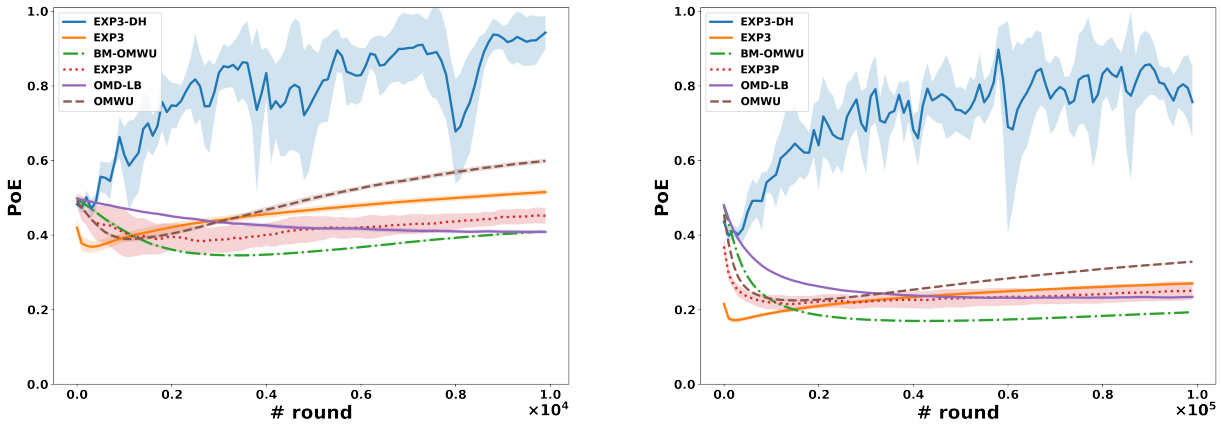


Figure 5.2: Progress of Elimination (PoE) in **The Market for "Lemons"** with 50 sellers (Left) or 200 sellers (Right). In this game, i.i.d. Gaussian noise with std. 0.1 is added onto agents' payoffs. The lightly shaded region displays the error bar of each convergence trend (by one standard deviation over 5 runs).

let each $q_i = N/2 + i$, $\mathcal{P} = \{N/2, \cdots, 3N/2\}$. We let the buyer and sellers apply no-regret learning algorithm to learn the equilibrium price and listing decisions from only the noisy bandit feedback in a repeated game. For `Exp3-DH`, we set $b = 0.5, \beta \approx L_0$, since N is of the same order with K in this game. In Figure 5.2, as predicted by our theoretical results, with all learning agents running `Exp3-DH` algorithm, the convergence can be polynomial, whereas the convergences from existing no-regret learning algorithms are comparatively slow.

Performance in Presence of Adversarial Agents We would also like to investigate the robustness of `Exp3-DH` beyond the multi-agent learning setup. We conduct test the learning algorithms in two setups: one is to interface with the adversarial opponent in the DIR game that randomly plays a fixed action for every 1000 rounds, the other is to run in the non-stationary environment where the arms' reward distributions change every 1000 rounds. Such a switching frequency of 1000 is particularly chosen for the adversarial purpose such that the learning algorithms are not best responding to a completely random reward distribution nor adapting to a periodic distribution change.

We plot the average regret incurred different algorithms in these two setups in Figure 5.3.

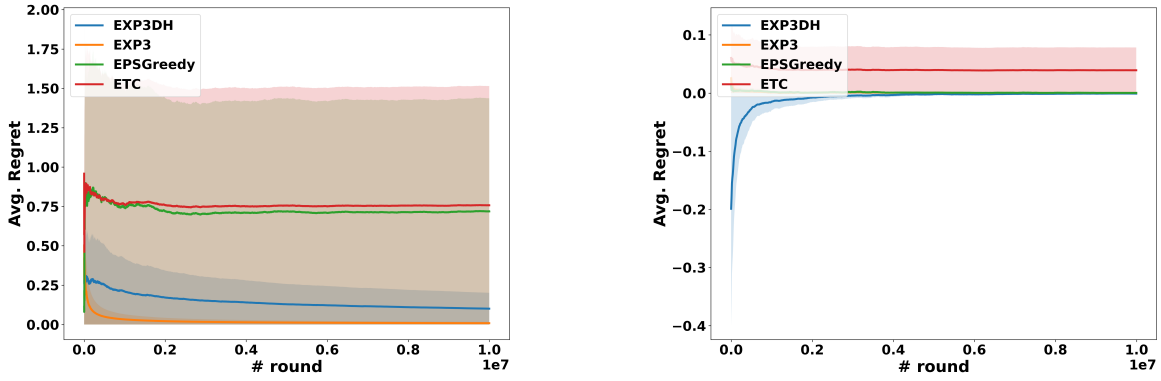


Figure 5.3: The history of average regret facing adversarial opponent in **DIR**$(10, 20)$ game (Left), in non-stationary environment (Right). In these games, i.i.d. Gaussian noise with std. 0.1 is added onto agents' payoffs. The lightly shaded region displays the error bar of each convergence trend (by one standard deviation over 10 runs).

To showcase the robustness of `Exp3-DH` over the iterative best response approach by [Wang et al., 2023], we also implement the ϵ-greedy (EPSGreedy) algorithm and the explore-then-commit (ETC) algorithm that both periodically explores and commit to the best arm.[21] As demonstrated by the empirical results in Figure 5.3, both ETC and EPSGreedy suffer in these adversarial environments, limiting their applications to well-specified multi-agent learning setups — almost linear mean regret with large variance in both setups. In contrast, the proposed `Exp3-DH` shows strong performance in both settings, matching or even outperforming the Exp3 algorithm, which has the provable no-regret guarantee in adversarial settings. That said, we are still able to construct special instances where even `Exp3-DH` suffers linear regret, and this points us towards an important open direction to design algorithms with better robustness guarantee in adversarial settings.

21. The exact algorithm in [Wang et al., 2023] is specified for a "centralized" learning setup (i.e., asking one agent to explore while the remaining agents follow the same action profile) and cannot be directly used for this experiment under the uncoupled learning setup. The ϵ-greedy and ETC algorithms are arguably the closest variants of their iterative best response approach.

5.8 Final Remarks

Our study formalizes the price of "over-hedging" of standard no-regret learning algorithms in the process of iterated dominance elimination. Such price is especially expensive in games whose equilibria are hidden in the "rough". In order to overcome this pitfall, we design a diminishing-history mechanism that deliberately balance the exploitation of the existing knowledge and indifference to history. However, there are still several open questions that remain for future works.

One direction is to understand the lower bound of convergence rate and whether there exists uncoupled learning algorithms with provably faster convergence guarantee. In addition, Wang et al. [2023] propose a learning algorithm with better convergence rate yet relying on agent's coordination, it is also interesting to understand the trade-off between the communication bandwidth and convergence rate. Moreover, despite we have proved the learning barrier of the merit-based algorithm, it is unclear whether online mirror descent (OMD), another family of no regret learning algorithm, suffers the similar learning barrier.[22] Finally, another important direction is to design the "best of both worlds" algorithm such that the rate is optimal both in regret and in last-iterate convergence to equilibrium. This raises a fundamental question on the existence of an intrinsic gap between the single-agent no-regret learning objective and the multi-agent equilibrium learning.

22. We conjecture the answer is negative, at least for a subset of regularizers within OMD that we have already excluded: note that the DA class and OMD class overlap as EW can be interpreted as an OMD method with an entropic regularizer. Additionally, in Appendix 5.7, we empirically show that OMD with log-barrier regularizer [Foster et al., 2016] also suffers from exponentially slow convergence.

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation develops a foundational framework for addressing strategic alignment problems in AI—issues that arise when learning algorithms interact with rational agents in environments shaped by incentives and information asymmetries. As AI systems gain influence over high-stakes decisions, misalignments between algorithmic objectives and stakeholder interests pose significant challenges to social welfare, robustness, and trust. To mitigate these risks, we integrate insights from machine learning, game theory, and economic design, advancing both theoretical models and practical algorithms that align incentives in complex, multi-agent environments. Collectively, these contributions aim to build more responsible, rational, and cooperative AI systems. Looking ahead, this dissertation opens several promising avenues for future work.

Towards Sustainable AI Ecosystems As AI systems increasingly operate as economic agents, a central question emerges: how can we model the *economics of intelligence and data*, and how can this understanding guide the design of systems that promote efficiency, fairness, and long-term sustainability?

The first step is to develop theoretical models that capture how intelligence is produced, valued, and distributed across the whole AI ecosystems. This involves understanding the marginal utility of agentic services, the economic roles of data, and the implications of model fine-tuning, knowledge transfer and competitions.

- **Values of agentic capabilities.** A foundational task is to establish principled models for pricing AI services based on their marginal contribution, task complexity, and substitutability with human labor. For example, in digital marketplaces where AI agents generate summaries or visual designs, pricing mechanisms should reflect performance variation while ensuring competitive balance and fair compensation.

- **Tensions in data ecology.** A critical distinction exists between the *training data* that serves as input capital and the *generated content* that functions as productive output. Understanding this separation is essential for addressing attribution, licensing, and competitive dynamics between human and synthetic content producers.

- **Competitions under model distillation.** Techniques such as model distillation and fine-tuning redistribute value across models and institutions. Analyzing trade-offs along the Pareto frontier (e.g., balancing size, latency, and capability) can inform licensing strategies and help mitigate distortions in competitive dynamics.

Building on these modeling foundations, the next step is to develop actionable infrastructure: pricing mechanisms, marketplaces, attribution tools, and contractual frameworks that govern how intelligence is exchanged, monetized, and regulated.

- **Optimal pricing of AI services.** Designing dynamic and incentive-compatible pricing mechanisms requires balancing factors such as demand elasticity, user heterogeneity, and information asymmetry. Tools from contract theory and market design can help prevent monopolistic practices while supporting innovation and fair access.

- **Attribution and profit sharing.** Cooperative game-theoretic methods, such as Shapley value–based attribution, enable the fair distribution of value among contributors in collaborative learning settings, including federated learning and synthetic data generation.

- **Contractual and licensing mechanisms.** Long-term relationships between data providers, model developers, and downstream users require robust contractual frameworks. These may include usage-based royalties, performance-contingent licensing, and strategies to limit strategic leakage through model reverse-engineering or distillation.

Together, these research directions aim to enable the design of sustainable, auditable marketplaces in which intelligence is treated as a tradable economic asset — aligning incentives across model developers, data contributors, and end users within the broader AI ecosystem.

Mechanism Design with AI Agents Two complementary directions lie at the intersection of agentic AI and mechanism design: developing AI agents that assist in designing mechanisms, and designing mechanisms to coordinate the behavior of AI agents.

On one front, recent advances open the door to a new design paradigm in which AI agents not only participate in mechanisms but actively support their development. This approach integrates economic theory with simulation-based experimentation and large language model (LLM)–driven agent modeling, fostering a tighter feedback loop between theoretical insight and empirical evaluation. Two key applications exemplify this perspective:

- **AI to help humans make rational decisions.** AI agents can assist human decision-making in complex environments — such as civic deliberation, contract negotiation, or public policy design — by forecasting outcomes, simulating tradeoffs, and detecting inconsistencies. For instance, language agents can help participants in participatory budgeting understand the consequences of proposed allocations.

- **AI agent simulation for empirical studies of mechanism design.** One can develop language-based and reinforcement-learning-based simulations of strategic agents to empirically evaluate mechanisms under realistic, boundedly rational behavior. These simulations allow us to test auction designs or deliberation protocols at scale before deployment, improving robustness and fairness.

On the other front, as AI agents increasingly make decisions on behalf of users — such as in bidding, routing, or information exchange — multi-agent coordination and incentive alignment become central design challenges. Mechanisms must be developed to ensure that

the collective behavior of these agents remains efficient, fair, and aligned with broader societal goals. Illustrative applications include:

- **Ad auctions with autonomous bidders.** Strategic interactions between platforms and LLM-powered advertisers in sponsored search and display advertising raise new questions about revenue optimization, fairness, and market efficiency. Mechanism design must evolve to address the increasing sophistication of bidders and their impact on platform dynamics.

- **Route allocation in autonomous mobility systems.** Self-driving vehicles operated by decentralized AI agents can contribute to congestion or adversarial behavior in traffic networks. Coordination mechanisms—such as congestion pricing, adaptive tolling, or intent-sharing protocols—can promote safe, efficient, and cooperative routing.

- **Platform design in the presence of AI intermediaries.** As users increasingly access digital platforms via AI assistants (e.g., Copilot, ChatGPT), platform incentives and API governance must be designed to withstand strategic behavior by intermediaries, ensuring user utility while safeguarding against manipulation.

In conclusion, this dissertation advances a unified and principled framework for aligning incentives in learning systems. These future applications reflect my broader agenda: to weave strategic reasoning and incentive alignment into the very fabric of agentic AI design. As AI agents grow more capable, we face the opportunity and the responsibility to shape them not only as intelligent but also as cooperative participants in our digital institutions. By integrating incentive design with information structure and embedding strategic reasoning into agent behaviors, it charts a path toward AI systems that promote positive societal outcomes, uphold human values, and serve the broader public good.

REFERENCES

Bennett's inequality for martingales. URL http://www.stat.yale.edu/~pollard/Books/Mini/BasicMG.pdf.

Creator earnings report breakdown, where are we in the creator economy? https://neoreach.com/creator-earnings/. Accessed: 2024-05-18.

The creator economy. https://www.goldmansachs.com/intelligence/pages/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html. Accessed: 2024-05-18.

Exploring the potential of the creator economy. https://about.fb.com/news/2022/11/exploring-the-potential-of-the-creator-economy/. Accessed: 2024-05-18.

Tiktok lite, a new app quietly released in france that rewards screen time. https://www.lemonde.fr/en/pixels/article/2024/04/13/tiktok-lite-a-new-app-quietly-released-in-france-that-rewards-screen-time_6668286_13.html. Accessed: 2024-05-18.

How surge pricing works. https://www.uber.com/us/en/drive/driver-app/how-surge-works/. Accessed: 2022-09-06.

Getty Images (US), Inc. v. Stability AI, Inc. (1:23-cv-00135). Online, 2023. URL http://tinyurl.com/getty-image.

The New York Times Company v. Microsoft Corporation (1:23-cv-11195). Online, 2023. URL http://tinyurl.com/ms-to-nyt.

Chatgpt-maker openai signs deal with ap to license news stories. Online, 2023. URL http://tinyurl.com/chatgpt-newsAP.

Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data. Online, 2023. URL https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year.

Introducing the gpt store. Online, 2024. URL https://openai.com/blog/introducing-the-gpt-store.

Reddit in AI content licensing deal with Google. Online, 2024. URL https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Dilip Abreu and Hitoshi Matsushima. Virtual implementation in iteratively undominated strategies: complete information. *Econometrica: Journal of the Econometric Society*, pages 993–1008, 1992.

Balazs Aczel, Barnabas Szaszi, and Alex O Holcombe. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6: 1–8, 2021.

Sydney N Afriat. The construction of utility functions from expenditure data. *International economic review*, 8(1):67–77, 1967.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pages 10–4, 2019.

Michele Aghassi and Dimitris Bertsimas. Robust game theory. *Mathematical programming*, 107(1):231–273, 2006.

Rajeev Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

AJE. Peer review: how we found 15 million hours of lost time, 2018. URL https://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time.

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier, 1978.

Jef Akst. I hate your paper. *The Scientist*, 24(8):36–41, 2010.

Jeff Allen. Misinformation amplification analysis and tracking dashboard. *Integrity Institute, October*, 13, 2022.

Noga Alon and Joel H Spencer. *The probabilistic method*. John Wiley & Sons, 2016.

Noga Alon, Kirill Rudov, and Leeat Yariv. Dominance solvability in random games, 2021a.

Tal Alon, Paul Dütting, and Inbal Talgam-Cohen. Contracts with private cost per unit-of-effort. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 52–69, 2021b.

Ingo Althöfer. On sparse approximations to randomized strategies and convex combinations. *Linear Algebra and its Applications*, 199:339–355, 1994.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1169–1177, 2013.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 622–630, 2014.

Kareem Amin, Rachel Cummings, Lili Dworkin, Michael Kearns, and Aaron Roth. Online learning and profit maximization from revealed preferences. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Bo An, David Kempe, Christopher Kiekintveld, Eric Shieh, Satinder Singh, Milind Tambe, and Yevgeniy Vorobeychik. Security games with limited surveillance. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*, 2012.

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022a.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems*, 2022b.

Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *International Conference on Machine Learning*, pages 536–581. PMLR, 2022c.

Gary Anderson. US R&D increased by $72 billion in 2021 to $789 billion; estimate for 2022 indicates further increase to $886 billion. NSF 24-317. *National Science Foundation*, 2024. URL https://ncses.nsf.gov/pubs/nsf24317/.

Simon P Anderson and Andre De Palma. The logit as a model of product differentiation. *Oxford Economic Papers*, 44(1):51–67, 1992.

Gabriel P Andrade, Rafael Frongillo, and Georgios Piliouras. Learning in matrix games can be arbitrarily complex. In *Conference on Learning Theory*, pages 159–185. PMLR, 2021.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Robert Aumann and Adam Brandenburger. Epistemic conditions for Nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1161–1180, 1995.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Robert J Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

Robert Axelrod and William D Hamilton. The evolution of cooperation. *science*, 211(4489): 1390–1396, 1981.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Yaron Azrieli and Dan Levin. Dominance-solvable common-value large auctions. *Games and Economic Behavior*, 73(2):301–309, 2011.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 554–562, 2013.

Moshe Babaioff, Brendan Lucier, and Noam Nisan. Bertrand networks. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 33–34, 2013.

Moshe Babaioff, Shaddin Dughmi, Robert Kleinberg, and Aleksandrs Slivkins. Dynamic pricing with limited supply. *ACM Transactions on Economics and Computation (TEAC)*, 3(1):1–26, 2015.

Michael Bacharach. Beyond individual choice. In *Beyond Individual Choice*. Princeton University Press, 2018.

Yoram Bachrach, Tor Lattimore, Marta Garnelo, Julien Perolat, David Balduzzi, Thomas Anthony, Satinder Singh, and Thore Graepel. Multiagent reinforcement learning in games with an iterated dominance solution, 2020. URL https://openreview.net/forum?id=ryl1r1BYDS.

Ashwinkumar Badanidiyuru, Kshipra Bhawalkar, and Haifeng Xu. Targeting and signaling in ad auctions. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2545–2563. SIAM, 2018a.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018b.

Gal Bahar, Omer Ben-Porat, Kevin Leyton-Brown, and Moshe Tennenholtz. Fiduciary bandits. In *International Conference on Machine Learning*, pages 518–527. PMLR, 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of Stackelberg equilibria in general-sum games. *Advances in Neural Information Processing Systems*, 34: 25799–25811, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Maria-Florina Balcan, Amit Daniely, Ruta Mehta, Ruth Urner, and Vijay V Vazirani. Learning economic parameters from revealed preferences. In *International Conference on Web and Internet Economics*, pages 338–353. Springer, 2014.

Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pages 61–78, 2015.

Melinda Baldwin. Scientific autonomy, public accountability, and the rise of "peer review" in the cold war united states. *Isis*, 109(3):538–558, 2018.

Siddhartha Banerjee, Kamesh Munagala, Yiheng Shen, and Kangning Wang. Fair price discrimination. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2023.

U Bardi. The decline of science: why scientists are publishing too many papers. cassandra's legacy. 2014, 2014. URL https://cassandralegacy.blogspot.com/2014/08/the-decline-of-science-we-are.html.

Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Siddharth Barman, Umang Bhaskar, Federico Echenique, and Adam Wierman. The empirical implications of rank in bimatrix games. In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 55–72, 2013.

Siddharth Barman, Umang Bhaskar, Federico Echenique, and Adam Wierman. On the existence of low-rank explanations for mixed strategy behavior. In *International Conference on Web and Internet Economics*, pages 447–452. Springer, 2014.

Stephen Bates, Michael I Jordan, Michael Sklar, and Jake A Soloff. Principal-agent hypothesis testing. *arXiv preprint arXiv:2205.06812*, 2022.

Paulina Beato and Andreu Mas-Colell. On marginal cost pricing with given tax-subsidy rules. *Journal of Economic Theory*, 37(2):356–365, 1985.

Curtis Bechtel, Shaddin Dughmi, and Neel Patel. Delegated pandora's box. *arXiv preprint arXiv:2202.10382*, 2022.

Eyal Beigman and Rakesh Vohra. Learning from revealed preference. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 36–42, 2006.

Richard Bellman. On the theory of dynamic programming. *Proceedings of the national Academy of Sciences*, 38(8):716–719, 1952.

Richard Bellman. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

Alexandre Belloni, Tengyuan Liang, Hariharan Narayanan, and Alexander Rakhlin. Escaping the local minima via simulated annealing: Optimization of approximately convex functions. In *Conference on Learning Theory*, pages 240–265, 2015.

Yoshua Bengio. Time to rethink the publication process in machine learning. *Retrieved January*, 21:2021, 2020.

Dirk Bergemann and Stephen Morris. Belief free incomplete information games. 2007.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Ulrich Berger. Two more classes of games with the continuous-time fictitious play property. *Games and Economic Behavior*, 60(2):247–261, 2007.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, Alberto Marchesi, Francesco Trovò, and Nicola Gatti. Optimal rates and efficient algorithms for online bayesian persuasion. In *International Conference on Machine Learning*, pages 2164–2183. PMLR, 2023.

B Douglas Bernheim. Rationalizable strategic behavior. *Econometrica: Journal of the Econometric Society*, pages 1007–1028, 1984.

Joseph Bertrand. Review of "theorie mathematique de la richesse sociale" and of "recherches sur les principles mathematiques de la theorie des richesses.". *Journal de savants*, 67:499, 1883.

Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. Is google getting worse? a longitudinal investigation of seo spam in search engines. In *Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024)(Lecture Notes in Computer Science). Springer*, 2024.

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

Hemant K Bhargava. The creator economy: Managing ecosystem supply, revenue sharing, and platform design. *Management Science*, 68(7):5233–5251, 2022.

Jakub Bielawski, Thiparat Chotibut, Fryderyk Falniowski, Grzegorz Kosiorowski, Michał Misiurewicz, and Georgios Piliouras. Follow-the-regularized-leader routes to chaos in routing games. In *International Conference on Machine Learning*, pages 925–935. PMLR, 2021.

R Selten Bielefeld. Reexamination of the perfectness concept for equilibrium points in extensive games. In *Models of Strategic Rationality*, pages 1–31. Springer, 1988.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A*, 5:147–154, 1946.

Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco Marmolejo, Ninad Rajgopal, and Alexandros Voudouris. Optimally deceiving a learning leader in Stackelberg games. *Advances in Neural Information Processing Systems*, 33, 2020.

Avrim Blum and Yishay Mansour. From external to internal regret. In *International Conference on Computational Learning Theory*, pages 621–636. Springer, 2005.

Avrim Blum, Vijay Kumar, Atri Rudra, and Felix Wu. Online learning in online auctions. *Theoretical Computer Science*, 324(2-3):137–146, 2004.

Avrim Blum, Eyal Even-Dar, and Katrina Ligett. Routing without regret: On convergence to Nash equilibria of regret-minimizing algorithms in routing games. *Theory of Computing*, 6(1):179–199, 2010.

Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems*, pages 1826–1834, 2014.

Avrim Blum, Nika Haghtalab, MohammadTaghi Hajiaghayi, and Saeed Seddighin. Computing Stackelberg equilibria of large general-sum games. In *International Symposium on Algorithmic Game Theory*, pages 168–182. Springer, 2019.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

John Bohannon. Who's afraid of peer review? *Science*, 342(6154):60–65, 2013. doi:10.1126/science.2013.342.6154.342_60. URL https://www.science.org/doi/abs/10.1126/science.2013.342.6154.342_60.

Tilman Börgers. Pure strategy dominance. *Econometrica: Journal of the Econometric Society*, pages 423–430, 1993.

Tilman Börgers and Maarten CW Janssen. On the dominance solvability of large cournot games. *Games and Economic Behavior*, 8(2):297–321, 1995.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis. *Journal of the Association for Information Science and Technology*, 28, 2014.

Holly P Borowski, Jason R Marden, and Jeff S Shamma. Learning to play efficient coarse correlated equilibria. *Dynamic Games and Applications*, 9:24–46, 2019.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Adam Brandenburger and Eddie Dekel. Rationalizability and correlated equilibria. *Econometrica: Journal of the Econometric Society*, pages 1391–1402, 1987.

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, 2018.

Mark Braverman, Jieming Mao, Jon Schneider, and S Matthew Weinberg. Multi-armed bandit problems with strategic arms. In *Conference on Learning Theory*, pages 383–416. PMLR, 2019.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. In *Advances in Neural Information Processing Systems*, pages 5661–5671, 2018.

Elise S Brezis and Aliaksandr Birukou. Arbitrariness in the peer review process. *Scientometrics*, 123(1):393–411, 2020.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. arxiv. *arXiv preprint arXiv:1606.01540*, 10, 2016.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Donald J Brown and Caterina Calsamiglia. The nonparametric approach to applied welfare analysis. *Economic Theory*, 31(1):183–188, 2007.

George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1829–1836, 2019.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85, 2017.

Peter Burger, Soeradj Kanhai, Alexander Pleijter, and Suzan Verberne. The reach of commercially motivated junk news on facebook. *PloS one*, 14(8):e0220446, 2019.

Federico Cacciamani, Matteo Castiglioni, and Nicola Gatti. Online information acquisition: Hiring multiple agents. *arXiv preprint arXiv:2307.06210*, 2023.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33904–33919. Curran Associates, Inc., 2022.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games. *arXiv preprint arXiv:2303.02738*, 2023.

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. IEEE, 2020.

Colin F Camerer. *Behavioral game theory: Experiments in strategic interaction*. Princeton university press, 2011.

Lingjing Cao and Zhiqiang Zhang. Developing science and technology policies for high risk-high reward research. *Bulletin of Chinese Academy of Sciences (Chinese Version)*, 37(5): 661–673, 2022.

Jean Cardinal, Martine Labbé, Stefan Langerman, and Belén Palop. Pricing of geometric transportation networks. In *CCCG*, pages 92–96. Citeseer, 2005.

Hans Carlsson and Eric Van Damme. Global games and equilibrium selection. *Econometrica: Journal of the Econometric Society*, pages 989–1018, 1993.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16188–16198. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/ba5451d3c91a0f982f103cdbe249bc78-Paper.pdf.

Matteo Castiglioni, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020b.

Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Designing menus of contracts efficiently: The power of randomization. *arXiv preprint arXiv:2202.10966*, 2022.

Marco Castillo and Mikhail Freer. A general revealed preference test for quasi-linear preferences: Theory and experiments. *Available at SSRN 3397767*, 2019.

Jakub Cernỳ, Viliam Lisỳ, Branislav Bosanskỳ, and Bo An. Dinkelbach-type algorithm for computing quantal Stackelberg equilibrium. In *IJCAI*, pages 246–253, 2020.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1): 1–11, 2021.

Kani Chen, Inchi Hu, and Zhiliang Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155–1163, 1999.

Le Chen, Alan Mislove, and Christo Wilson. Peeking beneath the hood of uber. In *Proceedings of the 2015 internet measurement conference*, pages 495–508, 2015.

Siyu Chen, Jibang Wu, Yifan Wu, and Zhuoran Yang. Learning to incentivize information acquisition: Proper scoring rules meet principal-agent model. In *International Conference on Machine Learning*, pages 5194–5218. PMLR, 2023.

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

Xi Chen, Xiaotie Deng, and Shang-Hua Teng. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Yifang Chen, Simon Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. In *International Conference on Machine Learning*, pages 1561–1570. PMLR, 2021.

Yiling Chen, Haifeng Xu, and Shuran Zheng. Selling information through consulting. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2412–2431. SIAM, 2020.

Ashish Cherukuri, Bahman Gharesifard, and Jorge Cortes. Saddle-point dynamics: conditions for asymptotic stability of saddle points. *SIAM Journal on Control and Optimization*, 55(1):486–511, 2017.

Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pages 807–834. PMLR, 2019.

Yun Kuen Cheung and Georgios Piliouras. Chaos, extremism and optimism: Volume analysis of learning in games. *Advances in Neural Information Processing Systems*, 33:9039–9049, 2020.

Yun Kuen Cheung and Yixin Tao. Chaos of learning beyond zero-sum and coordination via game decompositions. *arXiv preprint arXiv:2008.00540*, 2020.

Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. The route to chaos in routing games: When is price of anarchy too optimistic? *Advances in Neural Information Processing Systems*, 33:766–777, 2020.

Thiparat Chotibut, Fryderyk Falniowski, Michał Misiurewicz, and Georgios Piliouras. Family of chaotic maps from game theory. *Dynamical Systems*, 36(1):48–63, 2021.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Hedging under uncertainty: regret minimization meets exponentially fast convergence. In *International Symposium on Algorithmic Game Theory*, pages 252–263. Springer, 2017a.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *Proceedings of the 31th International Conference on Neural Information Processing Systems*, 2017b.

Michele Conforti, Marco Di Summa, and Giacomo Zambelli. Minimally infeasible set-partitioning problems with balanced constraints. *Mathematics of Operations Research*, 32(3):497–507, 2007.

Vincent Conitzer. On stackelberg mixed strategies. *Synthese*, 193(3):689–703, 2016.

Vincent Conitzer and Dmytro Korzhyk. Commitment to correlated strategies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 632–637, 2011.

Vincent Conitzer and Caspar Oesterheld. Foundations of cooperative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15359–15367, 2023.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

Vincent Conitzer and Tuomas Sandholm. New complexity results about Nash equilibria. *Games and Economic Behavior*, 63(2):621–641, 2008.

Publishing Research Consortium et al. Publishing research consortium peer review survey 2015. london, england: Mark ware consulting, 2016.

Gerard Cornuejols, Marshall L Fisher, and George L Nemhauser. Exceptional paper—location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management science*, 23(8):789–810, 1977.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Antoine Augustin Cournot. *Recherches sur les principes mathématiques de la théorie des richesses*. L. Hachette, 1838.

Jacques Crémer and Richard P McLean. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53(2):345–361, 1985.

Giovanni Paolo Crespi, Davide Radi, and Matteo Rocca. Robust games: theory and application to a Cournot duopoly model. *Decisions in economics and finance*, 40(1-2):177–198, 2017.

Giovanni Paolo Crespi, Davide Radi, and Matteo Rocca. Insights on the theory of robust games. *arXiv preprint arXiv:2002.00225*, 2020.

Alex Csiszar. Peer review: Troubled from the start. *Nature*, 532(7599):306–308, 2016.

Rachel Cummings, Federico Echenique, and Adam Wierman. The empirical implications of privacy-aware choice. *Operations Research*, 64(1):67–78, 2016.

Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In *Conference on Learning Theory*, pages 287–316. PMLR, 2014.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 207–232. JMLR Workshop and Conference Proceedings, 2011.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Constantinos Daskalakis and Qinxuan Pan. A counter-example to karlin's strong conjecture for fictitious play. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 11–20. IEEE, 2014.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Constantinos Daskalakis, Rafael Frongillo, Christos H Papadimitriou, George Pierrakos, and Gregory Valiant. On learning algorithms for Nash equilibria. In *International Symposium on Algorithmic Game Theory*, pages 114–125. Springer, 2010.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34, 2021.

Eddie Dekel and Marciano Siniscalchi. Epistemic game theory. In *Handbook of Game Theory with Economic Applications*, volume 4, pages 619–702. Elsevier, 2015.

Eddie Dekel, Drew Fudenberg, and Stephen Morris. Interim correlated rationalizability. *Theoretical Economics*, 2007.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

Erik D Demaine and Martin L Demaine. Every author as first author. *arXiv preprint arXiv:2304.01393*, 2023.

Stephan Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

Arnoud V den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

Arnoud V den Boer and Bert Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management science*, 60(3):770–783, 2014.

Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. In *Advances in Neural Information Processing Systems*, pages 1577–1585, 2019.

Richard A Derrig. Insurance fraud. *Journal of Risk and Insurance*, 69(3):271–287, 2002.

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*, pages 983–1002. SIAM, 2014.

Douglas W Diamond and Philip H Dybvig. Bank runs, deposit insurance, and liquidity. *Journal of political economy*, 91(3):401–419, 1983.

Ilgin Dogan, Zuo-Jun Max Shen, and Anil Aswani. Estimating and incentivizing imperfect-knowledge agents with hidden rewards. *arXiv preprint arXiv:2308.06717*, 2023a.

Ilgin Dogan, Zuo-Jun Max Shen, and Anil Aswani. Repeated principal-agent games with unobserved agent rewards and perfect-knowledge agents. *arXiv preprint arXiv:2304.07407*, 2023b.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

Joseph L Doob. Topics in the theory of markoff chains. *Transactions of the American Mathematical Society*, 52(1):37–64, 1942.

Alexey Drutsa. Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In *Proceedings of the 26th International Conference on World Wide Web*, pages 33–42, 2017.

Alexey Drutsa. Weakly consistent optimal pricing algorithms in repeated posted-price auctions with strategic buyer. In *International Conference on Machine Learning*, pages 1319–1328, 2018.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Miroslav Dudík, Nika Haghtalab, Haipeng Luo, Robert E Schapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. *Journal of the ACM (JACM)*, 67(5):1–57, 2020.

Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models, 2023.

Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. *SIAM Journal on Computing*, (0):STOC16–68, 2019.

Jetlir Duraj and Kevin He. Dynamic information design with diminishing sensitivity over news. *arXiv preprint arXiv:1908.00084*, 2019.

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387, 2019.

Paul Dütting, Tim Roughgarden, and Inbal-Talgam Cohen. The complexity of contracts. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2688–2707. SIAM, 2020.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259, 2007.

Susan A Elmore and Eleanor H Weston. Predatory journals: what they are and how to avoid them. *Toxicologic pathology*, 48(4):607–610, 2020.

Jeffrey C Ely. Beeps. *American Economic Review*, 107(1):31–53, 2017.

Jeffrey C Ely and Marcin Pęski. Hierarchies of belief and interim rationalizability. *Theoretical Economics*, 1(1):19–65, 2006.

Kousha Etessami, Christos H Papadimitriou, Aviad Rubinstein, and Mihalis Yannakakis. Tarski's theorem, supermodular games, and the complexity of equilibria. In *11th Innovations in Theoretical Computer Science Conference*, 2020.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

James Fallows. Want to see how crazy a bot-run market can be?, Apr 2011. URL https://www.theatlantic.com/technology/archive/2011/04/want-to-see-how-crazy-a-bot-run-market-can-be/237773/.

Farzaneh Farhadi and Demosthenis Teneketzis. Dynamic information design: a simple problem on optimal sequential information disclosure. *Dynamic Games and Applications*, pages 1–42, 2021.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, 35:39076–39089, 2022.

J Farkas. Ober die theorie der einfachen ungleichungen. *J. Reine Angew. Math*, 124:1–24, 1902.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. Online advertisements with llms: Opportunities and challenges. *arXiv preprint arXiv:2311.07601*, 2023.

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 AAAI Conference on Artificial Intelligence*, volume 35, pages 5399–5406, 2021.

Andrew Ferdowsian, Muriel Niederle, and Leeat Yariv. Decentralized matching with aligned preferences. Technical report, Working paper, Department of Economics, Princeton University.[1225], 2020.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*, pages 101–109, 2019.

Miroslav Fiedler and Vlastimil Ptak. On matrices with non-positive off-diagonal elements and positive principal minors. *Czechoslovak Mathematical Journal*, 12(3):382–400, 1962.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. Political sectarianism in america. *Science*, 370(6516):533–536, 2020.

Richard Florida. The rise of the creator economy. 2022.

Dean Foster and Hobart Peyton Young. Regret testing: Learning to play Nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1(3):341–367, 2006.

Dean P Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. *Advances in Neural Information Processing Systems*, 29, 2016.

Eitan Frachtenberg and Noah Koster. A survey of accepted authors in computer systems conferences. *PeerJ Computer Science*, 6:e299, 2020.

Chiara Franzoni. Encouraging high-risk high-reward research at nih. 2023.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22, 2014.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Drew Fudenberg and Annie Liang. Predicting and understanding initial play. *American Economic Review*, 109(12):4112–41, 2019.

Drew Fudenberg and Alexander Peysakhovich. Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)*, 4(4):1–18, 2016.

David Gale. A theory of n-person games with perfect information. *Proceedings of the National Academy of Sciences of the United States of America*, 39(6):496, 1953.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Jiarui Gan, Qingyu Guo, Long Tran-Thanh, Bo An, and Michael Wooldridge. Manipulating a learning defender and ways to counteract. In *Advances in Neural Information Processing Systems*, pages 8272–8281, 2019a.

Jiarui Gan, Haifeng Xu, Qingyu Guo, Long Tran-Thanh, Zinovi Rabinovich, and Michael Wooldridge. Imitative follower deception in Stackelberg games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 639–657, 2019b.

Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. *arXiv preprint arXiv:2106.05137*, 2021.

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, 2023.

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Generalized principal-agency: Contracts, information, games and beyond, 2024.

Jose A García, Rosa Rodriguez-Sánchez, and Joaquín Fdez-Valdivia. The principal-agent problem in peer review. *Journal of the Association for Information Science and Technology*, 66(2):297–308, 2015.

Navid Ghaffarzadegan, Joshua Hawley, Richard Larson, and Yi Xue. A note on PhD population growth in biomedical sciences. *Systems research and behavioral science*, 32(3): 402–405, 2015.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2): 153–161, 2002.

Angeliki Giannou, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Conference on Learning Theory*, pages 2147–2148. PMLR, 2021.

Michael T Gibbons. Universities report largest growth in federally funded r&d expenditures since fy 2011. infobrief. NSF 23-303. *National Science Foundation*, 2022. URL https://ncses.nsf.gov/pubs/nsf23303.

Paul Ginsparg. Lessons from arxiv's 30 years of information sharing. *Nature Reviews Physics*, 3(9):602–603, 2021.

Marco Giordan, Attila Csikasz-Nagy, Andrew M Collings, and Federico Vaggi. The effects of an editor serving as one of the reviewers during the peer-review process. *F1000Research*, 5, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Ashish Goel and Adam Meyerson. Simultaneous optimization via approximate majorization for concave profits or convex costs. *Algorithmica*, 44:301–323, 2006.

Denizalp Goktas and Amy Greenwald. Convex-concave min-max Stackelberg games. *Advances in Neural Information Processing Systems*, 34:2991–3003, 2021.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *arXiv preprint arXiv:2311.09497*, 2023.

Itay Goldstein and Yaron Leitner. Stress tests and information disclosure. *Journal of Economic Theory*, 177:34–69, 2018.

Robert E Gropp, Scott Glisson, Stephen Gallo, and Lisa Thompson. Peer review: A system under stress. *BioScience*, 67(5):407–410, 2017.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics*, pages 302–340. Springer, 1992.

Harish Guda and Upender Subramanian. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*, 65(5):1995–2014, 2019.

Qingyu Guo, Jiarui Gan, Fei Fang, Long Tran-Thanh, Milind Tambe, and Bo An. On the inducibility of Stackelberg equilibrium for security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2020–2028, 2019.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

Guru Guruganesh, Jon Schneider, and Joshua R Wang. Contracts under moral hazard and adverse selection. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 563–582, 2021.

James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

Mark A Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. The strain on scientific publishing. *arXiv preprint arXiv:2309.15884*, 2023.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

John C Harsanyi. Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182, 1967.

John C Harsanyi, Reinhard Selten, et al. A general theory of equilibrium selection in games. *MIT Press Books*, 1, 1988.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Sergiu Hart and Andreu Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93(5):1830–1836, 2003.

Jason D Hartline. Mechanism design and approximation. *Book draft. October*, 122, 2013.

Kenji Hata, Ranjay Krishna, Li Fei-Fei, and Michael S Bernstein. A glimpse far into the future: Understanding long-term crowd worker quality. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 889–901, 2017.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *Advances in Neural Information Processing Systems*, pages 6369–6378, 2017.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, 55:317–359, 2016.

9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, 2021. ICLR, OpenReview.net. URL https://openreview.net/group?id=ICLR.cc/2021/Conference.

The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, 2022. ICLR, OpenReview.net. URL https://openreview.net/group?id=ICLR.cc/2022/Conference.

The 11th International Conference on Learning Representations, ICLR 2023, Kigali Rwanda, May 1-5, 2023, 2023. ICLR, OpenReview.net. URL https://openreview.net/group?id=ICLR.cc/2023/Conference.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Nicole Immorlica, Meena Jagadeesan, and Brendan Lucier. Clickbait vs. quality: How engagement-based optimization shapes the content landscape in online platforms. *arXiv preprint arXiv:2401.09804*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Amir Jafari, Amy Greenwald, David Gondek, and Gunes Ercal. On no-regret learning, fictitious play, and Nash equilibrium. In *ICML*, volume 1, pages 226–233, 2001.

Manish Jain, Fernando Ordónez, James Pita, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Robust solutions in Stackelberg games: Addressing boundedly rational human preference models. In *Proc. of the AAAI 4th Multidiciplinary Workshop on Advances in Preference Handling*, 2008.

Ole Jann and Christoph Schottmüller. Correlated equilibria in homogeneous good bertrand competition. *Journal of Mathematical Economics*, 57:31–37, 2015.

Steven Jecmen, Nihar B Shah, Fei Fang, and Vincent Conitzer. Tradeoffs in preventing manipulation in paper bidding for reviewer assignment. *arXiv preprint arXiv:2207.11315*, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Sham Kakade, Akshay Krishnamurphy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *arXiv:2006.12466*, 2020.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, October 2011a. doi:10.1257/aer.101.6.2590. URL https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, August 2023. ISSN 2666-3899. doi:10.1016/j.patter.2023.100804. URL https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9. Publisher: Elsevier.

Richard M Karp. *Reducibility among combinatorial problems*. Springer, 2010.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Ron S Kenett and Bernard G Francq. Helping reviewers assess statistical analysis: A case study from analytic methods. *Analytical Science Advances*, 3(5-6):212–222, 2022.

Ron S Kenett and Galit Shmueli. Helping authors and reviewers ask the right questions: The infoq framework for reviewing applied research. *Statistical Journal of the IAOS*, 32 (1):11–19, 2016.

Wolfgang E Kerzendorf, Ferdinando Patat, Dominic Bordelon, Glenn van de Ven, and Tyler A Pritchard. Distributed peer review enhanced with natural language processing and machine learning. *Nature Astronomy*, 4(7):711–717, 2020.

N Bora Keskin and Assaf Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations Research*, 62(5):1142–1167, 2014.

Evan D Kharasch, Michael J Avram, J David Clark, Andrew J Davidson, Timothy T Houle, Jerrold H Levy, Martin J London, Daniel I Sessler, and Laszlo Vutskits. Peer review matters: research quality and the public trust. *Anesthesiology*, 134(1):1–6, 2021.

Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordónez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 689–696, 2009.

Christopher Kiekintveld, Milind Tambe, and Janusz Marecki. Robust bayesian methods for Stackelberg security games. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 1467–1468. Citeseer, 2010.

Christopher Kiekintveld, Towhidul Islam, and Vladik Kreinovich. Security games with interval uncertainty. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, page 231–238, Richland, SC, 2013. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450319935.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

Jon Kleinberg and Robert Kleinberg. Delegated search approximates efficient search. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 287–302, 2018.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Nitish Korula, Vahab Mirrokni, and Hamid Nazerzadeh. Optimizing display advertising markets: Challenges and directions. *IEEE Internet Computing*, 20(1):28–35, 2015.

Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, 41:297–327, 2011.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

Vijay Krishna. *Auction theory*. Academic press, 2009.

Christian Kroer and Tuomas Sandholm. Imperfect-recall abstractions with bounds in games. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 459–476, 2016.

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Robust Stackelberg equilibria in extensive-form games and extension to limited lookahead. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Volodymyr Kuleshov and Okke Schrijvers. Inverse game theory: Learning utilities in succinct games. In *International Conference on Web and Internet Economics*, pages 413–427. Springer, 2015.

Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017.

Jean-Jacques Laffont and David Martimort. The theory of incentives. In *The Theory of Incentives*. Princeton university press, 2009.

Kate Lajtha and Philippe C Baveye. How should we deal with the growing peer-review problem? *Biogeochemistry*, 101:1–3, 2010.

Nicolas S Lambert. Elicitation and evaluation of statistical forecasts. *Preprint*, 2011.

John Langford. When the bubble bursts. . . ., 2018. URL https://hunch.net/?p=9604328.

John Langford and Mark Guzdial. The arbitrariness of reviews, and advice for school administrators. *Communications of the ACM*, 58(4):12–13, 2015.

Rida Laraki and Panayotis Mertikopoulos. Higher order game dynamics. *Journal of Economic Theory*, 148(6):2666–2695, 2013.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-tuning games: Bargaining and adaptation for general-purpose models, 2023.

Neil Lawrence and Corinna Cortes. The NIPS experiment, 2014.

Edward P Lazear and Sherwin Rosen. Rank-order tournaments as optimum labor contracts. *Journal of political Economy*, 89(5):841–864, 1981.

An Le Thi Hoai and Pham Dinh Tao. Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of global optimization*, 11(3):253–285, 1997.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and MDPs. *Advances in Neural Information Processing Systems*, 33, 2020.

SangMok Lee. The testable implications of zero-sum games. *Journal of Mathematical Economics*, 48(1):39–46, 2012.

Ehud Lehrer and Dimitry Shaiderman. Markovian persuasion. *arXiv preprint arXiv:2111.14365*, 2021.

Renato Paes Leme and Jon Schneider. Contextual search via intrinsic volumes. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 268–282. IEEE, 2018.

Joshua Letchford and Vincent Conitzer. Computing optimal strategies to commit to in extensive-form games. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 83–92, 2010.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International symposium on algorithmic game theory*, pages 250–262. Springer, 2009.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

Jun Li, Nelson Granados, and Serguei Netessine. Are consumers strategic? structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137, 2014.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017a.

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017b.

Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The world wide web conference*, pages 983–994, 2019a.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019b.

Enming Liang, Kexin Wen, William HK Lam, Agachai Sumalee, and Renxin Zhong. An integrated reinforcement learning and centralized programming approach for online taxi dispatching. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Raz Lin, Sarit Kraus, Jonathan Wilkenfeld, and James Barry. Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, 172(6-7):823–851, 2008.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*, pages 6161–6171. PMLR, 2020b.

Chun Kai Ling, Fei Fang, and J Zico Kolter. What game are we playing? end-to-end learning in normal and extensive form games. *arXiv preprint arXiv:1805.02777*, 2018.

Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. Playing large games using simple strategies. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 36–41, 2003.

Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77, 2019.

Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

Shuze Liu, Weiran Shen, and Haifeng Xu. Optimal pricing of information. *arXiv preprint arXiv:2102.13289*, 2021.

Ilan Lobel, Renato Paes Leme, and Adrian Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 66(5):1346–1361, 2018.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

Robert S MacKay, Ralph Kenna, Robert J Low, and Sarah Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society open science*, 4(2):160760, 2017.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

Yishay Mansour, Alex Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. *Operations Research*, 2021.

Jieming Mao, Renato Leme, and Jon Schneider. Contextual pricing for lipschitz buyers. *Advances in Neural Information Processing Systems*, 31, 2018.

Janusz Marecki, Gerry Tesauro, and Richard Segal. Playing repeated Stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 821–828, 2012.

Albert W Marshall, Ingram Olkin, and Barry C Arnold. Inequalities: theory of majorization and its applications. 1979.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

Eric Maskin. Borda's rule and arrow's independence of irrelevant alternatives. 2023.

Jiri Matousek and Bernd Gärtner. *Understanding and using linear programming*. Springer Science & Business Media, 2007.

Eric Mazumdar, Lillian J Ratliff, and S Shankar Sastry. On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1):103–131, 2020.

Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Alison McCook. Is peer review broken? submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. what's wrong with peer review? *The scientist*, 20(2):26–35, 2006.

Ryan C McDevitt. "A" business by any other name: firm name choice as a signal of firm quality. *Journal of Political Economy*, 122(4):909–944, 2014.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Brian McManus, Jessica Howell, and Michael Hurwitz. Strategic disclosure of test scores: Evidence from us college admissions. edworkingpaper no. 23-843. *Annenberg Institute for School Reform at Brown University*, 2023.

Hardik Meisheri, Vinita Baniwal, Nazneen N Sultana, Harshad Khadilkar, and Balaraman Ravindran. Using reinforcement learning for a large variable-dimensional inventory management problem. In *Adaptive Learning Agents Workshop at AAMAS*, 2020.

Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

Panayotis Mertikopoulos and William H Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018a.

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*, pages 2703–2717. SIAM, 2018b.

Jeffrey Mervis. Peering into peer review. *Science*, 343(6171):596–598, 2014. doi:10.1126/science.343.6171.596. URL https://www.science.org/doi/abs/10.1126/science.343.6171.596.

Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *Ai & Society*, 35(4):957–967, 2020.

Paul Milgrom and John Roberts. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, pages 1255–1277, 1990.

Mehryar Mohri and Andres Munoz. Optimal regret minimization in posted-price auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, pages 1871–1879, 2014.

Mehryar Mohri and Andres Munoz. Revenue optimization against strategic buyers. In *Advances in Neural Information Processing Systems*, pages 2530–2538, 2015.

Dov Monderer and Lloyd S Shapley. Fictitious play property for games with identical interests. *Journal of economic theory*, 68(1):258–265, 1996.

Hervé Moulin. Dominance solvable voting schemes. *Econometrica: Journal of the Econometric Society*, pages 1337–1351, 1979.

Adrian Mulligan, Louise Hall, and Ellen Raphael. Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, 64(1):132–161, 2013.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 681–699. Springer, 2020.

S. Muthukrishnan. Ad exchanges: Research issues. In Stefano Leonardi, editor, *Internet and Network Economics*, pages 1–12, 2009. ISBN 978-3-642-10841-9.

Vidya Muthukumar and Anant Sahai. Robust commitments and partial reputation. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 637–638, 2019.

Roger B Myerson. Refinements of the Nash equilibrium concept. *International journal of game theory*, 7(2):73–80, 1978.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

Roger B Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81, 1982.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in neural information processing systems*, pages 5585–5595, 2017.

John F Nash et al. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

Robert F. Nau, Sabrina Gomez Canovas, and Pierre Hansen. On the geometry of Nash equilibria and correlated equilibria. *International Journal of Game Theory*, 32:443–453, 2003. URL https://api.semanticscholar.org/CorpusID:7609307.

Thomas Nedelec, Marc Abeille, Clément Calauzènes, Noureddine El Karoui, Benjamin Heymann, and Vianney Perchet. The bidder's standpoint: a simple way to improve bidding strategies in revenue-maximizing auctions. *arXiv preprint arXiv:1808.06979*, 2018.

Thomas Nedelec, Jules Baudet, Vianney Perchet, and Noureddine El Karoui. Adversarial learning for revenue-maximizing auctions. *arXiv preprint arXiv:1909.06806*, 2019a.

Thomas Nedelec, Noureddine El Karoui, and Vianney Perchet. Learning to bid in revenue maximizing auction. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 934–935, 2019b.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1): 265–294, 1978.

Maresi Nerad, David Bogle, Ulrike Kohl, Conor O'Carroll, Christian Peters, and Beate Scholz. *Towards a global Core value system in doctoral education*. UCL Press, 2022.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances on Neural Information Processing Systems 28 (NIPS 2015)*, pages 3150–3158, 2015.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

Thanh Nguyen and Haifeng Xu. Imitative attacker deception in Stackelberg security games. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 528–534. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi:10.24963/ijcai.2019/75. URL https://doi.org/10.24963/ijcai.2019/75.

Thanh H Nguyen and Haifeng Xu. Imitative attacker deception in Stackelberg security games. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 528–534. AAAI Press, 2019b.

Thanh H Nguyen, Francesco M Delle Fave, Debarun Kar, Aravind S Lakshminarayanan, Amulya Yadav, Milind Tambe, Noa Agmon, Andrew J Plumptre, Margaret Driciru, Fred Wanyama, et al. Making the most of our regrets: Regret-based solutions to handle payoff uncertainty and elicitation in green security games. In *International Conference on Decision and Game Theory for Security*, pages 170–191. Springer, 2015.

Thanh Hong Nguyen, Amulya Yadav, Bo An, Milind Tambe, and Craig Boutilier. Regret-based optimization and preference elicitation for Stackelberg security games with uncertainty. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

David Nicholas, Anthony Watkinson, Hamid R Jamali, Eti Herman, Carol Tenopir, Rachel Volentine, Suzie Allard, and Kenneth Levine. Peer review: Still king in the digital age. *Learned Publishing*, 28(1):15–21, 2015.

Muriel Niederle and Leeat Yariv. Decentralized matching with aligned preferences. Technical report, National Bureau of Economic Research, 2009.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 129–140, 1999.

Syavash Nobarany, Kellogg S Booth, and Gary Hsieh. What motivates people to review articles? the case of the human-computer interaction community. *Journal of the Association for Information Science and Technology*, 67(6):1358–1371, 2016.

Francesco Orabona. Yet another icml award fiasco, 2023. URL https://parameterfree.com/2023/08/30/yet-another-icml-award-fiasco/.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pages 2701–2710. PMLR, 2017.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

W.R. Paczkowski. *Pricing analytics: Models and advanced quantitative techniques for product pricing*. 06 2018. doi:10.4324/9781315178349.

Gerasimos Palaiopanos, Ioannis Panageas, and Georgios Piliouras. Multiplicative weights update with constant step-size in congestion games: Convergence, limit cycles and chaos. *Advances in Neural Information Processing Systems*, 30, 2017.

Praveen Paruchuri, Jonathan P Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games for security: An efficient exact algorithm for solving bayesian Stackelberg games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 895–902, 2008.

David G Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, pages 1029–1050, 1984.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2149–2156, 2019.

Vianney Perchet. Finding robust Nash equilibria. In *Algorithmic Learning Theory*, pages 725–751. PMLR, 2020.

Motty Perry and Philip J Reny. How to count citations if you must. *American Economic Review*, 106(9):2722–2741, 2016.

James Pita, Manish Jain, Fernando Ordóñez, Milind Tambe, Sarit Kraus, and Reuma Magori-Cohen. Effective solutions for real-world Stackelberg games: When agents must deal with human uncertainties. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 369–376. Citeseer, 2009.

James Pita, Manish Jain, Milind Tambe, Fernando Ordónez, and Sarit Kraus. Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171, 2010.

Bary SR Pradelski and H Peyton Young. Learning efficient Nash equilibria in distributed systems. *Games and Economic behavior*, 75(2):882–897, 2012.

Simon Price and Peter A Flach. Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, 60(3):70–79, 2017.

Publons. Global state of peer review 2018., 2018. URL https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Zhiwei Qin, Xiaocheng Tang, Yan Jiao, Fan Zhang, Zhe Xu, Hongtu Zhu, and Jieping Ye. Ride-hailing order dispatching at didi via reinforcement learning. *INFORMS Journal on Applied Analytics*, 50(5):272–286, 2020.

Zhiwei Tony Qin, Hongtu Zhu, and Jieping Ye. Reinforcement learning for ridesharing: A survey. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2447–2454. IEEE, 2021.

Matthew Rabin. Incorporating fairness into game theory and economics. *The American economic review*, pages 1281–1302, 1993.

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*, pages 645–653. Citeseer, 2015.

Howard Raiffa and Robert Duncan Luce. *Games and Decisions: Introduction and Critical Survey*. John Wiley, New York, 1957.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3066–3074, 2013.

Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Secure-ucb: Saving stochastic bandits from poisoning attacks via limited data verification. *arXiv preprint arXiv:2102.07711*, 2021.

Charvi Rastogi, Ivan Stelmakh, Alina Beygelzimer, Yann N Dauphin, Percy Liang, Jennifer Wortman Vaughan, Zhenyu Xue, Hal Daumé III, Emma Pierson, and Nihar B Shah. How do authors' perceptions of their papers compare with co-authors' perceptions and peer-review decisions? *Plos one*, 19(4):e0300710, 2024.

Lillian J Ratliff, Samuel A Burden, and S Shankar Sastry. Characterization and computation of local Nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924. IEEE, 2013.

Lillian J Ratliff, Shreyas Sekar, Liyuan Zheng, and Tanner Fiez. Incentives in the dark: multi-armed bandits for evolving users with unknown type. *arXiv preprint arXiv:1803.04008*, 55, 2018.

Jérôme Renault, Eilon Solan, and Nicolas Vieille. Optimal dynamic information provision. *Games and Economic Behavior*, 104:329–349, 2017.

Drummond Rennie. Let's make peer review scientific. *Nature*, 535(7610):31–33, 2016.

R Tyrrell Rockafellar. Second-order convex analysis. *J. Nonlinear Convex Anal*, 1(1-16):84, 1999.

Ralph Rockafellar. Characterization of the subdifferentials of convex functions. *Pacific Journal of Mathematics*, 17(3):497–510, 1966.

Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pages 520–534, 1965.

Sherwin Rosen. Prizes and incentives in elimination tournaments. 1985.

Amy Ross Arguedas, Craig Robertson, Richard Fletcher, and Rasmus Nielsen. Echo chambers, filter bubbles, and polarisation: A literature review. 2022.

Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Watch and learn: Optimizing from revealed preferences feedback. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 949–962, 2016.

Aaron Roth, Aleksandrs Slivkins, Jonathan Ullman, and Zhiwei Steven Wu. Multidimensional dynamic pricing for welfare maximization. *ACM Transactions on Economics and Computation (TEAC)*, 8(1):1–35, 2020.

Tim Roughgarden and Inbal Talgam-Cohen. Why prices need algorithms. In *Proceedings of the sixteenth acm conference on economics and computation*, pages 19–36, 2015.

Ariel Rubinstein. *Lecture notes in microeconomic theory: the economic agent*. Princeton University Press, 2012.

Aviad Rubinstein. Settling the complexity of computing approximate two-player Nash equilibria. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265. IEEE Computer Society, 2016.

Alessio Russo. Some ethical issues in the review process of machine learning conferences. *arXiv preprint arXiv:2106.00810*, 2021.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

Eden Saig, Inbal Talgam-Cohen, and Nir Rosenfeld. Delegated classification. *Advances in Neural Information Processing Systems*, 36, 2024.

Paul A Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17): 61–71, 1938.

Cláudia S Sarrico. The expansion of doctoral education and the changing nature and purpose of the doctorate. *Higher Education*, 84(6):1299–1315, 2022.

Umran Sarwar and Marios Nicolaou. Fraud and deceit in medical research. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, 17(11):1077, 2012.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Antoine Scheid, Daniil Tiapkin, Etienne Boursier, Aymeric Capitaine, El Mahdi El Mhamdi, Éric Moulines, Michael I Jordan, and Alain Durmus. Incentivized learning in principal-agent bandit games. *arXiv preprint arXiv:2403.03811*, 2024.

Aaron Schlenker, Omkar Thakoor, Haifeng Xu, Fei Fang, Milind Tambe, Long Tran-Thanh, Phebe Vayanos, and Yevgeniy Vorobeychik. Deceiving cyber adversaries: A game theoretic approach. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 892–900. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

James Schummer and Rakesh V Vohra. Mechanism design without money. *Algorithmic game theory*, 10:243–299, 2007.

D Sculley, Jasper Snoek, and Alex Wiltschko. Avoiding a tragedy of the commons in the peer review process. *arXiv preprint arXiv:1901.06246*, 2018.

David Segal. Fake online locksmiths may be out to pick your pocket, too. *New York Times*, 1(30):19–21, 2016.

Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the nips 2016 review process. *Journal of machine learning research*, 19(49):1–34, 2018.

Virag Shah, Ramesh Johari, and Jose Blanchet. Semi-parametric dynamic contextual pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning. *arXiv preprint arXiv:2103.00360*, 2021.

Deeksha Sinha, Karthik Abinav Sankararaman, Abbas Kazerouni, and Vashist Avadhanula. Multi-armed bandits with cost subsidy. In *International Conference on Artificial Intelligence and Statistics*, pages 3016–3024. PMLR, 2021.

Petr Slavík. A tight analysis of the greedy algorithm for set cover. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 435–441, 1996.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Richard Smith. Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine*, 99(4):178–182, 2006.

Stephen A Smith. *Contract theory*. OUP Oxford, 2004.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Ray Spier. The history of the peer-review process. *TRENDS in Biotechnology*, 20(8):357–358, 2002.

Yves Sprumont. On the testable implications of collective choice theories. *Journal of Economic Theory*, 93(2):205–232, 2000.

Flaminio Squazzoni, Giangiacomo Bravo, and Károly Takács. Does incentive provision increase the quality of peer review? an experimental study. *Research Policy*, 42(1):287–294, 2013.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4785–4793, 2021.

Gilles Stoltz and Gábor Lugosi. Internal regret in on-line portfolio selection. *Machine Learning*, 59(1):125–159, 2005.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

Buxin Su, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Kyunghyun Cho, Jianqing Fan, Aaron Roth, and Weijie J Su. Analysis of the icml 2023 ranking data: Can authors' opinions of their own papers assist peer review in machine learning? *arXiv preprint arXiv:2408.13430*, 2024.

Weijie J Su. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems*, 34:27929–27939, 2021.

Weijie J Su. A truthful owner-assisted scoring mechanism. *arXiv preprint arXiv:2206.08149*, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28: 2989–2997, 2015.

Milind Tambe. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge university press, 2011.

Kok-Keong Tan, Jian Yu, and Xian-Zhi Yuan. Existence theorems of Nash equilibria for non-cooperative n-person games. *International Journal of Game Theory*, 24(3):217–222, 1995.

Sijun Tan, Jibang Wu, Xiaohui Bei, and Haifeng Xu. Least square calibration for peer reviews. *Advances in Neural Information Processing Systems*, 34:27069–27080, 2021.

Pingzhong Tang and Yulong Zeng. The price of prior dependence in auctions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 485–502, 2018.

Tiffany Ya Tang and Pinata Winoto. I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia*, 22 (1-2):111–138, 2016.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

Pham Dinh Tao and Le Thi Hoai An. A dc optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Pham Dinh Tao et al. The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems. *Annals of operations research*, 133(1-4):23–46, 2005.

Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific journal of Mathematics*, 5(2):285–309, 1955.

Steve H Tijs. Nash equilibria for noncooperative n-person games in normal form. *Siam Review*, 23(2):225–237, 1981.

Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

Donald M Topkis. Equilibrium points in nonzero-sum n-person submodular games. *Siam Journal on control and optimization*, 17(6):773–787, 1979.

Donald M Topkis. *Supermodularity and complementarity*. Princeton university press, 1998.

Long Tran-Thanh, Archie Chapman, Enrique Munoz De Cote, Alex Rogers, and Nicholas R Jennings. Epsilon–first policies for budget–limited multi-armed bandits. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1134–1140, 2012.

Ralph Turvey. Marginal cost. *The Economic Journal*, 79(314):282–299, 1969.

Amos Tversky and Derek J Koehler. Support theory: a nonextensional representation of subjective probability. *Psychological review*, 101(4):547, 1994.

Takashi Ui. Correlated equilibrium and concave games. *International Journal of Game Theory*, 37(1):1–13, 2008.

Unknown. Stability analysis of a 2-d dynamical system. https://neurophysics.ucsd.edu/courses/physics_171/2d_dyn_sys_notes.pdf.

J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

Ngo Van Long and Gerhard Sorger. A dynamic principal-agent problem as a feedback Stackelberg differential game. *Central European Journal of Operations Research*, 18(4): 491–509, 2010.

Richard Van Noorden. More than 10,000 research papers were retracted in 2023—a new record. *Nature*, 624(7992):479–481, 2023.

Arsenii Vanunts and Alexey Drutsa. Optimal pricing in repeated posted-price auctions with different patience of the seller and the buyer. In *Advances in Neural Information Processing Systems*, pages 939–951, 2019.

Hal R Varian. Revealed preference. *Samuelsonian economics and the twenty-first century*, pages 99–115, 2006.

Thorstein Veblen. *The theory of the leisure class*. Routledge, 2017.

Theo MM Verhallen. Scarcity and consumer choice behavior. *Journal of Economic Psychology*, 2(4):299–322, 1982.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Yannick Viossat. The replicator dynamics does not lead to correlated equilibria. *Games and Economic Behavior*, 59(2):397–407, 2007.

Yannick Viossat. Is having a unique equilibrium robust? *Journal of Mathematical Economics*, 44(11):1152–1160, 2008.

Yannick Viossat. Evolutionary dynamics and dominated strategies. *Economic Theory Bulletin*, 3(1):91–113, 2015.

Yannick Viossat and Andriy Zapechelnyuk. No-regret dynamics and fictitious play. *Journal of Economic Theory*, 148(2):825–842, 2013.

Caitlyn Vlasschaert, Joel M Topf, and Swapnil Hiremath. Proliferation of papers and preprints during the coronavirus disease 2019 pandemic: progress or problems with peer review? *Advances in chronic kidney disease*, 27(5):418–426, 2020.

Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. *Advances in Neural Information Processing Systems*, 33:1380–1391, 2020.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

Heinrich Von Stackelberg. *Market structure and equilibrium*. Springer Science & Business Media, 2010.

Bernhard Von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.

Caroline S Wagner and Jeffrey Alexander. Evaluating transformative research programmes: A case study of the nsf small grants for exploratory research programme. *Research Evaluation*, 22(3):187–197, 2013.

Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, 46(2):265–280, 1945. ISSN 0003486X.

Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B Shah. Debiasing evaluations that are biased by evaluations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10120–10128, 2021a.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34, 2021b.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yuanhao Wang, Dingwen Kong, Yu Bai, and Chi Jin. Learning rationalizable equilibria in multiplayer games. In *The Eleventh International Conference on Learning Representations*, 2023.

Zizhuo Wang, Shiming Deng, and Yinyu Ye. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2): 318–331, 2014.

M. Ware and Publishing Research Consortium. *Peer Review: Benefits, Perceptions and Alternatives*. PRC summary papers. Publishing Research Consortium, 2008.

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*, 2020.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Karen White. Publications output: US trends and international comparisons. science & engineering indicators 2020. nsb-2020-6. *National Science Foundation*, 2019.

David P. Williamson. *Network Flow Algorithms*. Cambridge University Press, 2019a. doi:10.1017/9781316888568.

D.P. Williamson. *Network Flow Algorithms*. Cambridge University Press, 2019b. ISBN 9781107185890. URL https://books.google.com/books?id=TKGYwgEACAAJ.

James R Wright and Kevin Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

Jibang Wu, Ashwinkumar Badanidiyuru, and Haifeng Xu. Auctioning with strategically reticent bidders. *arXiv preprint arXiv:2109.04888*, 2021.

Jibang Wu, Weiran Shen, Fei Fang, and Haifeng Xu. Inverse game theory for stackelberg games: the blessing of bounded rationality. *Advances in Neural Information Processing Systems*, 35:32186–32198, 2022a.

Jibang Wu, Haifeng Xu, and Fan Yao. Multi-agent learning for iterative dominance elimination: Formal barriers and new algorithms. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 543–543. PMLR, 02–05 Jul 2022b. URL https://proceedings.mlr.press/v178/wu22a.html.

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*, EC '22, page 471–472, New York, NY, USA, 2022c. Association for Computing Machinery. ISBN 9781450391504. doi:10.1145/3490486.3538313. URL https://doi.org/10.1145/3490486.3538313.

Jibang Wu, Haifeng Xu, Yifan Guo, and Weijie Su. An isotonic mechanism for overlapping ownership. *arXiv preprint arXiv:2306.11154*, 2023.

Jibang Wu, Siyu Chen, Mengdi Wang, Huazheng Wang, and Haifeng Xu. Contractual reinforcement learning: Pulling arms with invisible hands. *arXiv preprint arXiv:2407.01458*, 2024.

Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. Budgeted multi-armed bandits with multiple plays. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2210–2216, 2016.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010.

Shenke Xiao, Zihe Wang, Mengjing Chen, Pingzhong Tang, and Xiwang Yang. Optimal common contract with heterogeneous agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7309–7316, 2020.

Wenyi Xiao, Huan Zhao, Haojie Pan, Yangqiu Song, Vincent W Zheng, and Qiang Yang. Beyond personalization: Social content recommendation for creator equality and consumer satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 235–245, 2019.

Yuanzhang Xiao, Florian Dörfler, and Mihaela Van Der Schaar. Incentive design in peer review: Rating and repeated endogenous matching. *IEEE Transactions on Network Science and Engineering*, 6(4):898–908, 2018.

Haifeng Xu, Zinovi Rabinovich, Shaddin Dughmi, and Milind Tambe. Exploring information asymmetry in two-stage security games. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Yuling Yan, Weijie J Su, and Jianqing Fan. The isotonic mechanism for exponential family estimation. *arXiv preprint arXiv:2304.11160*, 2023.

Cenying Yang, Yihao Feng, and Andrew Whinston. Dynamic pricing and information disclosure for fresh produce: An artificial intelligence approach. *Production and Operations Management*, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Pu Yang, Krishnamurthy Iyer, and Peter Frazier. Information design in spatial resource competition. *arXiv preprint arXiv:1909.12723*, 2019.

Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *AAMAS*, pages 847–854, 2012.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: optimism in the face of large state spaces. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 13903–13916, 2020.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. How bad is top-k recommendation under competing content creators? In *International Conference on Machine Learning*, pages 39674–39701. PMLR, 2023.

Fan Yao, Chuanhao Li, Karthik Abinav Sankararaman, Yiming Liao, Yan Zhu, Qifan Wang, Hongning Wang, and Haifeng Xu. Rethinking incentives in recommender systems: Are monotone rewards always beneficial? *Advances in Neural Information Processing Systems*, 36, 2024.

Yue Yin, Bo An, Yevgeniy Vorobeychik, and Jun Zhuang. Optimal deceptive strategies in security games: A preliminary study. In *Proc. of AAAI*, 2013.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022.

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

Morteza Zadimoghaddam and Aaron Roth. Efficiently learning from revealed preference. In *International Workshop on Internet and Network Economics*, pages 114–127. Springer, 2012.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Andriy Zapechelnyuk et al. Limit behavior of no-regret dynamics. *Discussion Papers*, 21, 2009.

Cun-Hui Zhang. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528 – 555, 2002. doi:10.1214/aos/1021379864. URL https://doi.org/10.1214/aos/1021379864.

Yichi Zhang, Grant Schoenebeck, and Weijie Su. Eliciting honest information from authors using sequential review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9977–9984, 2024.

Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael Jordan. Online learning in stackelberg games with an omniscient follower. In *International Conference on Machine Learning*, pages 42304–42316. PMLR, 2023.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The ai economist: Improving equality and productivity with ai-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*, 2020.

Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I Jordan. The sample complexity of online contract design. *arXiv preprint arXiv:2211.05732*, 2022.

Banghua Zhu, Sai Praneeth Karimireddy, Jiantao Jiao, and Michael I Jordan. Online learning in a creator economy. *arXiv preprint arXiv:2305.11381*, 2023.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. *Advances in neural information processing systems*, 20:1729–1736, 2007.

Justin Zobel. When measurement misleads: The limits of batch assessment of retrieval systems. In *ACM SIGIR Forum*, volume 56, pages 1–20. ACM New York, NY, USA, 2023.

You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. EC '21, page 927–928, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385541. doi:10.1145/3465456.3467593. URL https://doi.org/10.1145/3465456.3467593.

Shiliang Zuo. New perspectives in online contract design: Heterogeneous, homogeneous, non-myopic agents and team production. *arXiv preprint arXiv:2403.07143*, 2024.

Appendices

APPENDIX A

A.1 Potential Applications of MPPs

In the main body of this paper, we use the ride-sharing platform as the example to motivate our model. In this section, we give more context on the application scenarios of MPPs and their potential impacts.

Recommendation for digital ads The Ad platforms nowadays collect massive Internet user data from their services such as search engines or social media. While the advertiser looks for the Ad slots of highest return, the platform instead optimizes for the total revenue (and social welfare). By strategically revealing certain information on the user demographics and preferences, the platform could influence the advertisers' value and consequently their offers on each Ad slot. Furthermore, as the availability changes after some slots get brought, the platform needs to design policies for its long-term objective.

Recommendation for online shopping Online shopping platforms are making use of learning tools such as reinforcement learning to manage inventory and ensure profitability [Giannoccaro and Pontrandolfo, 2002, Meisheri et al., 2020]. The platform cannot single-handedly manage its inventory, and information design plays an important role in its interactions with its suppliers and consumers. On the supply side, it could strategically reveal aspects of consumer sentiment (e.g., rough number of visits, search) to the suppliers in order to guide their sales expectation and negotiate for lower unit prices. On the demand side, it could tactically control displayed product information (e.g., last five remaining, editor's choice) so as to influence consumers' perception of products and consequently their purchase decisions.

Recommendation for content sharing A content sharing platform also needs to reconcile its misaligned interest with its users. On the one hand, it should recommend the most relevant items to its users for click-through and engagement. On the other hand, its recommendations are subject to misalignments with long-term objectives such as profits (e.g., from paid promotion), social impact (e.g., to prevent misinformation and filter bubbles) or development of a creator ecosystem [Tang and Winoto, 2016, Xiao et al., 2019, Milano et al., 2020].

A.2 Omitted Proofs and Descriptions

A.2.1 Formal Description of the `OP4`

The full description of the `OP4` for general MPPs is stated as follows:

ALGORITHM A.1: The Full Description of OP4 for MPPs

Input: Number of Episodes T, Number of Step H

Parameters: $\beta > 0$, $\rho > 0$, $\lambda \in \mathbb{R}^+$.

Output: $a_h^t \in \mathcal{A}$ for each $h \in [H], t \in [T]$ **for** *episode* $t = 1 \dots T$ **do**

Receive the initial state s_1^t and context $C^t = (c_1^t, \dots, c_H^t)$.

for *step* $h = H, \dots, 1$ **do**

Compute the constrained least square problem

$$\theta_h^t \leftarrow \underset{\|\theta_h\| \le L_\theta}{\operatorname{argmin}} \sum_{\tau \in [t-1]} \left[\omega_h^\tau - f(\phi(c_h^\tau)^\top \theta_h)\right]^2.$$

Calculate $\Sigma_h^t = \Phi^2 I_{d_\phi} + \sum_{\tau \in [t-1]} \phi(c_h^\tau)\phi(c_h^\tau)^\top$. Update $\mathcal{B}_h^t \leftarrow \mathrm{B}_{\Sigma_h^t}(\theta_h^t, \beta)$.

Set $\mu_h^t(\cdot|c)$ to the distribution of $f\big(\phi(c)^\top \theta_h^t\big) + z_h$.

Calculate $\begin{cases} \Gamma_h^t &= \lambda I_{d_\psi} + \sum_{\tau \in [t-1]} \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)\psi(s_h^\tau, \omega_h^\tau, a_h^\tau)^\top, \\ \iota_h^t &= \sum_{\tau \in [t-1]} \psi(s_h^\tau, \omega_h^\tau, a_h^\tau)[v_h^\tau + V_{h+1}^t(s_{h+1}^\tau; C^t)] \end{cases}$

Update $q_h^t \leftarrow (\Gamma_h^t)^{-1} \iota_h^t$.

Set $\begin{cases} Q_h^t(\cdot, \cdot, \cdot; C^t) \leftarrow \min\{\psi(\cdot, \cdot, \cdot)^\top q_h^t + \rho \|\psi(\cdot, \cdot, \cdot)\|_{(\Gamma_h^t)^{-1}}, H\}, \\ V_h^t(\cdot; C^t) \leftarrow \max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(\cdot; C^t). \end{cases}$

for *step* $h = 1, \dots, H$ **do**

Choose $\pi_h^t \in \arg\max_{\pi_h \in \mathrm{Pers}(\mu_{\mathcal{B}_h^t}, u_h)} \left\langle Q_h^t, \mu_h^t \otimes \pi_h \right\rangle_{\Omega \times \mathcal{A}}(s_h^t; C^t)$.

Execute π^t to sample a trajectory $\{(s_h^t, \omega_h^t, a_h^t, v_h^t)\}_{h \in [H]}$.

A.2.2 Proof of Theorem 2.3

In order to prove the sublinear regret for OP4, we construct a novel regret decomposition tailored to MPPs. Our proof starts from decomposing the regret into several terms, each of which indicates the regret loss either from estimation or from the randomness of trajectories. Next, we evaluate each term and then add them together to conclude the upper bound of

the regret of OP4. For simplicity of presentation, denote $\widetilde{V}_h^t(\cdot; C) = \left\langle Q_h^t, \mu_h^* \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(\cdot; C)$ as the expectation of Q_h^t with respect to the ground-truth prior μ_h^* and signaling scheme π_h^t at the h-th step. Then we can define the temporal-difference (TD) error as

$$\delta_h^t(s, \omega, a) = (v_h^t + P_h V_{h+1}^t - Q_h^t)(s, \omega, a; C^t). \tag{A.1}$$

Here δ_h^t is a function on $\mathcal{S} \times \Omega \times \mathcal{A}$ for all $h \in [H]$ and $t \in [T]$. Intuitively, $\{\delta_h^t\}_{h \in [H]}$ quantifies how far the Q-functions $\{Q_h^t\}_{h \in [H]}$ are from satisfying the Bellman optimality equation in equation (2.2). Moreover, define $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ for the trajectory $\{c_h^t, s_h^t, \omega_h^t, a_h^t\}_{h \in [H]}$ generated by Algorithm A.1 at the t-th episode as follows

$$\begin{aligned} \zeta_{t,h}^1 &= (\widetilde{V}_h^t - V_h^{\pi^t})(s_h^t; C^t) - (Q_h^t - Q_h^{\pi^t})(s_h^t, \omega_h^t, a_h^t; C^t), \\ \zeta_{t,h}^2 &= P_h(V_{h+1}^t - V_{h+1}^*)(s_h^t, \omega_h^t, a_h^t; C^t) - (V_{h+1}^t - V_{h+1}^*)(s_{h+1}^t; C^t). \end{aligned} \tag{A.2}$$

By definition, $\zeta_{t,h}^1$ capture the randomness of realizing the outcome $\omega_h^t \sim \mu_h^*(\cdot | c_h)$ and signaling the action $a_h^t \sim \pi_h^t(s_h^t, \omega_h^t, \cdot)$, while $\zeta_{t,h}^2$ captures the randomness of drawing the next state s_{h+1}^t from $P_h(\cdot | s_h^t, \omega_h^t, \cdot)$. With the notations above, we can decompose the regret into six parts to facilitate the establishment of the upper bound of the regret.

Lemma A.1 (Regret Decomposition)**.** *With the notations defines in equation* (A.1) *and*

(A.2), *we can write the regret as:*

$$\begin{aligned}\mathrm{Reg}(T,\mu^*) = &\underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\left\{\mathbb{E}_{\mu_h^*,\pi_h^*}[\delta_h^t(s_h,\omega_h,a_h^t)|s_1=s_1^t]-\delta_h^t(s_h^t,\omega_h^t,a_h^t)\right\}}_{\text{(i)}}+\underbrace{\sum_{t\in[T]}\sum_{h\in[H]}(\zeta_{t,h}^1+\zeta_{t,h}^2)}_{\text{(ii)}}\\ &+\underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\mathbb{E}_{\mu_h^*,\pi_h^*}\left[\left\langle Q_h^t,\mu_h^*\otimes\pi_h^*-\mu_h^t\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h;C^t)|s_1=s_1^t\right]}_{\text{(iii)}}\\ &+\underbrace{\sum_{t\in[T]}\sum_{h\in[H]}\left\langle Q_h^t,(\mu_h^t-\mu_h^*)\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h^t,C^t)}_{\text{(iv)}}.\end{aligned} \tag{A.3}$$

In this regret decomposition, term (i) indicates the optimism in `OP4`. Term (iii) corresponds to the pessimism in `OP4` for inducing a robust equilibria. The expectation $\mathbb{E}_{\mu_h^*,\pi_h^*}$ is taken for the random states realized in the trajectory generated under the ground-truth prior μ_h^* and optimal signaling policy π_h^*. While such expectation is difficult to evaluate, we are able to show δ_h^t in term (i) is always non-positive due to the optimistic Q-value estimation, and $\left\langle Q_h^t,\mu_h^*\otimes\pi_h^*-\mu_h^t\otimes\pi_h^t\right\rangle_{\Omega\times\mathcal{A}}(s_h;C^t)$ in term (iii) is independent of the realized state. These facts largely simplifies our analysis and are the key rationale behind this specific regret decomposition.

Unlike the regret decomposition in [Yang et al., 2020], Lemma A.1 also captures the randomness of realizing the outcome. We also add the term (iii) and (iv) to evaluate the further regret loss, since we have to estimate the prior of the outcome and choose a robustly persuasive policy in MPPs. The rigorous arguments turn out to be technical, and thus we shall defer the proof of most lemmas to the later part of the Appendix. We start by presenting the bound of each terms.

The term (i) in equation (A.3) can be bounded in regardless of the trajectories under

prior μ^* and signaling policy π^*, as is stated in Lemma A.2 below.

Lemma A.2 (Optimism). *There exists an absolute constant $c > 0$ such that, for any fixed $\delta \in (0,1)$, if we set $\lambda = \max\{1, \Psi^2\}$ and $\rho = cd_\psi H\sqrt{\iota}$ in Algorithm A.1 with $\iota = \log(2d_\psi\Psi^2 T/\delta)$, then with probability at least $1-\delta/2$,*

$$-2\rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}} \leq \delta_h^t(s,\omega,a) \leq 0.$$

for all $s \in \mathcal{S}, \omega \in \Omega, a \in \mathcal{A}, h \in [H]$ and $t \in [T]$.

The term (ii) in equation (A.3) can be bounded by Lemma 5.3 from [Yang et al., 2020] using martingale techniques and the Azuma-Hoeffding inequality [Azuma, 1967]. We state the upper bound for term (ii) in Lemma A.8.

The term (iii) in equation (A.3) evaluates the regret loss caused by estimating the prior and choosing a robustly persuasive signaling policy. It mostly rely on the robustness gap $\mathrm{Gap}(\cdot)$ introduced in the Appendix A.3.

Lemma A.3 (Bounding Term (iii)). *On the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, under Assumption 2.6 and 2.7,*

$$\sum_{t\in[T]}\sum_{h\in[H]} \mathbb{E}_{\mu_h^*,\pi_h^*}\big[\langle Q_h^t, \mu_h^*\otimes\pi_h^* - \mu_h^t\otimes\pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h;C^t)|s_1 = s_1^t\big] \leq \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \sum_{h\in[H]}\sum_{t\in[T]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

It remains to bound term (iv) in equation (A.3). This bound can be derived from Holder inequality and the property of the prior.

Lemma A.4 (Bounding Term (iv)). *On the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, under Assumption 2.6 and*

2.7,

$$\sum_{t\in[T]}\sum_{h\in[H]}\langle Q_h^t,(\mu_h^t-\mu_h^*)\otimes\pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h^t;C^t)\le HL_\mu K\beta\sum_{h\in[H]}\sum_{t\in[T]}\|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Now we are ready to prove Theorem 2.3. Given the regret bound of each term above, we pick $\beta = C(1+\kappa^{-1}\sqrt{K+M+d_\phi\sigma^2\log(HT)})$ and obtains the following upper bound for regret:

$$\begin{aligned}\mathrm{Reg}(T,\mu^*)\le&4\sqrt{2TH^3\log(2HT)}\\&+\sum_{t\in[T]}\sum_{h\in[H]}\left[2\rho\|\psi(s_h^t,\omega_h^t,a_h^t)\|_{(\Gamma_h^t)^{-1}}+\left(\frac{3HL_\mu K}{p_0D}+\frac{3HL_\mu K}{2}\right)\beta\|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}\right],\end{aligned}$$

With the probability of the given event by Lemma A.6 and appropriately chosen δ in previous lemmas, the above inequality holds for the probability at least $1-3H^{-1}T^{-1}$.

By Lemma A.9, we have

$$\begin{aligned}\mathrm{Reg}(T,\mu^*)\le&2\sqrt{2TH^3\log(2HT)}+2\rho H\sqrt{2d_\psi T\log\left(1+T\Psi^2/(\lambda d_\psi)\right)}\\&+\beta H^2L_\mu K\left(\frac{3}{p_0D}+\frac{3}{2}\right)\sqrt{2d_\phi T\log\left(1+T/(d_\phi)\right)}.\end{aligned}$$

Since β is in $\widetilde{O}(\sqrt{d_\phi})$ and ρ is in $\widetilde{O}(d_\psi H)$, we can conclude that the regret of Algorithm A.1 is $\widetilde{O}\big(d_\phi d_\psi^{3/2}H^2\sqrt{T}/(p_0D)\big)$.

In MPPs, we have to estimate the prior of the outcome since we cannot observe the ground-truth prior. However, the estimation may not satisfy the regularity conditions, which conflicts with the requirements for the prior when proving Lemma A.12. To address this problem, we give another upper bound of the robustness gap for the prior estimation in Lemma A.14. In addition, to handle the regret loss incurred by estimating the prior, we compute the difference in Q-functions when choosing respectively persuasive scheme for different priors in Lemma A.15.

We now prove that the above pessimism design guarantees persuasiveness w.r.t. the true prior with high probability. And it suffices to show that the estimation θ_h^t is close enough to the real parameter θ_h^* such that the confidence region $\mathcal{B}_h^t$ centered at θ_h^t given in Algorithm A.1 contains θ_h^*. If so, the signaling scheme chosen to be persuasive for the whole set $\mu_{\mathcal{B}_h^t}$ is also persuasive for μ_h^*, where $\mu_{\mathcal{B}} \coloneqq \{\mu_{\theta'} : \theta' \in \mathcal{B}\}$ denotes the set of priors that are determined by the parameters $\theta' \in \mathcal{B}$.

Lemma A.5. *There exists a constant $C > 0$ such that for $\beta = C(1+\kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(HT)})$, OP4 is persuasive with probability at least $1 - H^{-1}T^{-1}$, i.e.,*

$$\mathbb{P}_{\theta^*}\Big(\bigcup_{h\in[H]} \{\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t\}\Big) \le H^{-1}T^{-1}.$$

Proof. Proof. We first analyze the probability for being non-persuasive. For any $\|\theta_h^*\| \le L_\theta$, using the union bound, we have

$$\begin{aligned} P_{\theta^*}\Big(\bigcup_{t\in[T],h\in[H]} \{\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t\}\Big) &\le \sum_{t\in[T]} \sum_{h\in[H]} P_{\theta_h^*}(\theta_h^* \notin \cap_{t\in[T]} \mathcal{B}_h^t) \\ &\le \sum_{t\in[T]} \sum_{h\in[H]} P_{\theta_h^*}\big(\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} > \beta\big). \end{aligned}$$

The following lemma gives the belief of confidence region for the linear parameter θ_h^*. The proof can be directly derived from Lemma 6 in Wang et al. [2019].

Lemma A.6 (Belief of Confidence Region)**.** *For any $t \in [T]$ and $h \in [H]$, there exists a constant $C > 0$, such that for $\beta = C(1 + \kappa^{-1}\sqrt{K + M + d_\phi \sigma^2 \log(1/\delta)})$, given $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have $\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} \le \beta$.*

By Lemma A.6, taking $\delta = H^{-2}T^{-2}$, then we have $\mathbb{P}_{\theta^*}(\|\theta_h^t - \theta_h^*\|_{\Sigma_h^t} > \beta) \le H^{-2}T^{-2}$. Summing up the failure probabilities over $t \in [T]$, we have $\mathbb{P}_{\theta^*}(\theta^* \notin \cap_{t\in[T]} \mathcal{B}^t) \le H^{-1}T^{-1}$.

□

A.2.3 Proof of Lemma A.1 – Regret Decomposition

Proof. Proof. Before presenting the proof, we first define two operators $\mathbb{J}_h^*$ and $\mathbb{J}_h^t$:

$$(\mathbb{J}_h^* f)(s;C) = \langle f, \mu_h^* \otimes \pi_h^* \rangle_{\Omega\times\mathcal{A}}(s;C), \quad (\mathbb{J}_h^t f)(s;C) = \langle f, \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s;C), \tag{A.4}$$

for any $h \in [H], t \in [T]$ and any function $f(\cdot,\cdot,\cdot;C) : \mathcal{S}\times\Omega\times\mathcal{A} \to \mathbb{R}$ under the context C. Moreover, for any $h \in [H], t \in [T]$ and any state $s \in \mathcal{S}$, we define

$$\xi_h^t(s;C) = (\mathbb{J}_h^* Q_h^t)(s;C) - (\mathbb{J}_h^t Q_h^t)(s;C) = \langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s;C). \tag{A.5}$$

After introducing these notations, we decompose the instantaneous regret at the t-th episode into two terms,

$$V_1^*(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t) = \underbrace{V_1^*(s_1^t;C^t) - V_1^t(s_1^t;C^t)}_{\mathbf{p}_1} + \underbrace{V_1^t(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t)}_{\mathbf{p}_2}. \tag{A.6}$$

Then we consider these two terms separately. By the definition of value functions in (2.1) and the operator $\mathbb{J}_h^*$ in (A.4), we have $V_h^* = \mathbb{J}_h^* Q_h^*$. By the construction of Algorithm A.1, we have $V_h^t = \mathbb{J}_h^t Q_h^t$ similarly. Thus, for the first term $\mathbf{p}_1$ defined in equation (A.6), using ξ_h^t defined in (A.5), for any $h \in [H], t \in [T]$, we have

$$\begin{aligned} V_h^* - V_h^t &= \mathbb{J}_h^* Q_h^* - \mathbb{J}_h^t Q_h^t = (\mathbb{J}_h^* Q_h^* - \mathbb{J}_h^* Q_h^t) + (\mathbb{J}_h^* Q_h^t - \mathbb{J}_h^t Q_h^t) \\ &= \mathbb{J}_h^*(Q_h^* - Q_h^t) + \xi_h^t. \end{aligned}$$

Next, by the definition of the temporal-difference error δ_h^t in (A.1) and the Bellman optimality equation in equation (2.2), we have

$$Q_h^* - Q_h^t = (v_h + P_h V_{h+1}^*) - (v_h + P_h V_{h+1}^t - \delta_h^t) = P_h(V_{h+1}^* - V_{h+1}^t) + \delta_h^t.$$

Hence we get

$$V_h^* - V_h^t = \mathbb{J}_h^* P_h (V_{h+1}^* - V_{h+1}^t) + \mathbb{J}_h^* \delta_h^t + \xi_h^t.$$

Then, by recursively applying the above formula, we have

$$V_1^* - V_1^t = \Big(\prod_{h\in[H]} \mathbb{J}_h^* P_h \Big)(V_{H+1}^* - V_{H+1}^t) + \sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i \Big) \mathbb{J}_h^* \delta_h^t + \sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i \Big) \xi_h^t.$$

By the definition of ξ_h^t in equation (A.5), we get

$$\sum_{h\in[H]} \Big(\prod_{i\in[h]} \mathbb{J}_i^* P_i \Big) \xi_h^t(s_h; C^t) = \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*} \big\{ \big[\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) | s_1 = s_1^t \big] \big\}.$$

Notice that $V_{H+1}^* = V_{H+1}^t = 0$. Therefore, for any episode $t \in [T]$, we have

$$\begin{aligned} V_1^*(s_1^t; C^t) - V_1^t(s_1^t; C^t) = & \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*} \big\{ \big[\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) | s_1 = s_1^t \big] \big\} \\ & + \sum_{h\in[H]} \mathbb{E}_{\mu^*,\pi^*} \Big[\delta_h^t(s_h, \omega_h, a_h) | s_1 = s_1^t \Big]. \end{aligned}$$

Now we come to bound the second term $\mathbf{p}_2$ in equation (A.6). By the definition of the temporal-difference error δ_h^t in (A.1), for any $h \in [H], t \in [T]$, we note that

$$\begin{aligned} \delta_h^t(s_h^t, \omega_h^t, a_h^t) &= (v_h^t + P_h V_{h+1}^t - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t) \\ &= (v_h^t + P_h V_{h+1}^t - Q_h^{\pi^t})(s_h^t, \omega_h^t, a_h^t; C^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t; C^t) \\ &= (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s_h^t, \omega_h^t, a_h^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t, \omega_h^t, a_h^t). \end{aligned}$$

where the last equality follows the Bellman equation (2.1). Furthermore, using $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$

defined in (A.2), we have

$$
\begin{aligned}
&V_h^t(s_h^t;C^t) - V_h^{\pi^t}(s_h^t;C^t) \\
&\quad =(V_h^t - V_h^{\pi^t})(s_h^t;C^t) - \delta_h^t(s_h^t,\omega_h^t,a_h^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t,\omega_h^t,a_h^t;C^t) \\
&\qquad + (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s_h^t,\omega_h^t,a_h^t;C^t) \\
&\quad =(V_h^t - \widetilde{V}_h^t)(s_h^t;C^t) - \delta_h^t(s_h^t,\omega_h^t,a_h^t) + (\widetilde{V}_h^t - V_h^{\pi^t})(s_h^t;C^t) + (Q_h^{\pi^t} - Q_h^t)(s_h^t,\omega_h^t,a_h^t;C^t) \\
&\qquad + \big(P_h(V_{h+1}^t - V_{h+1}^{\pi^t})\big)(s_h^t,\omega_h^t,a_h^t;C^t) - (V_{h+1}^t - V_{h+1}^{\pi^t})(s_{h+1}^t;C^t) + (V_{h+1}^t - V_{h+1}^{\pi^t})(s_{h+1}^t;C^t) \\
&\quad =\big[V_{h+1}^t(s_{h+1}^t;C^t) - V_{h+1}^{\pi^t}(s_{h+1}^t;C^t)\big] + \big[V_h^t(s_h^t;C^t) - \widetilde{V}_h^t(s_h^t;C^t)\big] - \delta_h^t(s_h^t,\omega_h^t,a_h^t) + \zeta_{t,h}^1 + \zeta_{t,h}^2.
\end{aligned}
$$

Applying the above equation recursively, we get that

$$
\begin{aligned}
V_1^t(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t) =& V_{H+1}^t(s_H^t;C^t) - V_{H+1}^{\pi^t}(s_H^t;C^t) + \sum_{h\in[H]} \big[V_h^t(s_h^t;C^t) - \widetilde{V}_h^t(s_h^t;C^t)\big] \\
&- \sum_{h\in[H]} \delta_h^t(s_h^t,\omega_h^t,a_h^t) + \sum_{h\in[H]} (\zeta_{t,h}^1 + \zeta_{t,h}^2).
\end{aligned}
$$

Again by Bellman equation (2.1), we have,

$$
V_h^t(s_h^t;C^t) - \widetilde{V}_h^t(s_h^t;C^t) = \big\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \big\rangle_{\Omega\times\mathcal{A}}(s_h^t;C^t).
$$

Then we use $V_{H+1}^t = V_{H+1}^{\pi^t} = 0$ to simplify the decomposition to the following form:

$$
\begin{aligned}
V_1^t(s_1^t;C^t) - V_1^{\pi^t}(s_1^t;C^t) =& \sum_{h\in[H]} \big\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \big\rangle_{\Omega\times\mathcal{A}}(s_h^t;C^t) \\
&- \sum_{h\in[H]} \delta_h^t(s_h^t,\omega_h^t,a_h^t) + \sum_{h\in[H]} (\zeta_{t,h}^1 + \zeta_{t,h}^2).
\end{aligned}
$$

Therefore, combining $\mathbf{p}_1$ and $\mathbf{p}_2$, we can conclude the proof of this lemma.

$$\begin{aligned}
\mathrm{Reg}(T, \mu^*) &= \sum_{t\in[T]} \Big[V_1^*(s_1^t; C^t) - V_1^{\pi^t}(s_1^t; C^t)\Big] \\
&= \sum_{t\in[T]} \sum_{h\in[H]} \big\{\mathbb{E}_{\mu_h^*,\pi_h^*}[\delta_h^t(s_h, \omega_h, a_h)|s_1 = s_1^t] - \delta_h^t(s_h^t, \omega_h^t, a_h^t)\big\} + \sum_{t\in[T]} \sum_{h\in[H]} (\zeta_{t,h}^1, \zeta_{t,h}^2) \\
&\quad + \sum_{t\in[T]} \sum_{h\in[H]} \mathbb{E}_{\mu_h^*,\pi_h^*}\big[\big\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t\big\rangle_{\Omega\times\mathcal{A}}(s_h; C^t)|s_1 = s_1^t\big] \\
&\quad + \sum_{t\in[T]} \sum_{h\in[H]} \big\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t\big\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t).
\end{aligned}$$

Therefore, we conclude the proof of the lemma. □

A.2.4 Proof of Lemma A.2 – Optimism

Proof. Proof. In the following lemma, we firstly bound the difference between the Q-function maintained in Algorithm A.1 (without bonus) and the real Q-function of any policy π by their expected difference at next step, plus an error term. This error term can be upper bounded by our bonus with high probability. This lemma can be derived from Lemma B.4 in [Jin et al., 2020] with slight revisions.

Lemma A.7. *Set $\lambda = \max\{1, \Psi^2\}$. There exists an absolute constant c_ρ such that for $\rho = c_\rho d_\psi H\sqrt{\iota}$ where $\iota = \log(2d_\psi\Psi^2 T/\delta)$, and for any fixed policy π, with probability at least $1-\delta/2$, we have for all $s \in \mathcal{S}$, $\omega \in \Omega$, $a \in \mathcal{A}$, $h \in [H]$, $t \in [T]$,*

$$\psi(s, \omega, a)^\top q_h^t - Q_h^\pi(s, \omega, a) = P_h(V_{h+1}^t - V_{h+1}^\pi)(s, \omega, a) + \Delta_h^t(s, \omega, a),$$

for some $\Delta_h^t(s, \omega, a)$ that satisfies $|\Delta_h^t(s, \omega, a)| \le \rho\|\psi(s, \omega, a)\|_{(\Gamma_h^t)^{-1}}$.

Now we are ready to prove Lemma A.2. By the definition of δ_h^t in (A.1), we have $\delta_h^t = (v_h + P_h V_{h+1}^t - Q_h^t) = (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t}) + (Q_h^{\pi^t} - Q_h^t)$. Therefore, by the construction

of Q_h^t in Algorithm A.1, we obtain that

$$\begin{aligned}\delta_h^t(s,\omega,a) &\geq (P_h V_{h+1}^t - P_h V_{h+1}^{\pi^t})(s,\omega,a) + Q_h^{\pi^t}(s,\omega,a) - \left(\psi(s,\omega,a)^\top q_h^t + \rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}}\right) \\ &= -\Delta_h^t(s,\omega,a) - \rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}} \geq -2\rho\|\psi(s,\omega,a)\|_{(\Gamma_h^t)^{-1}},\end{aligned}$$

which concludes the proof. □

A.2.5 Proof of Lemma A.3 – Bounding Term (iii)

Proof. Proof. Denote the optimal signaling schemes corresponding to the real prior μ_h^* and the estimated prior μ_h^t respectively as

$$\pi_h' = \operatorname*{argmax}_{\pi_h \in \mathrm{Pers}(\mu_h^*)} \langle Q_h^t, \mu_h^* \otimes \pi_h \rangle_{\Omega\times\mathcal{A}}(\cdot; C^t) \quad \text{and} \quad \pi_h'' = \operatorname*{argmax}_{\pi_h \in \mathrm{Pers}(\mu_h^t)} \langle Q_h^t, \mu_h^t \otimes \pi_h \rangle_{\Omega\times\mathcal{A}}(\cdot; C^t),$$

where the Q-function Q_h^t is given by Algorithm A.1. Notably, π_h' is different from the truly optimal policy μ_h^*, since π_h' is computed based on the approximate Q-function Q_h^t. By definition, we can decompose the difference as follows:

$$\begin{aligned}\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) =& \langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^* \otimes \pi_h' \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) && \text{(A.7)} \\ &+ \langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) && \text{(A.8)} \\ &+ \langle Q_h^t, \mu_h^t \otimes \pi_h'' - \mu_h^t \otimes \pi_h^t \rangle_{\Omega\times\mathcal{A}}(s_h; C^t). && \text{(A.9)}\end{aligned}$$

By definition, equation (A.7) is always non-positive. Apply Lemma A.15 to equation (A.8) and we can get

$$\begin{aligned}\langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) \leq& \operatorname{Gap}\left(s_h, \mu_h^*(\cdot|c_h^t), \mathrm{B}_1\big(\mu_h^*(\cdot|c_h^t), \big\|\mu_h^*(\cdot|c_h^t) - \mu_h^t(\cdot|c_h^t)\big\|_1\big); Q_h^t\right) \\ &+ \frac{H}{2}\big\|\mu_h^*(\cdot|c_h^t) - \mu_h^t(\cdot|c_h^t)\big\|_1.\end{aligned}$$

According to Corollary A.13, we can bound the above equation with the norm of feature vector and the radius of confidence region for θ_h.

$$\langle Q_h^t, \mu_h^* \otimes \pi_h' - \mu_h^t \otimes \pi_h'' \rangle_{\Omega\times\mathcal{A}}(s_h; C^t) \leq \big(\frac{HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\big)\beta \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

We also note that equation (A.9) is equal to Gap $\big(s_h, \mu_h^t(\cdot|c_h^t), \mu_{\mathcal{B}_h^t}(\cdot|c_h^t); Q_h^t\big)$. By Lemma A.14, on the event $\{\theta_h^* \in \mathcal{B}_h^t\}$, we have

$$\text{Gap}\,\big(s_h, \mu_h^t(\cdot|c_h^t), \mu_{\mathcal{B}_h^t}(\cdot|c_h^t); Q_h^t\big) \leq \frac{2HL_\mu K}{p_0 D}\beta \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Therefore, on the given event, we have

$$\mathbb{E}_{\mu_h^*,\pi_h^*}\big[\langle Q_h^t, \mu_h^*\otimes\pi_h^* - \mu_h^t\otimes\pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h; C^t)|s_1 = s_1^t\big] \leq \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Summing up together, we get

$$\sum_{t\in[T]}\sum_{h\in[H]} \mathbb{E}_{\mu_h^*,\pi_h^*}\big[\langle Q_h^t, \mu_h^* \otimes \pi_h^* - \mu_h^t \otimes \pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h; C^t)|s_1 = s_1^t\big] \leq \left(\frac{3HL_\mu K}{p_0 D} + \frac{HL_\mu K}{2}\right)\beta \sum_{t\in[T]}\sum_{h\in[H]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Therefore, we conclude the proof of Lemma A.3. □

A.2.6 Proof of Lemma A.4 – Bounding Term (iv)

Proof. Proof. By definition, we can rewrite the difference in Lemma A.4 as

$$\begin{aligned}\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t\rangle_{\Omega\times\mathcal{A}}(s_h^t; C^t) &= \int_{\Omega\times\mathcal{A}} \big[\mu_h^t(\omega|c_h^t) - \mu_h^*(\omega|c_h^t)\big]\pi_h^t(s,\omega,a)Q_h^t(s,\omega,a)\mathrm{d}a\mathrm{d}\omega \\ &= \int_\Omega \big[\mu_h^t(\omega|c_h^t) - \mu_h^*(\omega|c_h^t)\big] \int_{\mathcal{A}} \pi_h^t(s,\omega,a)Q_h^t(s,\omega,a)\mathrm{d}a\mathrm{d}\omega.\end{aligned}$$

By Holder's inequality, we have

$$\left|\left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(s_h^t; C^t)\right| \leq \left\|\mu_h^t(\cdot|c_h^t) - \mu_h^*(\cdot|c_h^t)\right\|_1 \sup_{\omega \in \Omega} \left| \int_{\mathcal{A}} \pi_h^t(s, \omega, a) Q_h^t(s, \omega, a) \mathrm{d}a \right|.$$

Since $Q_h^t \leq H$ for any $h \in [H]$ and $t \in [T]$, the inequality can be simplified to

$$\left|\left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(s_h^t; C^t)\right| \leq H \left\|\mu_h^t(\cdot|c_h^t) - \mu_h^*(\cdot|c_h^t)\right\|_1.$$

With the assumption of the prior and link function, on the given event, we obtain that

$$\sum_{t \in [T]} \sum_{h \in [H]} \left\langle Q_h^t, (\mu_h^t - \mu_h^*) \otimes \pi_h^t \right\rangle_{\Omega \times \mathcal{A}}(s_h^t; C^t) \leq H L_\mu K \beta \sum_{h \in [H]} \sum_{t \in [T]} \|\phi(c_h^t)\|_{(\Sigma_h^t)^{-1}}.$$

Therefore, we conclude the proof of Lemma A.4. □

A.2.7 Auxiliary Lemmas

This section presents several auxiliary lemmas and their proofs.

Lemma A.8 (Martingale Bound; [Cai et al., 2020])**.** *For $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ defined in (A.2) and for any fixed $\delta \in (0, 1)$, with probability at least $1 - \delta/2$, we have*

$$\sum_{t \in [T]} \sum_{h \in [H]} (\zeta_{t,h}^1 + \zeta_{t,h}^2) \leq \sqrt{16 T H^3 \log(4/\delta)}.$$

Proof. Proof. See [Cai et al., 2020] for a detailed proof. □

Lemma A.9. *Suppose that $\phi_1, \phi_2, \ldots, \phi_T \in R^{d_\phi \times d}$ and for any $1 \leq i \leq T$, there exists a constant $\Phi > 0$ such that $\|\phi_i\| \leq \Phi$. Let $\Sigma_t = \lambda I_{d_\phi} + \sum_{i \in [t-1]} \phi_i \phi_i'$ for some $\lambda \geq \Phi^2$. Then,*

$$\sum_{t \in [T]} \|\phi_t\|_{(\Sigma_t)^{-1}} \leq \sqrt{2 d_\phi T \log(1 + T\Phi^2/(\lambda d_\phi))}.$$

Proof. Proof. Firstly, we apply Cauchy-Schwartz inequality,

$$\sum_{t\in[T]} \|\phi_t\|_{(\Sigma_t)^{-1}} \leq \sqrt{T \sum_{t\in[T]} \|\phi_t\|^2_{(\Sigma_t)^{-1}}}.$$

Since $\|\phi_t\|_{(\Sigma_t)^{-1}} = \sqrt{\phi_t^\top (\Sigma_t)^{-1}\phi_t} \leq \sqrt{\lambda^{-1}\phi_t^\top \phi_t} \leq \Phi/\sqrt{\lambda} \leq 1$, we can use Lemma A.10 to bound the sum of squares:

$$\begin{aligned}\sum_{t\in[T]} \|\phi_t\|_{(\Sigma_t)^{-1}} &\leq \sqrt{2T \log(\det(\Sigma_T)\det(\Sigma_1)^{-1})} \\ &\leq \sqrt{2d_\phi T \log(1 + T\Phi^2/(\lambda d_\phi))}.\end{aligned}$$

The last inequality is derived from Lemma A.11. □

Lemma A.10 (Sum of Potential Function; [Agarwal et al., 2019])**.** *For any sequence of* $\{\phi_t\}_{t\in[T]}$*, let* $\Sigma_t = \lambda I_h + \sum_{t\in[t-1]} \phi_i \phi_i'$ *for some* $\lambda \geq 0$*. Then we have*

$$\sum_{t\in[T]} \min\{\|\phi_t\|^2_{(\Sigma_t)^{-1}}, 1\} \leq 2\log(\det(\Sigma_T)\det(\Sigma_1)^{-1}).$$

Proof. Proof. See [Agarwal et al., 2019] for a detailed proof. □

Lemma A.11 (Determinant-Trace Inequality)**.** *Suppose that* $\phi_1, \phi_2, \ldots, \phi_T \in R^{d_\phi \times d}$ *and for any* $1 \leq i \leq T$*, there exists a constant* $\Phi > 0$ *such that* $\|\phi_i\| \leq \Phi$*. Let* $\Sigma_t = \lambda I_{d_\phi} + \sum_{i\in[t-1]} \phi_i \phi_i'$ *for some* $\lambda \geq 0$*. Then,*

$$\det(\Sigma_t) \leq \left(\lambda + t\Phi^2/d_\phi\right)^{d_\phi}.$$

Proof. Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_h$ be the eigenvalues of Σ_t. Since Σ_t is positive definite, its eigenvalues are positive. Also, note that $\det(\Sigma_t) = \prod_{s=1}^{d_\phi} \lambda_s$ and $\mathrm{tr}\,(\Sigma_t) = \sum_{s=1}^{h} \lambda_s$. By

inequality of arithmetic and geometric means

$$\det(\Sigma_t) \leq (\operatorname{tr}(\Sigma_t)/d_\phi)^{d_\phi}.$$

It remains to upper bound the trace:

$$\operatorname{tr}(\Sigma_t) = \operatorname{tr}(\lambda I_{d_\phi}) + \sum_{i=1}^{t-1} \operatorname{tr}(\phi_i \phi_i') = d_\phi \lambda + \sum_{i=1}^{t-1} \|\phi_i\|^2 \leq d_\phi \lambda + t\Phi^2$$

and the lemma follows. □

A.3 Inducing Robust Equilibria via Pessimism

One necessary prerequisite is that the signaling policy given by `OP4` has to be persuasive to ensure receivers to take recommended actions. However, the optimal signaling policy that is persuasive for the estimated prior can hardly be also persuasive for the true prior, even if the estimation is quite close to it. To ensure persuasiveness under the prior estimation error, we adopt pessimism principle to select a signaling policy that is robustly persuasive for all the priors in the confidence region. And we shall quantify the extra utility loss suffered by the pessimism principle. In this subsection, we start by showing that there exists a robust signaling scheme that suffers only $O(\epsilon)$ utility loss compared to the optimal expected utility of persuasion algorithm designed with precise knowledge of the prior. Formally, in basic MPP, given any fixed Q-function $Q(\cdot,\cdot,\cdot)$, we define the *robustness gap* for some state $s \in \mathcal{S}$ and any prior $\mu \in \mathcal{B} \subseteq \Delta(\Omega)$ as

$$\operatorname{Gap}\left(s, \mu, \mathcal{B}; Q\right) \triangleq \max_{\pi \in \operatorname{Pers}(\mu,u)} \left\langle Q, \mu \otimes \pi \right\rangle_{\Omega \times \mathcal{A}}(s) - \max_{\pi \in \operatorname{Pers}(\mathcal{B},u)} \left\langle Q, \mu \otimes \pi \right\rangle_{\Omega \times \mathcal{A}}(s). \quad \text{(A.10)}$$

We let $\mathrm{B}(\mu, \epsilon) = \{\mu' \in \Delta(\Omega) : \left\|\mu - \mu'\right\|_1 \leq \epsilon\}$ be the ℓ_1-norm ball centered the prior distribution μ with radius ϵ.

Lemma A.12 (Pessimism)**.** *Under (p_0, D)-regularity, for all $\epsilon > 0$, given a Q-function Q, for any state $s \in \mathcal{S}$, we have*

$$\mathrm{Gap}\left(s, \mu^*, \mathrm{B}(\mu^*, \epsilon); Q\right) \leq \frac{H\epsilon}{p_0 D}.$$

The proof is given in Appendix A.3.1. This result extends Proposition 1 in [Zu et al., 2021]. Notice that the upper bound of $\mathrm{Gap}(\cdot;\cdot)$ does not depend on the value of Q, which is important for our analysis. Once given a signaling algorithm, at each episode $t \in [T]$ and each step $h \in [H]$, we are able to obtain an estimation of Q-function with an explicit form. It is equivalent to the "known" Q-function mentioned in equation (A.10). Using $\mathrm{Gap}(\cdot;\cdot)$, we can estimate the expected sender's utility loss for choosing a signaling mechanism that is persuasive for all priors in a subset. Moreover, if we consider the dependence on context for priors and add the linear assumption of priors to the proceeding lemma, we can bound $\mathrm{Gap}(\cdot;\cdot)$ by the difference of linearity parameter θ.

Corollary A.13. *Under (p_0, D)-regularity and Assumption 2.6 and 2.7, given a Q-function Q and context c, for any state $s \in \mathcal{S}$, prior $\mu_{\theta^*}(\cdot|c)$ and confidence region $\mathcal{B} = \{\mu_{\theta'}(\cdot|c) : \theta' \in \mathrm{B}_\Sigma(\theta^*, \epsilon)\}$, we have* $\mathrm{Gap}(s, \mu_{\theta^*}(\cdot|c), \mathcal{B}; Q) \leq HL_\mu K \|\phi(c)\|_{\Sigma^{-1}} \epsilon/(p_0 D)$.

A.3.1 Proof of Lemma A.12 – Pessimism

Proof. Proof. In this proof, let μ, u be the ground-truth prior and sender utility of the persuasion instance. Given the full knowledge of u, we prove with an explicit construction of a signaling scheme that is robustly persuasive for any prior in $\mathrm{B}(\mu, \epsilon)$ and achieve the expected utility at least $\max_{\pi \in \mathrm{Pers}(\mu,u)} \langle Q, \mu \otimes \pi \rangle_{\Omega \times \mathcal{A}}(s) - H\epsilon/(p_0 D)$. To simplify the notation, we omit the s in u, Q and $\mathcal{W}$.

Let $\pi^* = \arg\max_{\pi \in \mathrm{Pers}(\mu,u)} \langle Q, \mu \otimes \pi \rangle_{\Omega \times \mathcal{A}}$ be a direct scheme without loss of generality [Kamenica and Gentzkow, 2011b]. For each $a \in \mathcal{A}$, let $\mu_a(\cdot) \coloneqq \mu(\cdot) \odot \pi^*(a|\cdot)$ denote

the posterior of outcome (i.e., kernel [1]) that action a is recommended by π, so the prior can be composed as $\mu(\cdot) = \sum_{a\in\mathcal{A}} \mu_a(\cdot)$. Since π is persuasive, we know $\int_{\omega\in\Omega} \mu_a(\omega) \left[u(\omega, a) - u(\omega, a')\right] \geq 0, \forall a' \in \mathcal{A}$.

Let π^0 be the fully revealing signaling scheme that always recommends (signals) the action that maximizes the receivers' utility at the realized outcome. For each $a \in \mathcal{A}$, let $\eta_a(\cdot) \coloneqq \mu(\cdot) \odot \pi^0(a|\cdot)$ denote the posterior of outcome that action a is recommended by π^0, so the prior can be composed as $\mu(\cdot) = \sum_{a\in\mathcal{A}} \eta_a(\cdot)$. By regularity condition, we have

$$\int_{\omega\in\Omega} \eta_a(\omega) \left[u(\omega, a) - u(\omega, a')\right] \geq \int_{\omega\in\mathcal{W}_a(D)} \eta_a(\omega) \left[u(\omega, a) - u(\omega, a')\right] \geq p_0 D, \quad \forall a' \in \mathcal{A}.$$

We now show that the signaling scheme $\pi' = (1-\delta)\pi^* + \delta\pi^0$ is persuasive for any prior $\widetilde{\mu} \in \mathrm{B}(\mu, \epsilon)$ with $\delta = \frac{\epsilon}{p_0 D}$. One simple way to interpret this "compound" signaling scheme is to follow π^* with probability $(1-\delta)$ and follow π^0 with probability δ. Hence, given a recommended action a, the receiver would compute the posterior as $\mu'_a = (1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)$. Let $\mu'_a, \widetilde{\mu}_a$ be the outcome posterior of π' recommending action a under the true prior μ (resp. the perturbed prior $\widetilde{\mu}$). So $\mu'_a(\cdot) = \mu(\cdot) \odot \pi'(a|\cdot)$ and $\widetilde{\mu}_a(\cdot) = \widetilde{\mu}(\cdot) \odot \pi'(a|\cdot)$. By definition of persuasiveness, we need to show that for any recommended action (signal from π') $a \in \mathcal{A}$, the action a maximizes the receiver's utility under μ'_a. This follows from

1. In this proof, we will directly work with the posterior without normalization (kernel) to simplify our notations and derivations, because $\int_{\omega\in\Omega} \mu_a(\omega) \left[u(\omega, a) - u(\omega, a')\right] \geq 0 \iff \int_{\omega\in\Omega} \frac{\mu_a(\omega)}{\int_{\omega\in\Omega} \mu_a(\omega)} \left[u(\omega, a) - u(\omega, a')\right] \geq 0$. We use $\odot$ to denote the Hadamard product.

the decomposition below,

$$
\begin{aligned}
&\int_{\omega\in\Omega} \widetilde{\mu}_a \cdot \left[u(\omega,a) - u(\omega,a')\right] \\
&\geq \int_{\omega\in\Omega} \mu'_a \cdot \left[u(\omega,a) - u(\omega,a')\right] - \left\|\widetilde{\mu}_a - \mu'_a\right\|_1 \\
&\geq \int_{\omega\in\Omega} \left[(1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)\right] \cdot \left[u(\omega,a) - u(\omega,a')\right] - \|\widetilde{\mu}_a - \mu_a\|_1 \\
&= \int_{\omega\in\Omega} (1-\delta)\mu_a(\omega) \left[u(\omega,a) - u(\omega,a')\right] + \int_{\omega\in\Omega} \delta\eta_a(\omega) \left[u(\omega,a) - u(\omega,a')\right] - \|\widetilde{\mu}_a - \mu_a\|_1 \\
&\geq \delta p_0 D - \|\widetilde{\mu}_a - \mu_a\|_1 \\
&= \epsilon - \|\widetilde{\mu}_a - \mu_a\|_1 \geq 0.
\end{aligned}
$$

The first inequality is by the fact that $u(\omega,a) \in [0,1]$ for any ω, a and thus $\sum_a (\widetilde{\mu}_a - \mu'_a) \cdot \left[u(\omega,a) - u(\omega,a')\right] \leq \left\|\widetilde{\mu}_a - \mu'_a\right\|_1$. The second inequality is from $\mu'_a = (1-\delta)\mu_a(\omega) + \delta\eta_a(\omega)$. The third inequality is by construction of μ_a and η_a induced by signaling scheme π and π^0. The last inequality is by the fact that $\left\|\widetilde{\mu}_a - \mu'_a\right\|_1 = \left\|(\widetilde{\mu} - \mu') \odot \pi'(a|\cdot)\right\|_1 \leq \left\|\widetilde{\mu} - \mu'\right\|_1 = \epsilon$, since $\left\|\pi'(a|\cdot)\right\|_\infty \leq 1$

It remains to show the expected utility under signaling scheme π' is at least $\left\langle Q, \mu \otimes \pi^* \right\rangle_{\Omega\times\mathcal{A}} - H\epsilon/(p_0 D)$. This is due to the following inequalities,

$$
\begin{aligned}
\left\langle Q, \mu \otimes \pi' \right\rangle_{\Omega\times\mathcal{A}} - \left\langle Q, \mu \otimes \pi^* \right\rangle_{\Omega\times\mathcal{A}} &= \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega) \left[\pi'(a|\omega) - \pi^*(a|\omega)\right] Q(\omega,a) \\
&= \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega) \left[\delta\pi^0(a|\omega) - \delta\pi^*(a|\omega)\right] Q(\omega,a) \\
&\geq -\delta \int_{\omega\in\Omega, a\in\mathcal{A}} \mu(\omega)\pi(a|\omega) Q(\omega,a) \\
&\geq -H\delta = -\frac{H\epsilon}{p_0 D}.
\end{aligned}
$$

The first and second equalities use the definition and linearity. The third and last inequalities use the fact that $\mathbb{E}[Q(\omega,a)] \in [0,H]$ and remove the positive term. □

A.3.2 Proof of Corollary A.13

Proof. Proof. According to Assumption 2.6 for the prior, we can show that for any $\mu_{\theta'}(\cdot|c) \in \mathcal{B}$,

$$\|\mu_\theta(\cdot|c) - \mu_{\theta'}(\cdot|c)\|_1 \leq L_\mu \big\| f(\phi(c)^\top \theta) - f(\phi(c)^\top \theta') \big\|.$$

Moreover, by Assumption 2.7 for the link function $f(\cdot)$, we have

$$\|\mu_\theta(\cdot|c) - \mu_{\theta'}(\cdot|c)\|_1 \leq L_\mu K \big\| \phi(c)^\top (\theta - \theta') \big\| \leq L_\mu K \, \|\phi(c)\|_{\Sigma^{-1}} \, \epsilon.$$

Therefore, $\mathcal{B} \subseteq \mathrm{B}(\mu_\theta(\cdot|c), L_\mu K \|\phi(c)\|_{\Sigma^{-1}} \epsilon)$, and by Lemma A.12, we can conclude the result.

□

A.3.3 Properties for the Robustness Gap

We present the robustness gap Gap for the ground-truth prior in Lemma A.12. For the estimation of prior μ_h^t given in Algorithm A.1 which may not satisfy the regularity condition, we also have corresponding robustness gap.

Lemma A.14. *For any $h \in [H], t \in [T]$ and $s \in \mathcal{S}$, on the event of $\{\theta_h^* \in \mathcal{B}_h^t\}$, we have*

$$\mathrm{Gap}(s, \mu_h^t, \mathrm{B}(\mu_h^t, \epsilon_h^t); Q_h^t) \leq \frac{2H\epsilon}{p_0 D}.$$

Proof. Proof. For any fixed action $a \in \mathcal{A}$, on the given event, we have

$$\begin{aligned}
\mathbb{P}_{\omega \sim \mu_h^t(\cdot)}[\omega \in \mathcal{W}_{s,a}(D)] &= \int_{\omega \in \Omega} \mu_h^t(\omega) \mathbb{I}(\omega \in \mathcal{W}_{s,a}(D)) \mathrm{d}\omega \\
&= \int_{\omega \in \Omega} \mu_h^*(\omega) \mathbb{I}(\omega \in \mathcal{W}_{s,a}(D)) \mathrm{d}\omega + \int_{\omega \in \Omega} [\mu_h^t(\omega) - \mu_h^*(\omega)] \mathbb{I}(\omega \in \mathcal{W}_{s,a}(D)) \mathrm{d}\omega \\
&\geq \int_{\omega \in \Omega} \mu_h^*(\omega) \mathbb{I}(\omega \in \mathcal{W}_{s,a}(D)) \mathrm{d}\omega + \|\mu_h^t - \mu_h^*\|_1 \\
&\geq p_0 - \epsilon_h^t,
\end{aligned}$$

where $\mathbb{I}$ is the indicating function. The last inequality uses the regularity condition for the real prior μ_h^*. For $\epsilon_h^t \le p_0/2$, we have $\mathbb{P}_{\omega\sim\mu_h^t(\cdot)}[\omega \in \mathcal{W}_{s,a}] \le p_0/2$. Then by Lemma A.12, we have

$$\mathrm{Gap}(s, \mu_h^t, \mathrm{B}(\mu_h^t, \epsilon_h^t); Q_h^t) \le \frac{2H\epsilon_h^t}{p_0 D}.$$

For $\epsilon_h^t > p_0/2$, the bound holds trivially since $2H\epsilon_h^t/(p_0 D) > H$. □

The robustness gap Gap defined in equation (A.10) measures the loss in value functions for being robustly persuasive for a subset of priors. In the following lemma, we show that we can also use Gap to bound the difference in expected optimal Q-functions between different priors.

Lemma A.15. *Denote $\mathcal{B}_{1,2} \coloneqq \mathrm{B}\big(\mu_1, \|\mu_1 - \mu_2\|_1\big)$ for any fixed state $s \in \mathcal{S}$ and $\mu_1, \mu_2 \in \Delta(\Omega)$. Then given a known Q-function $Q(\cdot,\cdot,\cdot)$, we have*

$$\max_{\pi_1 \in \mathrm{Pers}(\mu_1,u)} \big\langle Q, \mu_1 \otimes \pi_1 \big\rangle_{\Omega\times\mathcal{A}}(s) - \max_{\pi_2 \in \mathrm{Pers}(\mu_2,u)} \big\langle Q, \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) \le \mathrm{Gap}(s, \mu_1, \mathcal{B}_{1,2}; Q) + \frac{H}{2}\|\mu_1 - \mu_2\|_1.$$

Proof. Proof. Fix $\mu_1, \mu_2 \in \Delta(\Omega)$, we respectively choose the optimal signaling scheme

$$\pi_i = \operatorname*{argmax}_{\pi_i \in \mathrm{Pers}(\mu_i,u)} \big\langle Q, \mu_i \otimes \pi_i \big\rangle_{\Omega\times\mathcal{A}}(s),\ i = 1, 2.$$

Then among all the signaling schemes persuasive for all $\mathcal{B}_{1,2}$, let π_3 maximize $\big\langle Q, \mu_1 \otimes \pi \big\rangle_{\Omega\times\mathcal{A}}(s)$. Since π_3 is persuasive for μ_2, we know $\big\langle Q, \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) \ge \big\langle Q, \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s)$ by definition. Therefore, we have

$$\begin{aligned}
\big\langle Q, \mu_1 \otimes \pi_1 - \mu_2 \otimes \pi_2 \big\rangle_{\Omega\times\mathcal{A}}(s) &\le \big\langle Q, \mu_1 \otimes \pi_1 - \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) \\
&\le \big\langle Q, \mu_1 \otimes \pi_1 - \mu_1 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) + \big\langle Q, \mu_1 \otimes \pi_3 - \mu_2 \otimes \pi_3 \big\rangle_{\Omega\times\mathcal{A}}(s) \\
&= \mathrm{Gap}(s, \mu_1, \mathcal{B}_{1,2}; Q) + \frac{H}{2}\|\mu_1 - \mu_2\|_1.
\end{aligned}$$

The last equality uses the definition of Gap and Lemma A.16. □

Lemma A.16. *Given a Q-function $Q(\cdot,\cdot,\cdot) \in [0,H]$, for any fixed state $s \in \mathcal{S}$, $\mu_1, \mu_2 \in \Delta(\Omega)$ and any signaling scheme π, we have*

$$\left|\left\langle Q, \mu_1 \otimes \pi \right\rangle_{\Omega\times\mathcal{A}}(s) - \left\langle Q, \mu_2 \otimes \pi \right\rangle_{\Omega\times\mathcal{A}}(s)\right| \leq \frac{H}{2}\|\mu_1 - \mu_2\|_1.$$

Proof. Proof. Fix $\mu_1(\cdot), \mu_2(\cdot) \in \Delta(\Omega)$. For any $x \in \mathbb{R}$, we have

$$\begin{aligned}\left|\left\langle Q, \mu_1 \otimes \pi - \mu_2 \otimes \pi \right\rangle_{\Omega\times\mathcal{A}}(s) &= \left| \int_{\omega\in\Omega} [\mu_1(\omega) - \mu_2(\omega)] \left[\int_{a\in\mathcal{A}} \pi(a|s,\omega) Q(s,\omega,a) \mathrm{d}a - x \right] \mathrm{d}\omega \right| \\ &\leq \|\mu_1 - \mu_2\|_1 \cdot \sup_{\omega\in\Omega} \left| \int_{a\in\mathcal{A}} \pi(a|s,\omega) Q(s,\omega,a) \mathrm{d}a - x \right|,\end{aligned}$$

where the last inequality is derived from Holder's inequality. With Q-function taking values in $[0,H]$, we can set $x = H/2$ and achieve the optimality.

□

APPENDIX B

B.1 Omitted Proofs in Section 3.3

B.1.1 Proof of Proposition 3.1

Proof. We provide a constructive proof for the theorem by exhibiting an algorithm that computes the δ-RSE for any Stackelberg game. The main idea of our algorithm is to partition the simplex in sub-regions and search the δ-RSE candidate within each sub-region $\mathcal{X}_{(S,\widetilde{j},j)}$ by solving a linear program but with *relaxed* non-strict inequalities (thus in total we solve exponentially many LPs). The crux of our proof is to argue that, while in general the relaxation above is not tight, there always exists a region $\mathcal{X}_{(S,\widetilde{j},j)}$ for which the above relaxation is tight and moreover the leader achieves the best possible leader utility. This proves the existence of a δ-RSE.

We begin with the definition and analysis of different sub-regions of the leader's strategy space Δ^m. Each sub-region $\mathcal{X}_{(S,\widetilde{j},j)}$ is characterized by three factors: (1) Follower δ-optimal response action set $S \in 2^{[n]}$; (2) Follower action $\widetilde{j} \in S$ with maximal follower utility among actions in S; (3) Follower action $j \in S$ with the worst leader utility among actions in S (j and $\widetilde{j}$ are different in general). Mathematically, for any follower action set $S \in 2^{[n]}$ and any $\widetilde{j}, j \in S$,

$$\mathcal{X}_{(S,\widetilde{j},j)} := \left\{ \boldsymbol{x} \in \Delta^m \;\middle|\; \mathrm{BR}_\delta(\boldsymbol{x}) = S,\ u_f(\boldsymbol{x},\widetilde{j}) \geq u_f(\boldsymbol{x},k), u_l(\boldsymbol{x},j) \leq u_l(\boldsymbol{x},k), \forall k \in S \right\}.$$

Next, we make a few observations about $\mathcal{X}_{(S,\widetilde{j},j)}$.

First, $\bigcup_{\widetilde{j},j\in S,S\in 2^{[n]}} \mathcal{X}_{(S,\widetilde{j},j)} = \Delta^m$. This is because any $\boldsymbol{x} \in \Delta^m$ must induce some follower δ-optimal response set S, an optimal follower action as $\widetilde{j}$, and a follower response j that is the worst response for the leader. As a result, any $\boldsymbol{x}$ must belong to some $\mathcal{X}_{(S,\widetilde{j},j)}$.

Second, $\mathcal{X}_{(S,\widetilde{j},j)}$ is a (possibly open) *polytope* that can be expressed by linear constraints,

with *strict* inequality constraints. This is because both constraints $u_f(\boldsymbol{x}, \widetilde{j}) \geq u_f(\boldsymbol{x}, k)$ and $u_l(\boldsymbol{x}, j) \leq u_l(\boldsymbol{x}, k)$ are linear in variable $\boldsymbol{x} = (x_1, \cdots, x_m)$. Moreover, we have the following equivalent expression for the constraint $\mathrm{BR}_\delta(\boldsymbol{x}) = S$:

$$\mathrm{BR}_\delta(\boldsymbol{x}) = S \quad \Longleftrightarrow \quad \begin{cases} u_f(\boldsymbol{x}, k) > u_f(\boldsymbol{x}, \widetilde{j}) - \delta, \ \forall k \in S, & \text{and} \\ u_f(\boldsymbol{x}, k) \leq u_f(\boldsymbol{x}, \widetilde{j}) - \delta, \ \forall k \notin S. \end{cases} \tag{B.1}$$

Notably, the second set of constraints guarantees that any $k \notin S$ cannot be a δ-optimal follower response, which is essential for the definition of S. Different from typical linear constraints, the first constraint above is a strict inequality and thus the resultant polytope $\mathcal{X}_{(S,\widetilde{j},j)}$ might be an open set.

In search of the optimal leader strategy, the main idea of our algorithm is to solve multiple (in fact, exponentially many) linear programs by relaxing the strict inequality in (B.1). The correctness proof of this algorithm relies on a key insight that there always exists a tuple $(S, \widetilde{j}, j)$ such that the leader objective on this tuple achieves the best possible leader utility and moreover the above relaxation of the δ-optimal follower response constraints will be "tight" on this particular tuple (although this relaxation is not tight for other tuples in general).

For the convenience of exposure, we present the formal procedure in Algorithm B.1. At the high level, the algorithm enumerates all possible tuple $(S, \widetilde{j}, j)$ (Line 1), solves LP (B.2) for the optimal leader strategy $\boldsymbol{x}^*_{S,\widetilde{j},j}$ within (a relaxation of) the polytope $\mathcal{X}_{(S,\widetilde{j},j)}$ (Line 2), and finally picks the $LP(S^*, \widetilde{j}^*, j^*)$ whose solution $\boldsymbol{x}^*$ has the highest leader utility (Line 4). Notably, in order to optimize over a closed polytope to avoid the non-existence of optimal solutions, we had to relax the δ-optimal response constraint, i.e., the second constraint in LP (B.2), to be "$\geq$". Therefore, we thus suffer the risk of getting a solution $\boldsymbol{x}^*$ that cannot truly induce the expected follower δ-optimal response set S^*, i.e., those actions $k \in S^*$ such that $u_f(\boldsymbol{x}^*, k) = u_f(\boldsymbol{x}^*, \widetilde{j}^*) - \delta$. Here comes the crucial (though seemingly

ALGORITHM B.1: Computing δ-RSE via LP Relaxations

Input : leader utilities $u_l \in \mathbb{R}^{m\times n}$, follower utilities $u_f \in \mathbb{R}^{m\times n}$, parameter $\delta > 0$.
Output : δ-RSE $(\boldsymbol{x}^*, j^*)$.

1 **for** *any non-empty* $S \subseteq [n]$ *and any* $\widetilde{j}, j \in S$ **do**

2 Solve the following *relaxed* linear program, denoted by $LP(S, \widetilde{j}, j)$, to compute an optimal leader strategy $\boldsymbol{x}^*_{S,\widetilde{j},j}$ within (a relaxation of) $\mathcal{X}_{(S,\widetilde{j},j)}$:

$$\begin{array}{lll} \text{maximize} & \sum_{i=1}^m x_i \cdot u_l(i,j) & \\ \text{subject to} & \sum_{i=1}^m x_i \cdot u_f(i,\widetilde{j}) \geq \sum_{i=1}^m x_i u_f(i,k), & \text{for } k \in [n]. \\ & \sum_{i=1}^m x_i \cdot u_f(i,k) \geq \sum_{i=1}^m x_i \cdot u_f(i,\widetilde{j}) - \delta, & \text{for } k \in S \quad \text{(relaxed constraints).} \\ & \sum_{i=1}^m x_i \cdot u_f(i,k) \leq \sum_{i=1}^m x_i \cdot u_f(i,\widetilde{j}) - \delta, & \text{for } k \notin S. \\ & \sum_{i=1}^m x_i \cdot u_l(i,j) \leq \sum_{i=1}^m x_i \cdot u_l(i,k), & \text{for } k \in S. \\ & (x_1, \cdots, x_m) \in \Delta^m & \end{array} \tag{B.2}$$

3 Choose any tuple $(S^*, \widetilde{j}^*, j^*)$ that has the maximum leader utility. That is, let $\boldsymbol{x}^* = \boldsymbol{x}^*_{S^*,\widetilde{j}^*,j^*}$ then $u_l(\boldsymbol{x}^*, j^*) \geq u_l(\boldsymbol{x}^*_{S,\widetilde{j},j}, j)$ for any $S \subseteq [n]$ and any $\widetilde{j}, j \in S$.

4 Verify the second constraint of LP (B.2) for the chosen tuple $(S^*, \widetilde{j}^*, j^*)$, and let $\widehat{S}$ denote the set of all $k \in S^*$ such that the second constraint is *strict* at $\boldsymbol{x}^*$ for the k.

5 Let $\widehat{j} \in \widehat{S}$ denote the follower action with minimum leader utility, i.e., $u_l(\boldsymbol{x}^*, \widehat{j}) \leq u_l(\boldsymbol{x}^*, k)$ for any $k \in \widehat{S}$.

6 **return** $(\boldsymbol{x}^*, \widehat{j})$.

straightforward) step in Lines 5 and 6 to remove the invalid δ-optimal responses from the set S^* and ultimately construct a valid equilibrium strategy. It is easy to see that, with the adjustment at Line 5 and 6, Algorithm B.1 always outputs a valid strategy pair $(\boldsymbol{x}^*, \widehat{j})$ in the sense that $\widehat{j} \in \mathrm{BR}_\delta(\boldsymbol{x}^*)$ and $\widehat{j}$ is indeed the follower response that is worst for the leader in the δ-optimal response set under leader strategy $\boldsymbol{x}^*$. This follows simply by the construction of the algorithm, which removes all invalid follower δ-optimal responses $k \in S^*$, due to $u_f(\boldsymbol{x}^*, k) = u_f(\boldsymbol{x}^*, \widetilde{j}^*) - \delta$, out the set S^*. This adjustment leads to a strictly small subset $\widehat{S}$, which is truly the follower δ-optimal response set $\mathrm{BR}_\delta(\boldsymbol{x}^*)$. The algorithm then "re-appoints" $\widehat{j} \in \widehat{S}$ as the true worst follower response within $\widehat{S}$ in the sense of minimizing the leader utility.

What remains to argue is that the output strategy $(\boldsymbol{x}^*, \widehat{j})$ achieves at least the leader utility as that of the δ-RSE, and thus must be a δ-RSE strategy pair. First, since LP (B.2)

is a relaxation of the δ-RSE problem and $\cup_{\widetilde{j},j\in S,S\in 2^{[n]}}\mathcal{X}_{(S,\widetilde{j},j)} = \Delta^m$, the optimal objective value of LP (B.2) for the special tuple $(S^*, \widetilde{j}^*, j^*)$ picked at Line 4 of Algorithm B.1 must be at least the leader utility of the δ-RSE. Therefore, all we need to argue is that the leader utility under the valid strategy pair $(\boldsymbol{x}^*, \widehat{j})$ is at least the optimal objective of $LP(S^*, \widetilde{j}^*, j^*)$, which is exactly $u_l(\boldsymbol{x}^*, j^*)$. This is true because both $\widehat{j}, j^* \in S^*$, and the fourth constraint of $LP(S^*, \widetilde{j}^*, j^*)$ — i.e., j^* has the worst leader utility among all actions in S^* — implies $u_l(\boldsymbol{x}^*, j^*) \leq u_l(\boldsymbol{x}^*, \widehat{j})$ which is precisely the leader utility for the feasible strategy pair $(\boldsymbol{x}^*, \widehat{j})$.

Time Complexity Algorithm B.1 solves $O(n^2 2^n)$ linear programs of size $O(mn)$, each corresponding to a tuple $(S, \widetilde{j}, j)$. Hence, the stated time complexity follows. □

B.1.2 Proof of Proposition 3.2

Proof. Consider the Stackelberg game instance in Table B.1, where both the leader and follower have 3 actions. In this game, we can compute the SSE leader strategy and max-min leader strategy as follows:

$$\boldsymbol{x}_1 = (1,0,0) = \operatorname*{argmax}_{\boldsymbol{x}\in\Delta^m} \max_{j\in \mathrm{BR}(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \quad \text{and} \quad \boldsymbol{x}_2 = (0,0,1) = \operatorname*{argmax}_{\boldsymbol{x}\in\Delta^m} \min_{j\in[n]} u_l(\boldsymbol{x}, j)$$

Meanwhile, the δ-RSE leader strategy is $\boldsymbol{x}^* = (0,1,0)$ and leader utility is $u_{\mathrm{RSE}}(\delta) = 2c$. Next, we apply the SSE leader strategy and max-min leader strategy against the bounded rational follower.

$$\min_{j\in \mathrm{BR}_\delta(\boldsymbol{x}_1)} u_l(\ \boldsymbol{x}_1, j) = u_l(i_1, j_2) = c \quad \text{and} \quad \min_{j\in \mathrm{BR}_\delta(\boldsymbol{x}_2)} u_l(\boldsymbol{x}_2, j) = u_l(i_3, j_1) = c$$

while the δ-RSE leader utility is $u_{\mathrm{RSE}}(\delta) = 2c$, causing a constant gap of c between the δ-RSE leader utility and the leader utility of playing SSE strategy or max-min strategy under the setting that allows approximately optimal follower responses. As we can set c to be

arbitrarily close to 1/2, we have

$$\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}_1)} u_l(\boldsymbol{x}_1, j) < u_{\mathrm{RSE}}(\delta) - \frac{1}{2} \quad \textit{and} \quad \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}_2)} u_l(\boldsymbol{x}_2, j) < u_{\mathrm{RSE}}(\delta) - \frac{1}{2}.$$

u_l, u_f	j_1	j_2	j_3
i_1	$1, \delta$	c, δ	$0, 0$
i_2	$2c, \delta$	$2c, \delta$	$0, 0$
i_3	c, δ	c, δ	c, δ

Table B.1: A Stackelberg game instance whose SSE leader strategy is i_1, δ-RSE leader strategy is i_2 (for any $\delta > 0$), and max-min leader strategy is i_3, for any $c \in (0, 1/2)$.

□

B.1.3 Proof of Proposition 3.3

Proof. Consider a game with $m = 3, n = 2$ and utility matrices specified in Table B.2. Notice that the leader utility is maximized at the leader/follower action pair (i_2, j_1), while the leader must place probability mass at least $\frac{\delta}{\Delta}$ on i_1, resulting in the leader utility no more than $\Delta - \delta$. In contrast, if the follower plays the action j_2, the leader can achieve a strictly better utility $\Delta - c$ given that $c < \delta$. Hence, $\forall \delta \in (0, \Delta)$, the δ-RSE $(\boldsymbol{x}^*, j^*)$ have leader strategy $\boldsymbol{x}^* = (0, 1, 0)$ and follower response $j^* = j_2$. We can compute for $\boldsymbol{x}^*$ the leader and follower utility under different follower action, $u_l(\boldsymbol{x}^*, j_1) = \Delta, u_l(\boldsymbol{x}^*, j_2) = \Delta - c$, $u_f(\boldsymbol{x}^*, j_1) = u_f(\boldsymbol{x}^*, j_2) = \Delta$. Hence, the δ-best response to x^* is not unique, as $\mathrm{BR}_\delta(\boldsymbol{x}^*) = \{j_1, j_2\}$. Moreover, the leader utility strictly improves by some constant gap c if the follower does not follow the pessimistic tie-breaking rule, i.e.,

$$\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) < \max_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) - c.$$

As we can set c to be arbitrarily close to δ in the problem instance, we have

$$\min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) < \max_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) - \delta.$$

u_l, u_f	j_1	j_2
i_1	$0, \Delta$	$0, 0$
i_2	Δ, Δ	$\Delta - c, \Delta$
i_3	$0, 0$	$0, \Delta$

Table B.2: A class of game instances in which the δ-RSE leader strategy does not have unique δ-best response for any $0 < c < \delta < \Delta$.

□

B.1.4 Omitted Examples in the Proof of Theorem 3.4

Example for Property 1

We provide a Stackelberg game instance to show how the equality $u_{\mathrm{SSE}} = u_{\mathrm{RSE}}(0^+)$ fails to hold, if the non-degeneracy assumption $\Delta > 0$ condition is not satisfied. Consider the following game in Table B.3 where the leader only has a single action while the follower has two actions. It is easy to see that $\Delta = 0$ in this game. Clearly, the SSE leader utility would be 1 which is induced by follower responding with action j_2. In contrast, we have $\mathrm{BR}_{0^+}(i_1) = \{j_1, j_2\}$ and by definition we have $u_{\mathrm{RSE}}(0^+) = 0$ which is induced by follower responding with action j_1.

u_l, u_f	j_1	j_2
i_1	$0, 1$	$1, 1$

Table B.3: A Stackelberg game instance whose SSE leader utility is different from $u_{\mathrm{RSE}}(0^+)$.

However, note that $\Delta > 0$ is only a sufficient but not necessary condition in general game. More precisely, for the equality to be attainable, it is equivalent to whether the best

response region of the SSE follower action has non-zero measure, which ensures the existence of strategy $\boldsymbol{x}$ with $\mathrm{BR}_{0^+}(\boldsymbol{x}) = \{j^*\}$ and so is the limit towards the SSE $(\boldsymbol{x}^*, j^*)$.

Examples for Property 3

An example of continuous $u_{\mathbf{RSE}}(\delta)$. Table B.4 illustrates an instance whose inducibility gap $\Delta = 1$ and its $u_{\mathrm{RSE}}(\delta) = \begin{cases} 1 & \text{if } \delta \le \epsilon \\ \frac{1-\delta}{1-\epsilon} & \epsilon < \delta \le 1 \\ 0 & \delta > 1 \end{cases}$ is continuous for any $\delta > 0$.

u_l, u_f	j_1	j_2
i_1	$1, \frac{1+\epsilon}{2}$	$0, \frac{1-\epsilon}{2}$
i_2	$0, 0$	$0, 1$
i_3	$0, 1$	$0, 0$

Table B.4: A Stackelberg game instance whose $u_{\mathrm{RSE}}(\delta)$ is continuous in δ (ϵ is any constant within $[0, 1]$).

An example of discontinuous for $u_{\mathbf{RSE}}(\delta)$ at $\delta \ge \Delta$. Figure B.1 illustrates an example with discontinuous $u_{\mathrm{RSE}}(\delta)$. This directly showcases how $u_{\mathrm{RSE}}(\delta)$ *may* be discontinuous at $\delta \ge \Delta$.

u_l, u_f	j_1	j_2	j_3	j_4	j_5
i_1	$1, \epsilon$	0,0	0,0	0,0	0,0
i_2	0,0	$0.8, 2\epsilon$	0,0	0,0	0,0
i_3	0,0	0,0	$0.6, 3\epsilon$	0,0	0,0
i_4	0,0	0,0	0,0	$0.4, 4\epsilon$	0,0
i_5	0,0	0,0	0,0	0,0	$0.2, 1$

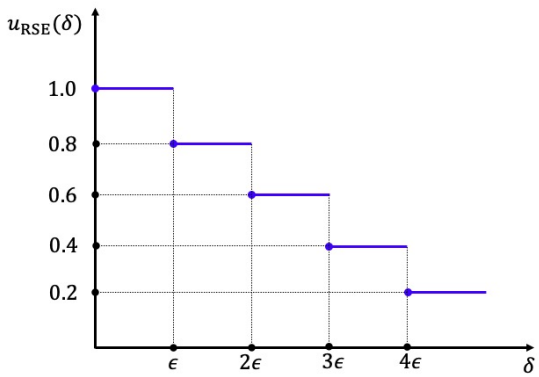


Figure B.1: A Stackelberg game instance with $\Delta = \epsilon$, and the corresponding $u_{\mathrm{RSE}}(\delta)$ as a function of δ.

B.1.5 Additional Discussions on Tie-breaking Rules and Uniqueness of δ-RSE

It is well known that SSE is almost always unique, and has generically unique leader payoff regardless of the follower's tie-breaking rules [Von Stengel and Zamir, 2010]. Hence, a natural question to ask is whether the δ-RSE shares any of these properties. The answer to the first part of the question is clear: δ-RSE is almost always unique in randomly generated game instances. This is because, as we fix the follower's pessimistic tie-breaking rule in Definition 2, the conditions for optimization problem in Equation (3.5) to admit multiple maximizers or minimizers have zero measure.

For the second part of the question, it is obvious that in general different tie-breaking rules result in different optimal leader strategies, and thus different utilities,

$$\max_{\boldsymbol{x}\in\Delta^m} \max_{j\in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) \neq \max_{\boldsymbol{x}\in\Delta^m} \min_{j\in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j).$$

Instead, a more interesting question is whether the δ-RSE leader strategy $\boldsymbol{x}^*$ has generically unique leader payoff if the follower could switch to different tie-breaking rules, i.e.,

$$\max_{j\in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) = \min_{j\in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j).$$

It is still obvious in the extreme cases where δ is larger than any utility difference between follower actions in the game, we have $\mathrm{BR}_\delta(\boldsymbol{x}) = [m], \forall \boldsymbol{x}$. However, if we were to restrict to reasonably small δ such that $\delta < \Delta$, it becomes a more difficult question. Lemma B.1 shows that δ-RSE do have generically unique leader payoff in this case when the number of leader actions $m = 2$. However, this observation is no longer true in general games. In Proposition 3.3, we are able to construct a class of instances that the follower's response to the δ-RSE leader strategy is not necessarily unique even in the case when δ is smaller than

the inducibility gap Δ. Hence, the follower's different tie-breaking rules could often result in radically different choices of the optimal leader strategy (with constant gaps among the corresponding leader payoffs). This is a shape contrast to the reckoning of SSE, where the follower's optimistic tie-breaking is almost without loss of generality, since the leader (in generic games) could improve his payoff by changing his strategy slightly so that the desired reply is unique in the equilibrium. As a result, we cannot apply the standard approaches used for SSE for the computation of δ-RSE. These differences are indeed rooted in the design of the two solution concepts: SSE serves as a standard solution concept (in generic games) where the follower's tie-breaking rule should not matter, whereas it is natural for δ-RSE, a robust solution concept, to focus on the worst case scenario and assume pessimistic tie-breaking rule.

Lemma B.1. *In any game with $m = 2$, $\Delta > 0, \delta \in (0, \Delta)$, at least one of its δ-RSE leader strategy $\boldsymbol{x}^*$ satisfies*

$$\max_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j) = \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}^*)} u_l(\boldsymbol{x}^*, j).$$

Proof. Pick one of δ-RSE leader strategy $\boldsymbol{x}^*$ in the game. The lemma clearly holds, when the δ-RSE leader strategy have $|\mathrm{BR}_\delta(\boldsymbol{x}^*)| = 1$. A key observation of this proof is the characterizations in Lemma B.2 such that it suffices to show that if the unique δ-RSE leader strategy $\boldsymbol{x}^*$ satisfies $\mathrm{BR}_\delta(\boldsymbol{x}^*) = \{j, j'\}$ for some $j, j' \in [n]$, then $u_l(\boldsymbol{x}^*, j) = u_l(\boldsymbol{x}^*, j')$. We prove this statement by contradiction:

Suppose there is a unique δ-RSE leader strategy $\boldsymbol{x}^*$ where $\mathrm{BR}_\delta(\boldsymbol{x}^*) = \{j, j'\}$, $u_l(\boldsymbol{x}^*, j) < u_l(\boldsymbol{x}^*, j')$. Under the pessimistic tie-breaking, the follower will respond with j. We first rule out the case that $u_l(\cdot, j)$ is a constant function, because we could find another leader strategy $\boldsymbol{x}$ such that $u_l(\boldsymbol{x}^*, j) = u_l(\boldsymbol{x}, j)$ and $\mathrm{BR}_\delta(\boldsymbol{x}) = \{j\}$, due to Property 1 in Lemma B.2.

Let $\boldsymbol{x}^L, \boldsymbol{x}^R$ be the two strategy on the boundary $\mathcal{X}(\delta; \{j, j'\})$. According to Property 3 in Lemma B.2, we can without loss of generality assume $\mathrm{BR}_\delta(\boldsymbol{x}^L) = \{j\}, \mathrm{BR}_\delta(\boldsymbol{x}^R) = \{j'\}$. We argue that $\boldsymbol{x}^*$ is not an RSE leader strategy, since $\max\left\{u_l(\boldsymbol{x}^L, j), u_l(\boldsymbol{x}^R, j')\right\} > u_l(\boldsymbol{x}^*, j)$.

Notice that, if we move $\boldsymbol{x}^*$ towards one of the two open boundary points of $\mathcal{X}(\delta; \{j, j'\})$ where $u_l(\cdot, j)$ is increasing, two possible cases can happen: 1) It reaches the point where $u_l(\boldsymbol{x}^*, j) = u_l(\boldsymbol{x}^*, j')$, and it violates the assumption. 2) We can set $\boldsymbol{x}^*$ to be arbitrarily close to the right of $\boldsymbol{x}^L$ or to the left of $\boldsymbol{x}^R$. If it is to the right of $\boldsymbol{x}^L$, then clearly $u_l(\boldsymbol{x}^L, j) > u_l(\boldsymbol{x}^*, j)$. If it is to the left of $\boldsymbol{x}^R$, then we have $u_l(\boldsymbol{x}^R, j') = \lim_{\boldsymbol{x}^* \to \boldsymbol{x}^R} u_l(\boldsymbol{x}^*, j') > \lim_{\boldsymbol{x}^* \to \boldsymbol{x}^R} u_l(\boldsymbol{x}^*, j)$. Hence, $\boldsymbol{x}^*$ is not an RSE leader strategy.

□

Lemma B.2 (A Characterization of δ-Best Response Regions)**.** *Denote by* $\mathcal{X}(\delta; S) := \{\boldsymbol{x} \in \Delta^m | \operatorname{BR}_\delta(x) = S\}$*, the* δ*-best response region for a subset of follower actions* $S \subseteq [n]$*.* $\mathcal{X}(\delta; S)$ *is a (possibly empty) convex set and satisfies the following properties when* $\Delta > 0, \delta \in (0, \Delta)$*:*

1. *If* $|S| = 1$*, then* $\mathcal{X}(\delta; S) \neq \varnothing$ *and* $\mathcal{X}(\delta; S) \supset \mathcal{X}(\delta'; S)$*,* $\forall \delta' \in (\delta, \Delta)$*.*

2. *If* $m = 2$ *and* $|S| \geq 3$*, then* $\mathcal{X}(\delta; S) = \varnothing$*.*

3. *If* $m = 2$ *and* $|S| = 2$*, then* $\mathcal{X}(\delta; S)$ *is an nonempty open set for any* $S = \{j, j'\}$ *that* $\exists \boldsymbol{x} \in \Delta^m$ *with* $u_f(\boldsymbol{x}, j) = u_f(\boldsymbol{x}, j')$*, the boundary of* $\mathcal{X}(\delta; S)$ *contains exactly two points* $\boldsymbol{x}^L, \boldsymbol{x}^R$ *with* $\operatorname{BR}_\delta(\boldsymbol{x}^L) = \{j\}, \operatorname{BR}_\delta(\boldsymbol{x}^R) = \{j'\}$*, and moreover,* $\mathcal{X}(\delta; S) \subset \mathcal{X}(\delta'; S)$*,* $\forall \delta' \in (\delta, \Delta)$*.*

Proof. Since $\operatorname{BR}_\delta(\boldsymbol{x})$ definition in Equation (3.4) specifies a set of convex constraints, $\mathcal{X}(\delta; S)$ forms a convex set as a result of finite intersection of convex sets. Below, we prove each one of the properties:

Property 1 It holds universally for any m. The first part can be directly verified using the definition of Δ. That is, $\forall j \in [n]$, there exists $\boldsymbol{x} \in \Delta^m$ such that $u_f(\boldsymbol{x}, j) - u_f(\boldsymbol{x}, j') \geq \Delta > \delta, \forall j \neq j'$, which means that $\mathcal{X}(\delta; \{j\}) \neq \varnothing$. For the second part, observe that the best response region of any action $j \in [n]$ can be written as $\mathcal{X}(\delta; j) = \{\boldsymbol{x} \in \Delta^m | u_f(\boldsymbol{x}, j) \geq \max_{j' \neq j} u_f(\boldsymbol{x}, j') + \delta\}$. It is thus clear that $\mathcal{X}(\delta; S) \supseteq \mathcal{X}(\delta'; S)$, $\forall \delta' \in (\delta, \Delta)$. To go from

$\supseteq$ to $\supset$, it suffices to show that $\mathcal{X}(\delta; S) \setminus \mathcal{X}(\delta'; S)$ is not empty for any $\delta' - \delta > 0$. That is, for any $\delta \in (0, \Delta)$, there exists $\boldsymbol{x}_j^\delta \in \Delta^m$ such that $u_f(\boldsymbol{x}_j^\delta, j) = \max_{j' \neq j} u_f(\boldsymbol{x}_j^\delta, j') + \delta$. This is because we have $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \Delta^m$ such that $u_f(\boldsymbol{x}_1, j) \geq \max_{j' \neq j} u_f(\boldsymbol{x}_1, j') + \Delta$ and $u_f(\boldsymbol{x}_2, j) \geq \max_{j' \neq j} u_f(\boldsymbol{x}_2, j')$ as $\Delta > 0$. So $\boldsymbol{x}_j^\delta \in \Delta^m$ as a convex combination of $\boldsymbol{x}_1, \boldsymbol{x}_2$. Hence, $\mathcal{X}(\delta; S) \supset \mathcal{X}(\delta'; S)$, $\forall \delta' \in (\delta, \Delta)$.

Property 2 We prove this by contradiction. Suppose there is such $\boldsymbol{x} \in \mathcal{X}(\delta; S)$ with $|S| > 2$. Pick three follower actions, say j_1, j_2, j_3 with the highest follower utility at $\boldsymbol{x}$, and we can assume without loss of generality that,

$$u_f(\boldsymbol{x}, j_1) \geq u_f(\boldsymbol{x}, j_2) \geq u_f(\boldsymbol{x}, j_3) > u_f(\boldsymbol{x}, j_1) - \delta.$$

We first argue that the above inequalities must be strict. That is, the following cases cannot hold:

1. $u_f(\boldsymbol{x}, j_1) = u_f(\boldsymbol{x}, j_2) = u_f(\boldsymbol{x}, j_3) > u_f(\boldsymbol{x}, j_1) - \delta$.
2. $u_f(\boldsymbol{x}, j_1) > u_f(\boldsymbol{x}, j_2) = u_f(\boldsymbol{x}, j_3) > u_f(\boldsymbol{x}, j_1) - \delta$.
3. $u_f(\boldsymbol{x}, j_1) = u_f(\boldsymbol{x}, j_2) > u_f(\boldsymbol{x}, j_3) > u_f(\boldsymbol{x}, j_1) - \delta$.

For the analysis below, we consider the gradient functions on the follower's utility over action j_1, j_2, j_3, as $f_1 = \frac{d[u_f((x,1-x),j_1)]}{dx}, f_2 = \frac{d[u_f((x,1-x),j_2)]}{dx}, f_3 = \frac{d[u_f((x,1-x),j_3)]}{dx}$. Since the utility functions are linear, we can treat f_1, f_2, f_3 as constants. Moreover, $f_1 \neq f_2 \neq f_3$, or it would indicate constant utility gap between two actions, contradicting with $\Delta > 0$. By symmetry, assume without loss of generality $f_1 - f_2 < 0$.

Case 1: Since $f_1 \neq f_2 \neq f_3$, there exists a strict ordering among them. Pick any ordering, say $f_1 < f_2 < f_3$. Then $u_f(\boldsymbol{x'}, j_2) < u_f(\boldsymbol{x'}, j_1)$, for any $\boldsymbol{x'}$ on the left of $\boldsymbol{x}$, and $u_f(\boldsymbol{x'}, j_2) < u_f(\boldsymbol{x'}, j_3)$, for any $\boldsymbol{x'}$ on the right of $\boldsymbol{x}$, which contradict with $\Delta > 0$.

Case 2: Note that $f_1 - f_2$ (or $f_1 - f_3$) captures the change of utility difference between j_1 and j_2 (or j_3). Since $f_1 \neq f_2 \neq f_3$, it is only possible that $(f_1 - f_2)(f_1 - f_3) > 0$ or $(f_1 - f_2)(f_1 - f_3) < 0$. If $(f_1 - f_2)(f_1 - f_3) < 0$, the maximum utility difference between j_1 and j_2, j_3 is no more than δ, which contradict with $\Delta > \delta$. If $(f_1 - f_2)(f_1 - f_3) > 0$ and $f_1 - f_2 < 0$ by assumption, we have $f_1 - f_3 < 0$. Pick any ordering between f_2, f_3, say $f_2 < f_3$. Then, j_2 can never be the follower's best response: for any leader strategy on the left of $\boldsymbol{x}$, the follower's utility of j_2, j_3 must be worse than that of j_1; for any leader strategy on the right of $\boldsymbol{x}$, the follower's utility of j_2 must be worse than that of j_3. This again contradicts $\Delta > 0$.

Case 3: Since $f_1 - f_2 < 0$, it is only possible that $f_1 - f_3 < 0$ (or $f_1 - f_3 > 0$ and $f_2 - f_3 > 0$). In this case, the follower's utility difference between j_3 and j_1 (or j_2) will always be less δ, and thus, $\mathcal{X}(\delta; \{j_1\})$ (or $\mathcal{X}(\delta; \{j_2\})$) will be empty, contradicting Property 1.

It now remains to show that even when the inequalities are not strict, the case $u_f(\boldsymbol{x}, j_1) > u_f(\boldsymbol{x}, j_2) > u_f(\boldsymbol{x}, j_3) > u_f(\boldsymbol{x}, j_1) - \delta$ is also impossible. Again since $f_1 - f_2 < 0$, it is only possible that $f_1 - f_3 < 0$, (or $f_1 - f_3 < 0$ and $f_2 - f_3 > 0$):

- If $f_1 - f_3 > 0$ and $f_2 - f_3 > 0$, then there exists a leader strategy $\boldsymbol{x'}$ on the left of $\boldsymbol{x}$ such that

$$u_f(\boldsymbol{x'}, j_1) \geq u_f(\boldsymbol{x'}, j_2) = u_f(\boldsymbol{x'}, j_3) > u_f(\boldsymbol{x'}, j_1) - \delta,$$

  which reduces to Case 1 or 2 above and reach the contradiction.

- If $f_1 - f_3 < 0$, then there exists a leader strategy $\boldsymbol{x'}$ on the right of $\boldsymbol{x}$ such that

$$u_f(\boldsymbol{x}, j_1) = u_f(\boldsymbol{x'}, j_2) \geq u_f(\boldsymbol{x'}, j_3) > u_f(\boldsymbol{x'}, j_1) - \delta,$$

  which reduces to Case 1 or 3 above and reach the contradiction.

Property 3 The first part of the property hinges on the observation that, when $m = 2$, for any $j, j' \in [n]$, if there exists $\boldsymbol{x}^{j,j'} \in \Delta^m$ such that $\text{BR}(\boldsymbol{x}^{j,j'}) = \{j, j'\}$, then $\boldsymbol{x}^{j,j'} \in \mathcal{X}(\delta; \{j, j'\})$ for any $\delta \in (0, \Delta)$. This is because $\boldsymbol{x}^{j,j'} \notin \mathcal{X}(\delta; S), \forall |S| = 1$ by definition, or $|S| > 2$ by Property 2. Given this observation, we prove by contradiction that $\mathcal{X}(\delta; \{j, j'\})$ must only contain points from the interior of $\mathcal{X}(0; j), \mathcal{X}(0; j')$, except for $\boldsymbol{x}^{j,j'}$. Suppose there exists another strategy $\boldsymbol{x} \in \mathcal{X}(\delta; \{j, j'\})$ such that $\text{BR}(\boldsymbol{x}) = \{j''\}$ and $j'' \neq j, j'$. Assume WLOG a left to right ordering of $\mathcal{X}(0; j), \mathcal{X}(0; \{j'\}), \mathcal{X}(0; \{j''\})$, i.e., the best response region of j, j', j'', respectively, in the 1-dimensional simplex. By convexity, this implies that $\mathcal{X}(\delta; \{j, j'\}) \supseteq \mathcal{X}(0; \{j'\} \supseteq \mathcal{X}(\delta; \{j'\})$. Since $\mathcal{X}(\delta; \{j, j'\}) \cap \mathcal{X}(\delta; \{j'\}) = \varnothing$, we get $\mathcal{X}(\delta; \{j'\}) = \varnothing$, contradicting Property 1. Using the same argument, $\mathcal{X}(\delta; \{j, j'\})$ cannot contain points from the boundary of $\mathcal{X}(0; j), \mathcal{X}(0; j')$, except for $\boldsymbol{x}^{j,j'}$. In addition, the two boundary point $\boldsymbol{x}^L, \boldsymbol{x}^R$ of $\mathcal{X}(\delta; \{j, j'\})$ must satisfy $u_f(\boldsymbol{x}^L, j) - u_f(\boldsymbol{x}^L, j') = \delta$ and $u_f(\boldsymbol{x}^R, j) - u_f(\boldsymbol{x}^R, j') = -\delta$ and thus $\boldsymbol{x}^L, \boldsymbol{x}^R \notin \mathcal{X}(\delta; \{j, j'\})$, $\mathcal{X}(\delta; \{j, j'\})$ is an open set. Suppose if $u_f(\boldsymbol{x}^L, j) - u_f(\boldsymbol{x}^L, j') = \delta'$. Clearly, if $\delta' > \delta$, the $\boldsymbol{x}^L$ must not be on the boundary of $\mathcal{X}(\delta; \{j, j'\})$ by definition. The case where $\delta' < \delta$ is also impossible, because this means $\boldsymbol{x}^L$ is on the left boundary of $\boldsymbol{x}$, or moving $\boldsymbol{x}^L$ leftward increases the utility gap δ'. Similar argument holds for $u_f(\boldsymbol{x}^R, j) - u_f(\boldsymbol{x}^R, j') = -\delta$.

The proof for the second part of the property now becomes straightforward. Consider $\delta' = \delta + \epsilon$ for any $\epsilon > 0, \delta' < \Delta$. We can verify by definition that $\mathcal{X}(\delta; \{j, j'\}) \subseteq \mathcal{X}(\delta'; \{j, j'\})$. That is, for any $\boldsymbol{x} \in \mathcal{X}(\delta; \{j, j'\})$, $|u_f(\boldsymbol{x}, j') - u_f(\boldsymbol{x}, j)| < \delta < \delta - \epsilon$. Moreover, observe that for the two points on the open boundary of $\mathcal{X}(\delta; \{j, j'\})$, we have $\boldsymbol{x}^R, \boldsymbol{x}^L \in \mathcal{X}(\delta + \epsilon; \{j, j'\})$. Therefore, $\mathcal{X}(\delta; \{j, j'\}) \subset \mathcal{X}(\delta'; \{j, j'\})$. □

B.1.6 Proof of Proposition 3.5

Proof. For the case when $m = 2$. Let $U(\delta; S) := \max_{\boldsymbol{x} \in \mathcal{X}(\delta; S)} \min_{j \in S} u_l(\boldsymbol{x}, j)$. Then, $u_{\text{RSE}}(\delta) = \max_{S \subseteq [n]} U(\delta; S) = \max_{S \subseteq [n], |S| \leq 2} U(\delta; S)$, by Lemma B.2. We first prove a

claim that if $U(\delta; S)$ is linearly for all $|S| = 1$, then $u_{\text{RSE}}(\delta)$ is convex at $\delta \in (0, \Delta)$.

From Lemma B.2, we know $U(\delta; S)$ is non-increasing in δ when $|S| = 1$, and is non-decreasing in δ when $|S| = 2$. This, along with Lemma B.1, implies that if for some $\delta_0 \in (0, \Delta)$, δ_0-RSE leader strategy $\boldsymbol{x}^* \in \mathcal{X}(\delta; S)$ with $|S| = 2$, then $\boldsymbol{x}^*$ is the δ-RSE for all $\delta \in [\delta_0, \Delta)$ — otherwise, it breaks the continuity of $u_{\text{RSE}}(\delta)$ when $\delta \in (0, \Delta)$. Hence, if such δ_0 exists, then we have

$$u_{\text{RSE}}(\delta) = \begin{cases} \max_{S \subseteq [n], |S|=1} U(\delta; S) & \delta \in (0, \delta_0) \\ u_{\text{RSE}}(\delta_0) & \delta \in [\delta_0, \Delta) \end{cases},$$

where $u_{\text{RSE}}(\delta)$ is convex non-increasing when $\delta \in (0, \delta_0)$ and is constant when $\delta \in [\delta_0, \Delta)$ — its gradient is non-decreasing and thus $u_{\text{RSE}}(\delta)$ is convex. Meanwhile, if such δ_0 does not exist, $u_{\text{RSE}}(\delta) = \max_{S \subseteq [n], |S|=1} U(\delta; S)$ is clearly convex.

It only remains to show that $U(\delta; S)$ is linearly for all $|S| = 1$. To see this, we can explicitly derive its gradient. Let $S = \{j_1\}$ and the leader utility gradient $g_1 = \frac{u_l((x,1-x),j_1)}{dx}$, follower utility gradient $f_i = \frac{u_f((x,1-x),j_i)}{dx}, \forall i \in [n]$. We assume WLOG $g_1 < 0$; otherwise, if $g_1 = 0$, $U(\delta; S)$ is a constant function and thus linear. The Lagrangian of $U(\delta; \{j_1\})$ is $L(\delta, x) = g_1 x - \sum_{i=2}^{n} \lambda_i [(f_1 - f_i)x - \delta]$, so its gradient function can be simplified using the envelope theorem as,

$$\frac{\partial U(\delta; \{j_1\})}{\partial \delta} = \frac{\partial L(\delta, x)}{\partial \delta} = \sum_{i=2}^{n} \lambda_i.$$

So it only remains to determine λ_i. First, if no constraint is tight, then $\sum_{i=2}^{n} \lambda_i = 0$ and $U(\delta; \{j_1\})$ is linear. Second, observe that, since the optimal solution x lies in a 1-dimensional space, there can be at most one constraint that is tight for any $\delta \in (0, \Delta)$, i.e., $u_f(\boldsymbol{x}, \{j_1\}) - u_f(\boldsymbol{x}, \{j_i\}) = \delta$ for some $j_i \neq j_1$. In this case, the gradient is constant $\frac{\partial U(\delta; \{j_1\})}{\partial \delta} = \lambda_i = \frac{g_1}{f_i - f_1}$ and $U(\delta; \{j_1\})$ is linear.

For the case when $m > 2$, we prove by construction. Consider a game with $m = 3, n = 2$

and utility matrices specified in Table B.5 where the parameter $\Delta \in (0, \frac{1}{2})$ and $c \in (0, 1)$.

u_l, u_f	j_1	j_2
i_1	$0, 1$	c, Δ
i_2	$1/2, 0$	c, Δ
i_3	$1, 1/2$	c, Δ

Table B.5: An instance where $u_{\text{RSE}}(\delta)$ is neither convex nor concave for $\Delta \in (0, \frac{1}{2}), c \in (0, 1)$.

We start with some observations on this game: 1) follower always have utility Δ for any leader strategy; 2) Since $c < 1$, the leader utility is maximized at the action profile (i_3, j_1). We now determine the δ-RSE at different interval of δ, see Figure B.2 for an illustration:

- For $\delta \in (0, \frac{1}{2} - \Delta]$, the δ-RSE leader strategy $\boldsymbol{x}^* = (0, 0, 1)$, as $\text{BR}_\delta(\boldsymbol{x}^*) = \{j_1\}$ and the leader receives utility 1 that are strictly better than any other possible leader/follower action profile.

- For $\delta \in (\frac{1}{2} - \Delta, 1 - \frac{c}{2} - \Delta]$, the δ-RSE leader strategy $\boldsymbol{x}^* = (\epsilon, 0, 1 - \epsilon)$, where $\epsilon = 2\Delta + 2\delta - 1$. We can verify that $\text{BR}_\delta(\boldsymbol{x}^*) = \{j_1\}$, since $\frac{1}{2}(1 - \epsilon) + \epsilon \geq \Delta + \delta$ and the leader receives utility $1 - \epsilon = 2 - 2\Delta - 2\delta$ that are no worse than c, which is the leader utility if the follower ever responds with j_2.

- For $\delta \in (1 - \frac{c}{2} - \Delta, \infty)$, the δ-RSE leader strategy $\boldsymbol{x}^* = (0, 1, 0)$. This is because leader cannot achieve a utility more than c, and thus the δ-RSE leader strategy $\boldsymbol{x}^*$ only needs to satisfy that $\text{BR}(\boldsymbol{x}^*) = \{j_2\}$.

Therefore, it is clear that $u_{\text{RSE}}(\delta)$ in this instance is neither convex nor concave.

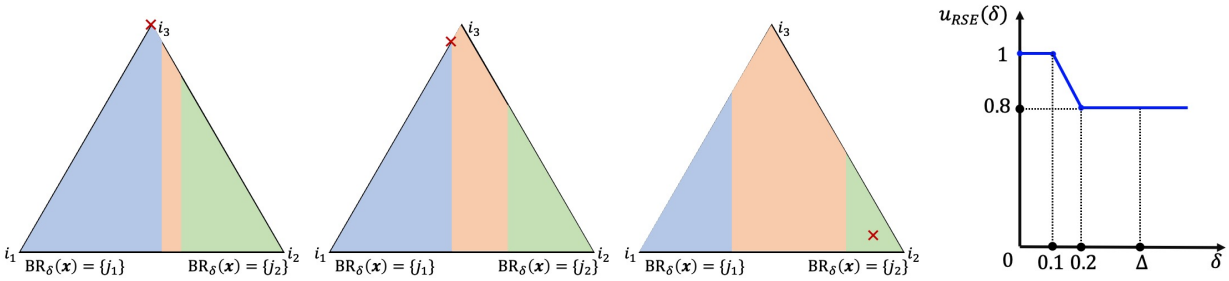


Figure B.2: An illustration of δ-RSE strategy $\boldsymbol{x}^*$ (highlighted with red marks) in the leader's strategy simplex and the plot of the function $u_{\mathrm{RSE}}(\delta)$, when $\Delta = 0.4, c = 0.8$. Each one of the first plots represents a linear segment of $u_{\mathrm{RSE}}(\delta)$.

B.2 Omitted Proofs in Section 3.4

B.2.1 Proof of Corollary 3.7

Proof. According to the proof of Property (1) in Theorem 3.4, we can construct a specific leader strategy $\widehat{\boldsymbol{x}} = (1 - \frac{\delta}{\Delta})\boldsymbol{x}^* + \frac{\delta}{\Delta}\boldsymbol{x}^{j^*}$ whose leader utility against a δ-rational follower is at least $(1 - \frac{\delta}{\Delta})u_{\mathrm{SSE}}$. Recall that $\langle \boldsymbol{x}^*, j^* \rangle$ is the SSE of the game, and $\boldsymbol{x}^{j^*}$ is the strategy such that $u_f(\boldsymbol{x}^{j^*}, j^*) \geq u_f(\boldsymbol{x}^{j^*}, j') + \Delta$ for all $j' \neq j^*$. It is easy to see that its utility $(1 - \frac{\delta}{\Delta})u_{\mathrm{SSE}} \geq (1 - \frac{\delta}{\Delta})u_{\mathrm{RSE}}(\delta) \geq u_{\mathrm{RSE}}(\delta) - \frac{\delta}{\Delta}$, which is $\frac{\delta}{\Delta}$-optimal for the δ-RSE leader utility.

It remains to show this construction is computationally efficient: it is well-known that SSE $(\boldsymbol{x}^*, j^*)$ can be computed by solving n linear programs [Conitzer and Sandholm, 2006]. What's more, finding $\boldsymbol{x}^{j^*}$ requires solving the following linear program:

$$\begin{aligned} \max_{\boldsymbol{x}} \quad & u_l(\boldsymbol{x}, j^*) \\ \text{subject to} \quad & u_f(\boldsymbol{x}, j^*) \geq u_f(\boldsymbol{x}, j') + \Delta, \forall j' \neq j^* \end{aligned} \tag{B.3}$$

Thus, constructing $\widehat{\boldsymbol{x}}$ requires solving $n + 1$ linear programs. □

B.3 Omitted Proofs in Section 3.5

B.3.1 Proof of Lemma 3.12

Let $\mathrm{BR}_\delta(\boldsymbol{x}, u_f)$ denote the set of δ-optimal response(s) of $\boldsymbol{x}$ for follower with utility function u_f. We start by showing that $\forall \boldsymbol{x} \in \Delta^m, \mathrm{BR}_{\delta+2\epsilon}(\boldsymbol{x}, \widetilde{u}_f) \supseteq \mathrm{BR}_\delta(\boldsymbol{x}, u_f)$, and it suffices to argue that for any $j \in [n]$, if $j \notin \mathrm{BR}_{\delta+2\epsilon}(\boldsymbol{x}, \widetilde{u}_f)$, then $j \notin \mathrm{BR}_\delta(\boldsymbol{x}, u_f)$. That is, given $j \notin \mathrm{BR}_{\delta+2\epsilon}(x, \widetilde{u}_f)$, we know there exists j' such that $\widetilde{u}_f(\boldsymbol{x}, j) - \widetilde{u}_f(\boldsymbol{x}, j') \leq -\delta - 2\epsilon$. Reusing the fact that for any $\|\widetilde{u} - u\|_\infty \leq \epsilon$, $\boldsymbol{x} \in \Delta^m$, $|\widetilde{u}(\boldsymbol{x}, j) - u(\boldsymbol{x}, j)| \leq \epsilon$, we have $u_f(\boldsymbol{x}, j) - u_f(\boldsymbol{x}, j') \leq \widetilde{u}_f(\boldsymbol{x}, j) - \widetilde{u}_f(\boldsymbol{x}, j') + 2\epsilon \leq -\delta$ and thus $j \notin \mathrm{BR}_\delta(\boldsymbol{x}, u_f)$.

Since $\mathrm{BR}_{\delta+2\epsilon}(\boldsymbol{x}, \widetilde{u}_f) \supseteq \mathrm{BR}_\delta(\boldsymbol{x}, u_f)$, we have by definition,

$$V(\boldsymbol{x}; u_l, u_f, \delta) = \min_{j \in \mathrm{BR}_\delta(\boldsymbol{x}, u_f)} u_l(\boldsymbol{x}, j) \geq \min_{j \in \mathrm{BR}_{\delta+2\epsilon}(\boldsymbol{x}, \widetilde{u}_f)} u_l(\boldsymbol{x}, j) = V(\boldsymbol{x}; u_l, \widetilde{u}_f, \delta + 2\epsilon).$$

This leads to the following inequalities,

$$V^*(u_l, \widetilde{u}_f, \delta+2\epsilon) = V(\boldsymbol{x}_1; u_l, \widetilde{u}_f, \delta+2\epsilon) \leq V(\boldsymbol{x}_1; u_l, u_f, \delta) \leq V(\boldsymbol{x}_2; u_l, u_f, \delta) = V^*(u_l, u_f, \delta),$$

where $\boldsymbol{x}_1$ is $(\delta + 2\epsilon)$-RSE strategy of Stackelberg game $u_l, \widetilde{u}_f$, and $\boldsymbol{x}_2$ is δ-RSE strategy of Stackelberg game u_l, u_f.

B.3.2 Proof of Proposition 3.14

Proof. We divide the proof into two parts: $\delta < \Delta$ or $\delta \geq \Delta$. The proofs for two parts are similar but require different instance construction.

$\Omega(1/\sqrt{T})$ lower bound instance when $\delta < \Delta$:

Table B.6 illustrates a game in which the inducibility gap is Δ. Suppose $\Delta > \delta$. Consider the following two Stackelberg games G_1, G_2 where $r_l(i, j) \sim \mathrm{Bern}(u_l(i, j))$ and $r_f(i, j) \sim \mathrm{Bern}(u_f(i, j))$, with mean values shown in Table B.6 and $\epsilon \in [0, 1]$ is a parameter to be

u_l, u_f	j_1	j_2
i_1	$1, \frac{1+\epsilon}{2} + \delta$	$0, \frac{1-\epsilon}{2}$
i_2	$0, 0$	$0, \Delta$
i_3	$0, \Delta$	$0, 0$

u_l, u_f	j_1	j_2
i_1	$1, \frac{1-\epsilon}{2} + \delta$	$0, \frac{1+\epsilon}{2}$
i_2	$0, 0$	$0, \Delta$
i_3	$0, \Delta$	$0, 0$

Table B.6: Example instances $G_1 = \{u_l, u_f\}$ (left) and $G_2 = \{u_l, u_f\}$ (right). Each table represents a Stackelberg game $u_l, u_f \in \mathbb{R}^{3\times 2}$ with inducibility gap Δ.

determined.

In Stackelberg game G_1, the leader strategy in δ-RSE can be computed as

$$\boldsymbol{x}_1^* = \operatorname*{argmax}_{\boldsymbol{x}\in\Delta^m} \min_{j\in \mathrm{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) = (1, 0, 0).$$

That is, in the δ-RSE of Stackelberg game G_1, the leader plays pure strategy i_1 and $\mathrm{BR}_\delta(i_1) = \{j_1\}$. As a result, we have $u_l^*(\delta) = 1$ for G_1.

On the other hand, for Stackelberg game G_2, note that $u_f(i_1, j_1) - u_f(i_1, j_2) = \delta - \epsilon < \delta$. Thus, j_2 is included in the follower's δ-optimal response set if the leader plays pure strategy i_1, i.e. the $\boldsymbol{x}_1^*$ for the other game G_1. Consequently, the leader receives 0 utility if the leader plays pure strategy i_1 in the Stackelberg game G_2. However, the leader can play a mixed strategy $\boldsymbol{x}_2^*$ that includes both pure actions i_1 and i_3 to increase the follower's utility difference between responding with j_1 and j_2. When the probability of playing i_3 is high enough, $\boldsymbol{x}_2^*$ can make $u_f(\boldsymbol{x}_2^*, j_1) - u_f(\boldsymbol{x}_2^*, j_2) = \delta$. Specifically, let $\boldsymbol{x}_2^* = (p, 0, 1-p)$, then we have the following constraint on p in order to exclude j_2 from the δ-optimal response set:

$$p \cdot \left(\frac{1-\epsilon}{2} + \delta\right) + (1-p)\cdot\Delta - p\cdot\frac{1+\epsilon}{2} = \delta \tag{B.4}$$

which makes $p = \frac{\Delta-\delta}{\Delta-\delta-\epsilon}$. As a result, for the Stackelberg game G_2, we have $\mathrm{BR}_\delta(\boldsymbol{x}_2^*) = \{j_1\}$ and $u_{\mathrm{RSE}}(\delta) = u_l(\boldsymbol{x}_2^*, j_1) = \frac{\Delta-\delta}{\Delta-\delta-\epsilon}$.

Therefore, if the learning algorithm cannot distinguish the above two Stackelberg games

with finite samples and makes a wrong estimation, the outputted leader strategy will suffer an error of at least $\frac{\epsilon}{\Delta-\delta+\epsilon}$. Specifically, the following incorrect estimations can happen:

1. If the true game is G_1, but the learning algorithm estimated the game as G_2 and outputs leader strategy as $\widehat{\boldsymbol{x}} = \boldsymbol{x}_2^*$. Then the leader utility of playing $\boldsymbol{x}_2^*$ will be $\frac{\Delta-\delta}{\Delta-\delta+\epsilon}$ while $u_{\text{RSE}}(\delta) = 1$, this has the learning error of $\frac{\epsilon}{\Delta-\delta+\epsilon}$.

2. If the true game is G_2, but the learning algorithm estimated the game as G_1 and outputs leader strategy as $\widehat{\boldsymbol{x}} = \boldsymbol{x}_1^*$. Then the leader utility of playing $\boldsymbol{x}_1^*$ will be 0 while $u_{\text{RSE}}(\delta) = \frac{\Delta-\delta}{\Delta-\delta+\epsilon}$, and this has the learning error of $\frac{\Delta-\delta}{\Delta-\delta+\epsilon}$ which is even larger than $\frac{\epsilon}{\Delta-\delta+\epsilon}$.

Consequently, if the learning algorithm can output a leader strategy $\widehat{\boldsymbol{x}}$ that violates equation (3.18), i.e.,

$$\min_{j \in \text{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) > u_{\text{RSE}}(\delta) - \frac{\epsilon}{\Delta - \delta + \epsilon},$$

then for the game G_1 we have $u_l(\widehat{\boldsymbol{x}}, j) > \frac{\Delta-\delta}{\Delta-\delta+\epsilon}$, while for the game G_2 we have $u_{\text{RSE}}(\delta) = \frac{\Delta-\delta}{\Delta-\delta+\epsilon}$. In other words, the learning algorithm can distinguish G_1 from G_2 with T samples.

Therefore, consider the learning algorithm that samples $r_f(i,j)$ for T times, and the goal is to identify if $u_f \in G_1$ or $u_f \in G_2$. We prove the proposition by contradiction. Suppose Proposition 3.14 is not correct, then with T samples, we can identify if $u_f \in G_1$ or $u_f \in G_2$ with probability more than $\frac{2}{3}$. Since $u_f(i_2, j_1)$, $u_f(i_2, j_2)$, $u_f(i_3, j_1)$, and $u_f(i_3, j_2)$ are the same for G_1 and G_2, the algorithm can only identify G_1 and G_2 by sampling $u_f(i_1, j_1)$ or $u_f(i_1, j_2)$. Formally, we define the problem as follows. Let $\Omega = [0,1]^T$ to be the sample space for the outcome of T samples of $r_f(i_1, j_1)$, our goal is to have the following decision rule

$$\text{Rule: } \Omega \to \{G_1, G_2\}, \tag{B.5}$$

u_l, u_f	j_1	j_2
i_1	$1, \frac{1+\epsilon}{2} + \delta$	$0, \frac{1-\epsilon}{2}$
i_2	$\frac{1}{2}, 1$	$\frac{1}{2}, 1$

u_l, u_f	j_1	j_2
i_1	$1, \frac{1-\epsilon}{2} + \delta$	$0, \frac{1+\epsilon}{2}$
i_2	$\frac{1}{2}, 1$	$\frac{1}{2}, 1$

Table B.7: Example instances $G_1 = \{u_l, u_f\}$ (left) and $G_2 = \{u_l, u_f\}$ (right). Each table represents a Stackelberg game $u_l, u_f \in \mathbb{R}^{2\times 2}$.

which satisfies the following two properties for any $\omega \in \Omega$:

$$\mathbf{Pr}[u_f \in G_1 | \text{Rule}(\omega) = G_1] > \frac{2}{3} \text{ and } \mathbf{Pr}[u_f \in G_2 | \text{Rule}(\omega) = G_2] > \frac{2}{3}. \tag{B.6}$$

As a result, let $\omega_o \in \Omega$ be the event this Rule returns G_1 (i.e., $\text{Rule}(\omega_o) = G_1$). Then we have:

$$\mathbf{Pr}[u_f \in G_1 | \omega_o] - \mathbf{Pr}[u_f \in G_2 | \omega_o] > \frac{1}{3} \tag{B.7}$$

Next, for any event $\omega \in \Omega$, let $P_k(\omega) = \mathbf{Pr}[u_f \in G_k | \omega]$ where $k = 1, 2$. Then we have $P_k = P_{k,1} \times \cdots \times P_{k,T}$ where $P_{k,t}$ is the distribution of t'th sample of $r_f(i_1, j_1)$. As a result, by applying a basic KL-divergence argument to distributions P_1 and P_2 [Slivkins et al., 2019, Lemma 2.5], for any event ω we have $|P_1(\omega) - P_2(\omega)| \leq \epsilon\sqrt{T}$. Plugging $\omega = \omega_0$ and $\epsilon = \frac{1}{3\sqrt{T}}$ we have $|P_1(\omega) - P_2(\omega)| \leq \frac{1}{3}$, contradicting with equation (B.7). Therefore, the Stackelberg game instances G_1 and G_2 in table B.6 with $\epsilon = \frac{1}{3\sqrt{T}}$ proves the proposition when $\delta < \Delta$.

$\Omega(1)$ **lower bound instance when** $\delta \geq \Delta$**:**

Consider the following two Stackelberg games G_1 and G_2, where $r_l(i, j) \sim \text{Bern}(u_l(i, j))$ and $r_f(i, j) \sim \text{Bern}(u_f(i, j))$, with mean values shown in Table B.7 and $\epsilon \in [0, 1]$ is a parameter to be determined. Note that in G_1 follower action j_2 is dominated by action j_1. Thus, we have $\Delta = 0$.

According to Table B.7, it is straightforward that the δ-RSE of Stackelberg game G_1 is $\langle i_1, j_1 \rangle$ where the leader plays pure strategy i_1 and follower responds with pure strategy j_1. Leader's utility at δ-RSE is $u_{\text{RSE}}(\delta) = 1$ for G_1. On the other hand, for Stackelberg game

G_2, we have $u_f(\boldsymbol{x}, j_1) - u_f(\boldsymbol{x}, j_2) < \delta$ for any leader strategy $\boldsymbol{x}$. Thus, j_2 is always in the follower's δ-optimal response set. To achieve the best utility, the leader plays pure strategy i_2 while the follower is indifferent between playing j_1 or j_2. As a result, $u_{\text{RSE}}(\delta) = \frac{1}{2}$ for G_2.

Suppose the learning algorithm can output leader strategy $\widehat{\boldsymbol{x}}$ that violates equation (3.18), i.e.,

$$\min_{j \in \text{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) > u_{\text{RSE}}(\delta) - 1/2.$$

Then for game G_1, we have $\min_{j \in \text{BR}_\delta(\widehat{\boldsymbol{x}})} u_l(\widehat{\boldsymbol{x}}, j) > \frac{1}{2}$ while

$$u_{\text{RSE}}(\delta) = \max_{\boldsymbol{x} \in \Delta^m} \min_{j \in \text{BR}_\delta(\boldsymbol{x})} u_l(\boldsymbol{x}, j) = \frac{1}{2}$$

for game G_2. Then this means the learning algorithm can identify G_1 from G_2.

Therefore, consider the learning algorithm that samples $r_f(i, j)$ for T times, and the goal is to identify if $u_f \in G_1$ or $u_f \in G_2$. We prove the proposition by contradiction. Suppose Proposition 3.14 is not correct, then with T samples, we can identify if $u_f \in G_1$ or $u_f \in G_2$ with probability more than $\frac{2}{3}$. Since $u_f(i_2, j_1)$ and $u_f(i_2, j_2)$ are the same for G_1 and G_2, the algorithm can only identify G_1 and G_2 by sampling $u_f(i_1, j_1)$ or $u_f(i_1, j_2)$. Formally, we define the problem as follows. Let $\Omega = [0, 1]^T$ to be the sample space for the outcome of T samples of $r_f(i_1, j_1)$ and our goal is to have the following decision rule

$$\text{Rule: } \Omega \to \{G_1, G_2\} \tag{B.8}$$

which satisfies the following two properties:

$$\begin{aligned} \mathbf{Pr}[u_f \in G_1 | \text{Rule(observations)} = G_1] > \frac{2}{3} \\ \mathbf{Pr}[u_f \in G_2 | \text{Rule(observations)} = G_2] > \frac{2}{3} \end{aligned} \tag{B.9}$$

Therefore, let $\omega_o \in \Omega$ be the event this Rule returns G_1. Then we have:

$$\mathbf{Pr}[u_f \in G_1|\omega_o] - \mathbf{Pr}[u_f \in G_2|\omega_o] > \frac{1}{3} \tag{B.10}$$

Next, for any event $\omega \in \Omega$, let $P_k(\omega) = \mathbf{Pr}[u_f \in G_k|\omega]$ where $k = 1, 2$. Then we have $P_k = P_{k,1} \times \cdots \times P_{k,T}$ where $P_{k,t}$ is the distribution of t'th sample of $r_f(i_1, j_1)$. As a result, by applying a basic KL-divergence argument to distributions P_1 and P_2 [Slivkins et al., 2019, Lemma 2.5], for any event ω we have $|P_1(\omega) - P_2(\omega)| \leq \epsilon\sqrt{T}$. Plugging $\omega = \omega_0$ and $\epsilon = \frac{1}{3\sqrt{T}}$ we have $|P_1(\omega) - P_2(\omega)| \leq \frac{1}{3}$, contradicting with equation (B.10).

Therefore, choosing the problem class to be G_1 and G_2 in table B.7 with $\epsilon = \frac{1}{3\sqrt{T}}$, finishes the proof.

□

APPENDIX C

C.1 Proof of Theorem 4.1

Definition 10 (Mixed Strategy Nash equilibrium)**.** *In a mixed strategy Nash equilibrium, each owner j adopts a mixed strategy x^j as a distribution supported on all possible rankings. We say a profile of owners' report $\{x^j\}_{j=1}^m$ forms a* mixed strategy *Bayes-Nash Equilibrium (NE) under mechanism $\mathcal{M}$ if for any owner $j \in [m]$, given others' randomized report profile $x^{-j} = \{x^{j'}\}_{j' \neq j}$, the expected utility under x^j is no worse than any other possible randomized report $\widetilde{x}^j$, i.e.,*

$$\mathop{\mathbf{E}}_{\boldsymbol{y}}\left[\mathop{\mathbf{E}}_{\pi^j \sim x^j}\mathop{\mathbf{E}}_{\pi^{-j} \sim x^{-j}} \mathrm{U}^j\left(\widehat{\mathrm{R}}_{\mathcal{M}}(x^j, x^{-j}; \boldsymbol{y})\right)\right] \geq \mathop{\mathbf{E}}_{\boldsymbol{y}}\left[\mathop{\mathbf{E}}_{\widetilde{\pi}^j \sim \widetilde{x}^j}\mathop{\mathbf{E}}_{\pi^{-j} \sim x^{-j}} \mathrm{U}^j\left(\widehat{\mathrm{R}}_{\mathcal{M}}(\widetilde{x}^j, x^{-j}; \boldsymbol{y})\right)\right].$$

Definition 11 (Majorization)**.** *We say $\boldsymbol{a}$ majorizes $\boldsymbol{b}$, denoted $\boldsymbol{a} \succeq \boldsymbol{b}$, if $\sum_{i=1}^k a_{(i)} \geq \sum_{i=1}^k b_{(i)}, \forall k < n$ and $\sum_{i=1}^n a_{(i)} = \sum_{i=1}^n b_{(i)}$, where $a_{(1)} \geq \cdots \geq a_{(n)}$ and $b_{(1)} \geq \cdots \geq b_{(n)}$ are sorted in descending order from $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively.*

Lemma C.1 (Hardy–Littlewood–Pólya inequality)**.** *For any vector $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, the inequality*

$$\sum_{i=1}^n h(a_i) \geq \sum_{i=1}^n h(b_i)$$

holds for all convex functions $h : \mathbb{R} \to \mathbb{R}$ if and only if $\boldsymbol{a} \succeq \boldsymbol{b}$.

Proof. Proof of Theorem 4.1 We first show that truthful reporting forms a Bayes-Nash equilibrium. Pick any owner and let its utility function be $\mathrm{U}(\widehat{\boldsymbol{R}}) = \sum_{i=1}^n U(\widehat{R}_i)$. Suppose that the other $m-1$ owners are already truth-telling and without loss of generality we can index them from 1 to $m-1$. With slight abuse of notation, we let $\widehat{\boldsymbol{R}}^j(\pi)$ be the random variable for the score estimates using owner j's reported ranking π w.r.t. the randomness of review noise $\boldsymbol{z}$, i.e, $\mathbf{Pr}[\widehat{\boldsymbol{R}}^j(\pi) = \widehat{\mathrm{R}}^j(\pi; \boldsymbol{R}+\boldsymbol{z})] = \mathbf{Pr}[\boldsymbol{z}]$. Suppose that the first $m-1$ estimates

use the true ranking and the true ranking $\pi^\star$ is the identity permutation for simplicity. It is clear that the item score estimates with all $m-1$ owners' truthfully reported rankings are all identical,

$$\widehat{\boldsymbol{R}}^1(\pi^\star) = \widehat{\boldsymbol{R}}^2(\pi^\star) = \cdots = \widehat{\boldsymbol{R}}^{m-1}(\pi^\star) := \widehat{\boldsymbol{R}}(\pi^\star).$$

Suppose the m-th owner reports π. Pick any $\{\alpha^j\}_{j=1}^m$ such that $\sum_{j=1}^m \alpha^j = 1$. Then, her expected utility is

$$\mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\mathrm{U}(\alpha^m \widehat{\boldsymbol{R}}^m(\pi) + \sum_{j=1}^{m-1} \alpha^j \widehat{\boldsymbol{R}}^j(\pi^\star))\right] = \mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\mathrm{U}(\alpha^m \widehat{\boldsymbol{R}}^m(\pi) + (1-\alpha^m)\widehat{\boldsymbol{R}}(\pi^\star))\right].$$

Since the noise terms are exchangeable in distribution, the expectation can be replaced by averaging the term above with $n!$ permuted noise terms. Let $\widehat{\boldsymbol{R}}(\pi; \boldsymbol{z}) := \widehat{\boldsymbol{R}}(\pi; \boldsymbol{R}+\boldsymbol{z})$ and $\rho \circ \boldsymbol{z}$ be the noise vector $\boldsymbol{z}$ after permutation ρ. By the linearity of expectation, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\mathrm{U}(\alpha^m \widehat{\boldsymbol{R}}^m(\pi) + (1-\alpha^m)\widehat{\boldsymbol{R}}(\pi^\star))\right] = \mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\sum_{\rho} \mathrm{U}\left(\alpha^m \widehat{\mathrm{R}}^m(\pi; \rho\circ\boldsymbol{z}) + (1-\alpha^m)\widehat{\mathrm{R}}(\pi^\star; \rho\circ\boldsymbol{z})\right)\right].$$

Note that $\widehat{\mathrm{R}}^m(\pi; \rho\circ\boldsymbol{z}) = \widehat{\mathrm{R}}(\pi; \rho\circ\boldsymbol{z})$ for any π. Then, it suffices to show that, for any vector $\boldsymbol{z} \in \mathbb{R}^n$,

$$\sum_{\rho} \mathrm{U}\left(\alpha^m \widehat{\mathrm{R}}(\pi; \rho\circ\boldsymbol{z}) + (1-\alpha^m)\widehat{\mathrm{R}}(\pi^\star; \rho\circ\boldsymbol{z})\right) \leq \sum_{\rho} \mathrm{U}\left(\alpha^m \widehat{\mathrm{R}}(\pi^\star; \rho\circ\boldsymbol{z}) + (1-\alpha^m)\widehat{\mathrm{R}}(\pi^\star; \rho\circ\boldsymbol{z})\right). \quad \text{(C.1)}$$

Let $\boldsymbol{a}^+$ be the projection of $\boldsymbol{a}$ onto the standard isotonic cone $\{\boldsymbol{a} | a_1 \geq a_2 \geq \cdots \geq a_n\}$. Recall that we assume without loss of generality that $R_1 \geq R_2 \cdots \geq R_n$. Following from the coupling argument in the proof of Theorem 1 in [Su, 2021], we have $\mathrm{R}(\pi^\star; \boldsymbol{z}) = (\boldsymbol{R}+\boldsymbol{z})^+$

and $\mathrm{R}(\pi; \boldsymbol{z}) = \pi^{-1} \circ (\pi \circ \boldsymbol{R} + \pi \circ \boldsymbol{z})^+, \forall \boldsymbol{z} \in \mathbb{R}^n$. So we can rewrite the inequality (C.1) as,

$$\sum_{\rho} \mathrm{U}\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})^+ + \alpha^m \pi^{-1} \circ (\pi \circ \boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+\right) \leq \sum_{\rho} \mathrm{U}\left((\boldsymbol{R} + \rho \circ \boldsymbol{z})^+\right).$$

We can prove this inequality with the following steps:

$$\begin{aligned}
&\sum_{\rho} \mathrm{U}\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})^+ + \alpha^m \pi^{-1} \circ (\pi \circ \boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+\right) \\
\leq &\sum_{\rho} \mathrm{U}\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})^+ + \alpha^m (\pi \circ \boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+\right) \\
\leq &\sum_{\rho} \mathrm{U}\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})^+ + \alpha^m (\boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+\right) \\
= &\sum_{i=1}^{n} \sum_{\rho} U\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})_i^+ + \alpha^m (\boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})_i^+\right) \\
\leq &\sum_{i=1}^{n} \sum_{\rho} U\left((1-\alpha^m)(\boldsymbol{R} + \rho \circ \boldsymbol{z})_i^+ + \alpha^m (\boldsymbol{R} + \rho \circ \boldsymbol{z})_i^+\right) \\
= &\sum_{\rho} \mathrm{U}\left((\boldsymbol{R} + \rho \circ \boldsymbol{z})^+\right)
\end{aligned}$$

The first inequality is due to Lemmas C.1 and C.2, as π^{-1} is some permutation on $\boldsymbol{b}^+ = (\pi \circ \boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+$. The second inequality follows from Lemmas C.1, C.3 as well as Lemma 2.4 of [Su, 2021], which shows that $(\boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+ \succeq (\pi \circ \boldsymbol{R} + \pi \circ \rho \circ \boldsymbol{z})^+, \forall \rho$. The third inequality follows from Lemma C.1 and C.2 where for each distinct permutation ρ_j, $(1-\alpha^m)(\boldsymbol{R} + \rho_j \circ \boldsymbol{z})_i^+, \alpha^m (\boldsymbol{R} + \rho_j \circ \boldsymbol{z})_i^+$ are the j-th entry of $\boldsymbol{a}^+, \boldsymbol{b}^+ \in \mathbb{R}^{n!}$ and π is some permutation on $\boldsymbol{b}^+$. The equalities are by the definition of the utility function and its linearity.

It now remains to prove that truthful NE gives every owner the highest equilibrium utility among all possible equilibria of the game. Again, we pick arbitrary owner j. Under truthful NE, the owner's utility is $\mathbf{E}_{\boldsymbol{z}}\left[\mathrm{U}^j(\widehat{\boldsymbol{R}}(\pi^\star))\right]$, where $\widehat{\boldsymbol{R}}(\pi^\star)$ is the item score estimate under the truthful ranking. We compare it to any (possibly mixed) Nash equilibrium

(see Definition 10). Consider any pure strategy profile in this mixed Nash equilibrium, $\boldsymbol{\pi} = (\pi^1, \pi^2, \ldots, \pi^m) \sim \boldsymbol{x} = (x^1, x^2, \ldots, x^m)$. This sampled profile of reported ranking corresponds to the weighted average of score estimates, $\sum_{j=1}^{m} \alpha^j \widehat{\boldsymbol{R}}(\pi^j)$. We can see that the owner's expected utility under this pure strategy profile is no larger than that under the truthful strategy profile,

$$\begin{aligned} \mathop{\mathbf{E}}_{\boldsymbol{z}} \left[\mathrm{U}^j \left(\sum_{j=1}^{m} \alpha^j \widehat{\boldsymbol{R}}(\pi^j) \right) \right] &\leq \sum_{j=1}^{m} \alpha^j \mathop{\mathbf{E}}_{\boldsymbol{z}} \left[\mathrm{U}^j (\widehat{\boldsymbol{R}}(\pi^j)) \right] \\ &\leq \max_{j \in [m]} \mathop{\mathbf{E}}_{\boldsymbol{z}} \left[\mathrm{U}^j (\widehat{\boldsymbol{R}}(\pi^j)) \right] \\ &\leq \mathop{\mathbf{E}}_{\boldsymbol{z}} \left[\mathrm{U}^j (\widehat{\boldsymbol{R}}(\pi^\star)) \right], \end{aligned} \tag{C.2}$$

where the first inequality is due to Jensen's inequality and linearity of expectation, the second inequality is by the property of arithmetic mean, last inequality is due to Theorem C.1.1. Hence, the owner's expected utility under this Nash equilibrium, taking an expectation over the distribution of pure strategy profile, $\mathbf{E}_{\boldsymbol{\pi} \sim \boldsymbol{x}} \mathbf{E}_{\boldsymbol{z}} \left[\mathrm{U}^j (\sum_{j=1}^{m} \alpha^j \widehat{\boldsymbol{R}}^j) \right] \leq \mathbf{E}_{\boldsymbol{z}} \left[\mathrm{U}^j (\widehat{\boldsymbol{R}}(\pi^\star)) \right]$, is no larger than the expected utility under the truthful NE. □

Remark C.1.1. *In fact, one may have noticed that the proof of payoff dominance above does not rely on whether the non-truthful strategy profile forms an equilibrium or not. Hence, a slightly stronger result holds, that is, the truthful NE gives every owner the highest utility among all possible strategy profiles of the game. However, we would like to also point out that this additional observation does not imply that the truthful report is necessarily a dominant strategy for any owner, i.e., the best response to any other strategy profile of the other owners. Nor is the truthful NE necessarily unique in this game. That said, given its payoff-dominant property, there is other strong evidence, especially from empirical human preference and behavioral theory, to expect the truthful equilibrium outcome, as pointed out in Remark 4.1.1.*

Theorem C.1.1 (Restatement of [Su, 2021, Theorem 1])**.** *Under Assumption 4.3, 4.4, 4.5,*

it is optimal for a single owner to truthfully report the ranking of its items $\pi^\star$,

$$\mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\mathrm{U}(\widehat{\boldsymbol{R}}(\pi^\star))\right] = \max_{\pi} \mathop{\mathbf{E}}_{\boldsymbol{z}}\left[\mathrm{U}(\widehat{\boldsymbol{R}}(\pi))\right].$$

Lemma C.2. *For any vector $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{a}^+ + \boldsymbol{b}^+ \succeq \boldsymbol{a}^+ + \rho \circ \boldsymbol{b}^+$ for any permutation ρ on $\boldsymbol{b}^+$.*

Proof. Proof of Lemma C.2 We prove it by contradiction. Suppose ρ maximizes the sum. Let $\boldsymbol{b'} = \rho \circ \boldsymbol{b}^+$. We have for some $i < j$, $b'_i < b'_j$. In this case, consider another vector $\widetilde{\boldsymbol{b}} = \widetilde{\rho} \circ \boldsymbol{b}^+$ with $\widetilde{b}_j = b'_i, \widetilde{b}_i = b'_j$ and $\widetilde{b}_\ell = b'_\ell, \forall \ell \neq i, j$. Given that $a_i^+ > a_j^+$, we have $a_i^+ + b'_j > a_j^+ + b'_i$. Hence, $\boldsymbol{a}^+ + \widetilde{\rho} \circ \boldsymbol{b}^+ \succeq \boldsymbol{a}^+ + \rho \circ \boldsymbol{b}^+$, which reaches the contradiction. □

Lemma C.3. *Suppose $\boldsymbol{a}, \boldsymbol{a'}, \boldsymbol{b}$ are in the same descending order and $\boldsymbol{a} \succeq \boldsymbol{a'}$, then $\boldsymbol{a} + \boldsymbol{b} \succeq \boldsymbol{a'} + \boldsymbol{b}$.*

Proof. Proof of Lemma C.3 We first verify that equality holds. Since $\sum_{i=1}^n a_i = \sum_{i=1}^n a'_i$,

$$\sum_{i=1}^{n}(a_i + b_i) = \sum_{i=1}^{n}(a'_i + b_i).$$

For the inequality at any $k < n$, we have $\sum_{i=1}^k a_i \geq \sum_{i=1}^k a'_i$. This readily implies that

$$\sum_{i=1}^{k}(a_i + b_i) \geq \sum_{i=1}^{k}(a'_i + b_i).$$

Therefore, by definition, we have $\boldsymbol{a} + \boldsymbol{b} \succeq \boldsymbol{a'} + \boldsymbol{b}$. □

C.2 Proof of Proposition 4.5

We prove that for any ownership instance $\mathcal{O} = \{\mathcal{I}^j\}_{j\in[m]}$, we can construct a different ownership instance $\mathcal{O}' = \{\overline{\mathcal{I}}^j\}_{j\in[m]}$ such that any L-strong partition $\mathcal{S}$ in $\mathcal{O}$ is a 1-strong partition in $\mathcal{O}'$.

Pick any bipartite graph $\mathcal{G} = (\mathcal{A}, \mathcal{P}, \mathcal{E})$ from the ownership instance $\mathcal{O} = \{\mathcal{I}^j\}_{j\in[m]}$. Suppose we were to find L-strong partition $\mathcal{S}$ in $\mathcal{G}$. We can construct a bipartite graph $\mathcal{G}' = (\mathcal{A}', \mathcal{P}', \mathcal{E}')$, where the item set remains the same $\mathcal{P}' = \mathcal{P}$, the owner set $\mathcal{A}'$ corresponds to all the L element subsets of $\mathcal{A}$, $\mathcal{A}' = \{\ell | u^\ell \subseteq \mathcal{A}, |u^\ell| = L\}$ and the edge set captures the commons items of every L owners, $\mathcal{E}' = \{(i, \ell) | i \in \bigcap_{j\in u^\ell} \mathcal{I}^j\}$. We show that for any L-strong partition $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K\}$ of $\mathcal{G}$, there is an equivalent 1-strong greedy partition $\mathcal{S}' = \{\mathcal{S}'_1, \mathcal{S}'_2, \ldots, \mathcal{S}'_K\}$ of $\mathcal{G}'$ such that $\mathcal{S}_k = \mathcal{S}'_k, \forall k \in K$.

The proof is a straightforward verification of this bipartite graph construction. Pick any $\mathcal{S}_k \in \mathcal{S}$ with $|\mathcal{T}_k| \geq L$ in $\mathcal{G}$, there exists an $\mathcal{S}'_k$ with $|\mathcal{T}'_k| = 1$ in $\mathcal{G}'$. That is, we have $\mathcal{T}'_k = \{\ell\}$ for some $u^\ell \subseteq \mathcal{T}_k$, then $\mathcal{S}'_k = \mathcal{S}_k \subseteq \bigcap_{j\in u^\ell} \mathcal{I}^j$ by construction. Therefore, $\mathcal{S}_k = \mathcal{S}'_k, \forall k \in [K]$.

C.3 Proof of Proposition 4.6

We consider the weight function $w(x) = \max\{x, 1\} - 1$. For any partition with K blocks, the objective function is $\text{obj}(\mathcal{S}) = n - K$. We show that finding the optimal partition under such an objective function is at least as hard as finding the minimum set. We prove via a standard hardness reduction argument: for any set cover instance, we can construct a bipartite graph instance in which the partition that maximizes the objective function is $\text{obj}(\mathcal{S}) = n - K$ can be turned into a minimum set cover.

Pick arbitrary set cover instance with m subsets of n item, $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m$. We construct a bipartite graph $\mathcal{G} = (\mathcal{A}, \mathcal{P}, \mathcal{E})$ where $\mathcal{A} = [m], \mathcal{P} = [n]$ and $\mathcal{E} = \{(p_i, a_j) | p_i \in \mathcal{V}_j, j \in [m]\}$. There exists an optimal partition $\mathcal{S}^*$ with $\text{obj}(\mathcal{S}^*) = n - K^*$, if and only if the minimum set cover has size K^*. For the "if" direction, we construct a partition $\mathcal{S}^*$ with $\text{obj}(\mathcal{S}^*) = n - K^*$ from the set cover of size K^*. For each subset selected by the set cover, we construct a partition block with items in the subset minus the items already in the partition. This partition must be valid and have K^* blocks. For the "only if" direction, we construct the set cover of size K^* from the optimal partition $\mathcal{S}^*$. For every block $\mathcal{S}^*_k$ of the optimal partition

$\mathcal{S}^*$, we pick any vertex $a_j \in \mathcal{T}_k^*$ from the corresponding owner set. Based on the K^* distinct vertices, we select the subsets with the same indices, and it forms a set cover of all n items by construction.

C.4 Proof of Theorem 4.3

Proof. Proof of Theorem 4.3 The proof is structured as follows. First, we characterize the necessary conditions under which Mechanism 4.3 is truthful, according to Lemma 4.4. Second, we show that for any Mechanism 4.3 under the necessary condition, a Mechanism 4.2 can be constructed to elicit as least as much ranking information.

Fix any ownership relation $\{\mathcal{I}^j\}_{j=1}^m$. For notational convenience, we let $\beta_i^j = 0, \forall i \notin \mathcal{I}^j$, as β_i^j does not affect the mechanism output anyway. Our characterization of truthful Mechanism 4.3 is based on the following lemma (restatement of Lemma), which shows that truthful condition holds for Mechanism 4.3 only if its parameters admit certain partition structure as detailed in Equation (C.3).

Lemma [Restatement of Lemma 4.4] *Under Assumptions 4.3, 4.4, and 4.5, if a Mechanism 4.3 with parameters* $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ *is truthful, then the following two conditions must both hold:*

(I) *Each owner has balanced influence on the items within each of its partition blocks for any input, i.e.,*

$$\text{for any } j \in [m], \mathcal{S} \in \mathfrak{S}^j, i, i' \in \mathcal{S}, \qquad \text{we have } \omega_i^j = \omega_{i'}^j,$$

where $\omega_i^j = \alpha^j \beta_i^j / \sum_{j' \in \mathcal{J}^i} \alpha^{j'} \beta_i^{j'}$ *denotes owner* j*'s relative influence on the output score of item* i*.*

(II) *The parameters* $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ *has a valid partition structure in the following sense,*

$$\text{for any } j, j' \in [m], \mathcal{S} \in \mathfrak{S}^j, \mathcal{S}' \in \mathfrak{S}^{j'}, \text{ if } \mathcal{S} \cap \mathcal{S}' \neq \varnothing, \text{ then } \beta_i^j = \beta_{i'}^j, \text{ for all } i, i' \in \mathcal{S} \cup \mathcal{S}'.^1 \tag{C.3}$$

Moreover, these two conditions are equivalent to each other.

We start by observing a few corollaries from Lemma 4.4, which will be useful for our proof of Theorem 4.3. First, Condition (II) of the lemma is applicable to the situation with $j = j'$, in which case $\mathcal{S} = \mathcal{S}'$ since other sets in $\mathfrak{S}^j$ does not overlap with $\mathcal{S}$. Hence we have $\beta_i^j = \beta_{i'}^j$ for each $i, i' \in \mathcal{S}$. Second, let us now consider the case $j \neq j'$ and let $\mathcal{S} \in \mathfrak{S}^j, \mathcal{S}' \in \mathfrak{S}^{j'}$ satisfying $\mathcal{S} \cap \mathcal{S}' \neq \varnothing$. Then the lemma implies $\beta_i^j = \beta_{i'}^j$, for all $i, i' \in \mathcal{S} \cup \mathcal{S}'$. A consequence of this is that either $\beta_i^j = 0$ (i.e., j's report does not affect item i), or j must own every item in $\mathcal{S} \cup \mathcal{S}'$ since j has a non-zero value for every item in $\mathcal{S} \cup \mathcal{S}'$. In the later case, if $\mathcal{S}' \not\subseteq \mathcal{S}$, then some element $\bar{i}$ in $\mathcal{S}' \setminus \mathcal{S}$ owned by j must belong to another set $\overline{\mathcal{S}} \in \mathfrak{S}^j$. Condition (II) of Lemma 4.4 then implies that the weights in $\overline{\mathcal{S}}, \mathcal{S}$ must all be the same, i.e., $\beta_i^j = \beta_{i'}^j$ for any $i \in \mathcal{S}$ and $i' \in \overline{\mathcal{S}}$.

Observations above hint at a constructive proof of the theorem by iteratively merging any two "unnecessarily isolated" partition blocks such as the $\mathcal{S}, \overline{\mathcal{S}}$ as above in the parameters $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$, until it becomes a Mechanism 4.2. Formally, we will show that for any Mechanism 4.3 with a valid partition structure specified in Equation (C.3), we can construct a Mechanism 4.2 with some item set partition that elicits no less ranking information and thus has no worse statistical efficiency.

Our construction hinges on a `MERGE` operator that merges all those "unnecessarily isolated" partition blocks like the $\mathcal{S}, \overline{\mathcal{S}}$ above. Specifically, if some owner $j \in [m]$ has two distinct blocks $\mathcal{S}, \overline{\mathcal{S}} \in \mathfrak{S}^j$ that satisfy the following conditions — for any $i \in \mathcal{S}, \bar{i} \in \overline{\mathcal{S}}$: (1) $\beta_i^j = \beta_{\bar{i}}^j$; and (2) $\beta_i^{j'} = \beta_{\bar{i}}^{j'} \neq 0$ for any other owner j' who fully owns all items in $\mathcal{S} \cup \overline{\mathcal{S}}$ — then `MERGE`

1. We remark that β_i^j is set to 0 for any $i \notin \mathcal{I}^j$ in this statement.

combines $\mathcal{S}, \overline{\mathcal{S}}$ as one set in j's $\mathfrak{S}^j$ with the same parameter β_i^j for any $i \in \mathcal{S} \cup \overline{\mathcal{S}}$. Notice that `MERGE` strictly increases the amount of ranking information elicited from the owners, as the orderings of items among merged blocks $\mathcal{S}, \overline{\mathcal{S}}$ are now used in the calibration.

We exhaustively apply `MERGE` on the parameters for every owner's every pair of item sets until there is no more valid merge, and then remove any block $\mathcal{S}$ with zero weights, i.e., $\beta_i^j = 0, \forall i \in \mathcal{S}$. The last procedure does not affect the mechanism but makes our following analysis cleaner. It is also obvious that the merge process must terminate, since the merged block can only be as large as each owner's item set. In addition, after each merge, the value of any $\boldsymbol{\beta}^j$ remains unchanged; it can be verified that the condition in Equation (C.3) holds for the updated partition blocks. Let the resulting parameters be $\left\{\widetilde{\mathfrak{S}}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$. We claim that $\left\{\widetilde{\mathfrak{S}}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ satisfies that, for any $j, j' \in [m]$, $\mathcal{S} \in \widetilde{\mathfrak{S}}^j, \mathcal{S}' \in \widetilde{\mathfrak{S}}^{j'}$, $\mathcal{S}, \mathcal{S}'$ are either identical or disjoint.

We prove by contradiction: if this is not the case, then there must exist at least one valid merge. Suppose $\mathcal{S} \cap \mathcal{S}' \neq \varnothing$. Since $\left\{\widetilde{\mathfrak{S}}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ satisfies the condition in Equation (C.3), we have shown above that for owner j, she must (1) either have $\beta_i^j = 0, \forall i \in \mathcal{S} \cup \mathcal{S}'$, (2) or j must own every item in $\mathcal{S} \cup \mathcal{S}'$ since j has a non-zero value for every item in $\mathcal{S} \cup \mathcal{S}'$. The first case is eliminated as blocks with zero weights are removed. In the second case, some element $\bar{i}$ in $\mathcal{S}' \setminus \mathcal{S}$ owned by j must belong to another set $\overline{\mathcal{S}} \in \mathfrak{S}^j$, and $\mathcal{S}, \overline{\mathcal{S}}$ can be merged, which would reach the contradiction. To check that the two condition for a valid *merge* is satisfied, we can observe the following, according to the condition in Equation (C.3):

1. For this owner j, we have $\beta_i^j = \beta_{i'}^j = \beta_{\bar{i}}^j$ for any $\bar{i} \in \overline{\mathcal{S}} \setminus \mathcal{S}'$. Hence, $\beta_i^j = \beta_{i'}^j, \forall i, i' \in \mathcal{S} \cup \overline{\mathcal{S}}$.

2. For any other owner j' who owns all items in $\mathcal{S} \cup \overline{\mathcal{S}}$, we pick two items $i_1 \in \mathcal{S} \cap \mathcal{S}'$ and $i_2 \in (\mathcal{S}' \setminus \mathcal{S}) \cap \overline{\mathcal{S}}$. Let $i_1 \in \mathcal{S}_1$ and $i_2 \in \mathcal{S}_2$ from $\mathfrak{S}^{j'}$. Since $\mathcal{S}_1 \cap \mathcal{S} \neq \varnothing$, $\beta_{i_1}^{j'} = \beta_i^{j'}, \forall i \in \mathcal{S}$. Since $\mathcal{S}_2 \cap \overline{\mathcal{S}} \neq \varnothing$, $\beta_{i_2}^{j'} = \beta_i^{j'}, \forall i \in \overline{\mathcal{S}}$. Since $\mathcal{S}_1 \cap \mathcal{S} \neq \varnothing$ and $i_1, i_2 \in \mathcal{S}$, $\beta_{i_1}^{j'} = \beta_{i_2}^{j'}$. Hence, $\beta_i^{j'} = \beta_{i'}^{j'} = \beta_{\bar{i}}^{j'}, \forall i, i' \in \overline{\mathcal{S}} \cup \mathcal{S}$.

Finally, consider $\mathfrak{S} := \bigcup_{j=1}^m \widetilde{\mathfrak{S}}^j$ (after removing repeated item sets). One can verify that

any two blocks $\mathcal{S}, \mathcal{S}' \in \mathfrak{S}$ are disjoint, though the union of all its blocks is not necessarily $[n]$ (recall that zero weight blocks have been removed). We claim that Mechanism 4.2 with parameter $\mathfrak{S}$ is at least as statistically efficient as Mechanism 4.3 with parameter $\left\{\widetilde{\mathfrak{S}}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$, and thus, Mechanism 4.3 with parameter $\left\{\mathfrak{S}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ before `MERGE`. To see this, let us first observe that Mechanism 4.3 under $\left\{\widetilde{\mathfrak{S}}^j, \boldsymbol{\beta}^j\right\}_{j=1}^m$ proceeds in the following steps: (1) pick each distinct block $\mathcal{S} \in \mathfrak{S} = \bigcup_{j=1}^m \widetilde{\mathfrak{S}}^j$, (2) elicit rankings of items in $\mathcal{S}$ from any owner j such that $\mathcal{S} \in \widetilde{\mathfrak{S}}^j$, (3) determine a weighted average of adjusted review scores based on the rank-calibrated score from those owners with weight $\alpha^j \beta_i^j$. These procedures are identical to Mechanism 4.2 with parameter $\mathfrak{S}$, except that Mechanism 4.2 elicits from all owners with complete ownership of a block $\mathcal{S} \in \mathfrak{S}$. This completes the proof of our theorem.

□

Proof. Proof of Lemma 4.4

We will prove that the truthfulness of Mechanism 4.3 implies condition (I), and condition (I) implies condition (II). Finally, we will prove condition (II) also implies condition (I), showing their equivalence.

We start by proving the first step, i.e., truthfulness of Mechanism 4.3 implies condition (I). Prove by contradiction. Suppose that for some owner $j \in [m]$ and for one of her partition blocks $\mathcal{S} \in \mathfrak{S}^j$, there exists $i, i' \in \mathcal{S}$ such that $\omega_i^j \neq \omega_{i'}^j$. Without loss of generality, let $\omega_i^j < \omega_{i'}^j$. We construct an instance with owner utility $\{\mathrm{U}^j\}_{j=1}^m$, review scores and ground-truth scores $\{y_i, R_i\}_{i=1}^n$ such that the unbalanced influence would give the owner j incentive to untruthfully report the pairwise order between i, i'. Let $y_i = R_i > R_{i'} = y_{i'}$. For any other item in the block $\mathcal{S}$, let its review score and ground-truth score be equal, and be smaller than the ground-truth scores of i, i' — it is now possible to only misreport the ranking between i, i', since the pairwise order of i, i' is independent of other items in this block. In addition, since the owner's report strategy in block $\mathcal{S}$ does not affect the items outside block $\mathcal{S}$, it suffices to compare the owner's report strategy in block $\mathcal{S}$. Let the utility of owner j be

$U^j(\widehat{R}) = \max\{\widehat{R} - c, 0\}$ for some constant c to be specified below. The utility function is convex w.r.t. the item's adjusted review score $\widehat{R}$, conforming to Assumption 4.5.

We now compare two reporting strategy of this owner: one is to truthfully report the ground-truth ranking, the other is to report that item i' is better than i, and the ground-truthful ranking for the rest of the items in this block. It suffices to show that the owner has strictly higher utility in the latter reporting strategy. Recall that in this problem instance, the review scores and ground-truth scores of the other items are equal, and are smaller than the ground-truth scores of i, i'. Isotonic regression implies that the adjusted score of all other items remain the same as the ground-truth score in both reporting strategy, so it suffices to analyze adjusted score of i, i' and compare the owner j's utility on these two items.

If the owner j is truthful, the adjusted score is $\widehat{R}_i = R_i, \widehat{R}_{i'} = R_{i'}$. If the owner j chooses the misreporting strategy, the calibrated scores based on the owner j's ranking are $\widetilde{R}^j_i = \widetilde{R}^j_{i'} = \frac{1}{2}(R_{i'} + R_i)$, resulting in the final adjusted scores, $\widetilde{R}_i = \frac{1}{2}\omega^j_i(R_{i'} + R_i) + (1 - \omega^j_i)R_i$ and $\widetilde{R}_{i'} = \frac{1}{2}\omega^j_{i'}(R_{i'} + R_i) + (1 - \omega^j_{i'})R_{i'}$. Let $\varepsilon = \frac{R_i - R_{i'}}{2}$; it can be verified that $\widetilde{R}_i = R_i - \omega^j_i\varepsilon$ and $\widetilde{R}_{i'} = R_{i'} + \omega^j_{i'}\varepsilon$. Notice that since $0 \le \omega^j_i < \omega^j_{i'} \le 1$, we have $R_i \ge \widetilde{R}_i > \widetilde{R}_{i'} > R_{i'}$ and $\omega^j_{i'}\varepsilon - \omega^j_i\varepsilon > 0$. Pick any $c \in [R_{i'}, R_{i'} + \omega^j_{i'}\varepsilon - \omega^j_i\varepsilon)$, we can derive the owner j's utility for the two items under the misreport and under the truthful report, respectively:

$$\begin{aligned}
U^j(\widetilde{R}_i) + U^j(\widetilde{R}_{i'}) &= \max\{R_i - \omega^j_i\varepsilon - c, 0\} + \max\{R_{i'} + \omega^j_{i'}\varepsilon - c, 0\} \\
&= R_i - \omega^j_i\varepsilon - c + R_{i'} + \omega^j_{i'}\varepsilon - c \\
U^j(\widehat{R}_i) + U^j(\widehat{R}_{i'}) &= \max\{R_i - c, 0\} + \max\{R_{i'} - c, 0\} \\
&= R_i - c
\end{aligned}$$

We can see that $U^j(\widetilde{R}_i) + U^j(\widetilde{R}_{i'}) - U^j(\widehat{R}_i) - U^j(\widehat{R}_{i'}) = R_{i'} + \omega^j_{i'}\varepsilon - \omega^j_i\varepsilon - c > 0$ (by our choice of c), so misreporting the ordering between j and j' is strictly better than truthful report.

We now prove that the condition (I) and (II) are equivalent and, therefore, are both necessary condition for Mechanism 4.3 to be truthful.

(I)⇒(II): We prove by contradiction. Suppose that there exists $j, j' \in [m], \mathcal{S} \in \mathfrak{S}^j, \mathcal{S}' \in \mathfrak{S}^{j'}$ such that $\mathcal{S} \cap \mathcal{S}' \neq \varnothing$ and there exists $i_1, i_2 \in \mathcal{S} \cup \mathcal{S}', \beta^j_{i_1} \neq \beta^j_{i_2}$. We consider two possible cases:

1. If $i_1, i_2 \in \mathcal{S}$, then we can construct the problem instance with $\alpha^j = 1$ being the only non-zero weight such that $\omega^j_{i_1} = \beta^j_{i_1} \neq \beta^j_{i_2} = \omega^j_{i_2}$, which is a contradiction to the given condition. Similar contradiction arises when $i_1, i_2 \in \mathcal{S}'$.

2. Otherwise, let $i_1 \in \mathcal{S}$ and $i_2 \in \mathcal{S}'$ without loss of generality. Since $\mathcal{S} \cap \mathcal{S}' \neq \varnothing$, let $i_0 \in \mathcal{S} \cap \mathcal{S}'$. Note that the argument above already shows items within each block must have the same β^j_i values, i.e., $\beta^j_{i_0} = \beta^j_{i_1}$ and $\beta^{j'}_{i_0} = \beta^{j'}_{i_2}$. Consequently, we have $\beta^j_{i_0} = \beta^j_{i_1} \neq \beta^j_{i_2}$ and $\beta^{j'}_{i_0} = \beta^{j'}_{i_2}$. We consider a problem instance with $\alpha^j = \alpha^{j'} = 1$ being only non-zero weight, thus $w^{j'}_{i_0} = \frac{\beta^{j'}_{i_0}}{\beta^j_{i_0} + \beta^{j'}_{i_0}}$ and $w^{j'}_{i_2} = \frac{\beta^{j'}_{i_2}}{\beta^j_{i_2} + \beta^{j'}_{i_2}}$. Since $i_0, i_2 \in \mathcal{S}'$, we have $\omega^{j'}_{i_0} \neq \omega^{j'}_{i_2}$ because $\beta^{j'}_{i_0} = \beta^{j'}_{i_2}$ but $\beta^j_{i_0} \neq \beta^j_{i_2}$. Contradiction is reached.

(II)⇒(I): We consider arbitrary owner $j \in [m]$ and any partition block $\mathcal{S} \in \mathfrak{S}^j$. Without loss of generality, suppose $|\mathcal{S}| \geq 2$ since singleton set does not have item rankings. Pick any two distinct items $i, i' \in \mathcal{S}$ in the block. We make the following observations from the condition in Equation (C.3):

1. $\beta^j_i = \beta^j_{i'}$ for the owner j herself. This is implied by instantiating Equation (C.3) to $j = j'$, in which case $\mathcal{S} = \mathcal{S}'$ since other sets in $\mathfrak{S}^j$ does not overlap with $\mathcal{S}$. The condition applies to any two items in $\mathcal{S}$.

2. $\beta^{j'}_i = \beta^{j'}_{i'}$, for any owner $j'(\neq j)$ who own any of i, i', i.e., $j' \in \mathcal{J}^i \cup \mathcal{J}^{i'}$. This is implied by instantiating Equation (C.3) to the j', j and some $\mathcal{S}' \in \mathfrak{S}^{j'}$ such that $\mathcal{S}' \cap \mathcal{S} \neq \varnothing$. Any items in $\mathcal{S}' \cup \mathcal{S}$ such as i, i' must have the same β values, i.e., $\beta^{j'}_i = \beta^{j'}_{i'}$.

3. $\beta_i^{j'} = \beta_{i'}^{j'} = 0$, for any owner j' who own neither of i, i', i.e., $j' \notin \mathcal{J}^i \cup \mathcal{J}^{i'}$. This follows from the mechanism construction where we set $\beta_i^{j'} = 0$ for any $i \notin \mathcal{I}^{j'}$.

The three cases above implies that, for any input credentials $\{\alpha^j\}_{j=1}^m$, we must always have $\omega_i^j = \omega_{i'}^j$ by definition for any two items i, i' in any of j's partition block $\mathcal{S}$, which proves the condition (I).

□

C.5 Proof of Theorem 4.7

A Convenient Bipartite Graph View of the Ownership Relation. To ease our discussion in the formal proofs regarding the optimal partition, we start by representing the overlapping ownership model as a bipartite graph $\mathcal{G} = (\mathcal{P}, \mathcal{A}, \mathcal{E})$. The vertex sets $\mathcal{P}, \mathcal{A}$, with $|\mathcal{P}| = n, |\mathcal{A}| = m$ correspond to the disjoint set of n items (papers) and m owners (authors), respectively. The edge set $\mathcal{E} \subseteq \mathcal{P} \times \mathcal{A}$ corresponds to the ownership relation: the owner j owns the item i, if and only if there is an edge $(p_i, a_j) \in \mathcal{E}$ between the vertices $p_i \in \mathcal{P}$ and $a_j \in \mathcal{A}$. The number of edges $|E| = \sum_{j \in [m]} |\mathcal{I}^j|$ is precisely the N in the theorem statement, which captures the order of the problem's input size. Let $E = (e_i^j)_{m \times n}$ be the $0-1$ bi-adjacency matrix of the ownership in which $e_i^j = 1$ for each edge $(p_i, a_j) \in \mathcal{E}$. Hence, the set of items owned by j is exactly the neighbor set of vertex a_j, i.e., $\mathcal{I}^j = \{p_i \in \mathcal{P} | e_i^j = 1\}$.

We now formalize the partition-based scheme in Mechanism 4.2 using the bipartite graph characterization. Specifically, we consider the partition $\mathfrak{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ on the vertex set $\mathcal{P}$ of the bipartite graph. By definition of a partition, we have $\bigcup_{k=1}^K \mathcal{S}_k = \mathcal{P}$ and $\mathcal{S}_{k'} \cap \mathcal{S}_k = \varnothing, \forall k \neq k' \in [K]$. For each partition block $\mathcal{S}_k$, we denote $\mathcal{T}_k = \{j \in \mathcal{A} | \mathcal{S}_k \subseteq \mathcal{I}^j\}$ as the set of owners who owns all items in $\mathcal{S}_k$. For the blocks with $|\mathcal{S}_k| \leq 1$ or $|\mathcal{T}_k| \leq 0$, the mechanism simply uses the raw review score. Otherwise, by construction, the sub-bipartite-graph of $\mathcal{G}$ induced by each $(\mathcal{S}_k, \mathcal{T}_k)$ and their incidence edges $\{(p_i, a_j) \in \mathcal{E} | p_i \in \mathcal{S}_k, a_j \in \mathcal{T}_k\}$ forms a complete bipartite graph, i.e., a complete overlapping ownership. Therefore, we

can also visualize the partition of the partially overlapping ownership using the bi-adjacency matrix.

We now utilize the notations above to formally prove Theorem 4.7.

Proof. Proof of Theorem 4.7 Pick any ownership instance represented as a bipartite graph $\mathcal{G}$ and any wellness function $w \in \mathcal{W}$. Let $\mathrm{OPT}(\mathcal{G}) := \mathrm{obj}(\mathfrak{S}^*)$, where $\mathfrak{S}^* = \{\mathcal{S}_k^*\}_{k=1}^K$ is the optimal 1-strong partition of $\mathcal{G}$. Consider the greedy algorithm that keeps picking the largest residual item set owned by some author, until all items are assigned to some block, as described in Algorithm 4.4. Let $\mathfrak{S}$ denote the partition obtained by Algorithm 4.4. By construction, $\mathfrak{S}$ is 1-strong — i.e., there is at least one owner who owns all items of each block in $\mathfrak{S}$. We let $\mathrm{ALG}(\mathcal{G}) := \mathrm{obj}(\mathfrak{S})$ be the objective value of the partition obtained by Algorithm 4.4. The remainder of the proof will argue that the output $\mathfrak{S}$ simultaneously satisfies,

$$\frac{\mathrm{ALG}(\mathcal{G})}{\mathrm{OPT}(\mathcal{G})} = \frac{\mathrm{obj}(\mathfrak{S})}{\mathrm{obj}(\mathfrak{S}^*)} \geq \inf\big\{\frac{w(x)}{w'_-(x)x}\big| w'_-(x) > 0, x \geq 2\big\},$$

for every $w \in \mathcal{W}$.

Let $\mathcal{G} \setminus j$ denote the induced subgraph of $\mathcal{G}$ by removing the vertex j as well as all its adjacent item nodes and corresponding edges from $\mathcal{G}$. Formally, in this induced subgraph $\mathcal{G} \setminus j = (\mathcal{P}', \mathcal{A}', \mathcal{E}')$, we have $\mathcal{P}' = \mathcal{P} \setminus \mathcal{I}^j, \mathcal{A}' = \mathcal{A} \setminus \{j\}$ and $\mathcal{E}' = \{(i,j)|(i,j) \in \mathcal{E} \text{ and } i \in \mathcal{P}', j \in \mathcal{A}'\}$.

We now prove the approximation ratio of the greedy algorithm by induction on the owner set's size $m = |\mathcal{A}|$. For the base case, for any bipartite graph $\mathcal{G} = (\mathcal{P}, \mathcal{A}, \mathcal{E})$ with $|\mathcal{A}| = 1$, it is clear that the greedy approach finds the exact optimal partition. For the inductive case, we show that if the approximation guarantee holds for any bipartite graph $\mathcal{G} = (\mathcal{P}, \mathcal{A}, \mathcal{E})$ with $|\mathcal{A}| \leq k$, then it must hold for any bipartite graph $\mathcal{G}$ with $|\mathcal{A}| = k + 1$ as well.

Let j be the very first owner picked by Algorithm 4.4 on the bipartite graph instance $\mathcal{G}$ with $|\mathcal{A}| = k + 1$. Thus, the first block (i.e., item set) picked by the algorithm must be $\mathcal{I}^j$. Let $\mathfrak{S}^*$ be the optimal partition of $\mathcal{G}$ under weight w. Note that the algorithm output $\mathfrak{S}$ is

independent of w, but $\mathfrak{S}^*$ typically depends on w. For each block $\mathcal{S}_k^* \in \mathfrak{S}^*$, let $p_k^* := \left|\mathcal{S}_k^*\right|$ be the number of items in this block and $a_k := \left|\mathcal{S}_k^* \cap \mathcal{I}^j\right|$ be the number of items covered by the greedy algorithm's first choice $\mathcal{I}^j$. Hence, $\left|\mathcal{I}^j\right| = \sum_{k=1}^K a_k$. We can then bound the approximation ratio of the greedy algorithm by decomposing the objective value as follows,

$$\begin{aligned}\frac{\mathrm{ALG}(\mathcal{G})}{\mathrm{OPT}(\mathcal{G})} =& \frac{w(\mathcal{I}^j) + \mathrm{ALG}(\mathcal{G} \setminus j)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k) + \sum_{k=1}^K w(p_k^* - a_k)} \\ \geq& \frac{w(\mathcal{I}^j) + \mathrm{ALG}(\mathcal{G} \setminus j)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k) + \mathrm{OPT}(\mathcal{G} \setminus j)},\end{aligned}$$

where the inequality uses the suboptimality of the partition $\{\mathcal{S}_k^* \setminus \mathcal{I}^j\}_{k=1}^K$ for the graph $\mathcal{G} \setminus j$.

We will bound the approximation ratio by considering the two parts $\frac{w\left(\sum_{k=1}^K a_k\right)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k)}$ and $\frac{\mathrm{ALG}(\mathcal{G} \setminus j)}{\mathrm{OPT}(\mathcal{G} \setminus j)}$ separately, and then apply the mediant inequality to derive the lower bound. The second part can be easily bounded according to the induction hypothesis. However, the first part could be unbounded in general (this may happen when p_k^*'s are all at most 1). Fortunately, the analysis for such special case can be carried out separately. Specifically, if $\max_{k \in [K]} p_k^* \leq 1$ in $\mathcal{G}$, by convexity of w function, this implies that any 1-strong partition of the items are at least as good as $\mathfrak{S}^*$, thus the greedy algorithm is optimal.

Next, we consider the more typical situation with $p := \max_{k \in [K]} p_k^* \geq 2$. Here, invoking the mediant inequality and a technical Lemma C.4 below with $p \geq 2$, we can lower bound the approximation ratio as follows,

$$\begin{aligned}&\frac{w(\mathcal{I}^j) + \mathrm{ALG}(\mathcal{G} \setminus j)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k) + \mathrm{OPT}(\mathcal{G} \setminus j)} \\ \geq& \min\left\{\frac{w\left(\sum_{k=1}^K a_k\right)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k)}, \frac{\mathrm{ALG}(\mathcal{G} \setminus j)}{\mathrm{OPT}(\mathcal{G} \setminus j)}\right\} \\ \geq& \inf\left\{\frac{w(x)}{w'_-(x)x} \middle| w'_-(x) > 0, x \geq 2\right\}.\end{aligned}$$

Finally, for running time analysis, Algorithm 4.4 can be efficiently implemented in almost

linear time using appropriately chosen data structures. Specifically, we employ the red-black tree data structure which can retrieve the maximum element $j^* = \arg\max_{j\in[m]} |\mathcal{I}^j \setminus \overline{\mathcal{I}}|$, i.e., the maximal residual paper sets among all authors, using $O(\log m)$ time. We will then update the paper sets of all the authors who have at least one paper in $\mathcal{I}^{j^*} \setminus \overline{\mathcal{I}}$. The number of operations is at most the number of edges connected to removed items from j^* in the bipartite graph. We then update the red-black tree structure using the new sizes of the residual paper sets of the involved papers. Note that the update of each author's paper set size takes $O(\log m)$ time by first deleting it in the red-black tree and then re-inserting it. The above recursion takes at most m rounds since each round at least one author's residual paper set will become empty and there are only m authors. In total, the total running time is thus $O(|\mathcal{E}| \log m)$.

□

Lemma C.4. *Consider any $\{p_k^*\}_{k\in[K]}$ and $\{a_k\}_{k\in[K]}$ such that $p_k^* \geq a_k \geq 0$ for any $k \in [K]$ and $\sum_{k=1}^K a_k \geq p := \max_{k\in[K]} p_k^*$. For any $w \in \mathcal{W}$ with $w(p) > 0$, we have*

$$\frac{w\left(\sum_{k=1}^K a_k\right)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k)} \geq \inf_{x\geq p} \frac{w(x)}{w'_-(x)x}.$$

Proof. Proof of Lemma C.4 Any $w \in \mathcal{W}$ is convex, so its left derivative w'_- is non-deceasing such that $w(p_k^*) - w(p_k^* - a_k) \leq w'_-(p_k^*)a_k$. In addition, $w'_-(p_k^*) \geq 0$ since $w(0) = 0, w(p_k^*) \geq 0$. Let $\sum_{k=1}^K a_k = a$, we have

$$\frac{w\left(\sum_{k=1}^K a_k\right)}{\sum_{k=1}^K w(p_k^*) - \sum_{k=1}^K w(p_k^* - a_k)} \geq \frac{w(a)}{\sum_{k=1}^K w'_-(p_k^*)a_k}.$$

Since $a \geq \max_{k\in[K]} p_k^* = p$, thus $w'_-(p_k^*) \leq w'_-(a)$ for any k by convexity of w. Since

$w(p) > 0$ and $w(0) = 0$, we must have $0 < w'_-(p) \le w'_-(x)$ for any $x \ge p$. Therefore,

$$\frac{w(a)}{\sum_{k=1}^{K} w'_-(p^*_k)a_k} \ge \frac{w(a)}{\sum_{k=1}^{K} w'_-(a)a_k} = \frac{w(a)}{w'_-(a)a} \ge \inf_{x \ge p} \frac{w(x)}{w'_-(x)x}.$$

□

C.5.1 Tightness of Greedy's Approximation Ratio in the Monomial Class

One implication of Theorem 4.7 is that for α-th degree polynomial function $w = |\mathcal{S}_k|^\alpha$, the greedy algorithm has $\frac{1}{\alpha}$-approximation. Next, we show that this approximation ratio is provably tight for every weight in the monomial class.

Consider an instance with $n = MN$ items and $m = M + L$ owners for some positive integer L, M, N such that $M^{L-1} \mid N$ and $N(1 - 1/M)^L \ge L$. The ownership relations are as follows (see also Figure C.1 for the visualization of the instance):

- There are M owners $1, \cdots, M$, each of whom owns a disjoint set of N items. Thus, collectively, they own all the MN items. Let the items owned by these owners as $\mathcal{I}^1, \ldots, \mathcal{I}^M$ respectively.

- We denote the item owned by the remaining L owners as $\mathcal{I}^{\ell+M}$ for $\ell \in [1, \ldots, L]$. Let $\left|\mathcal{I}^{\ell+M} \cap \mathcal{I}^{\ell}\right| = (1 - 1/M)^{\ell-1}N/M + 1$ and $\left|\mathcal{I}^{\ell+M} \cap \mathcal{I}^{\ell'}\right| = (1 - 1/M)^{\ell-1}N/M, \forall \ell' \ne \ell \in [M]$. Hence, $|\mathcal{I}^{\ell+M}| = (1 - 1/M)^{\ell-1}N + 1$.

Observe that $\mathcal{I}^{M+1}$ is the largest item set and $\mathcal{I}^{M+2}$ becomes the largest item set after $\mathcal{I}^{M+1}$ is removed. Therefore, we can see that the greedy algorithm proceeds by picking the ℓ-th owner of the last L owners from $\ell \in [1, \ldots, L]$ and lastly the first M owners, i.e., when each of whom has remaining item less than $N - \sum_{\ell=1}^{L}(1-1/M)^{\ell-1}N/M = (1-1/M)^L N$. As such, the greedy partition achieves the objective of at least $\mathrm{ALG}(\mathcal{G}) \le \sum_{\ell=1}^{L}[(1 - 1/M)^{\ell-1}N + 1]^\alpha + M(1 - 1/M)^{\alpha L}N^\alpha$.

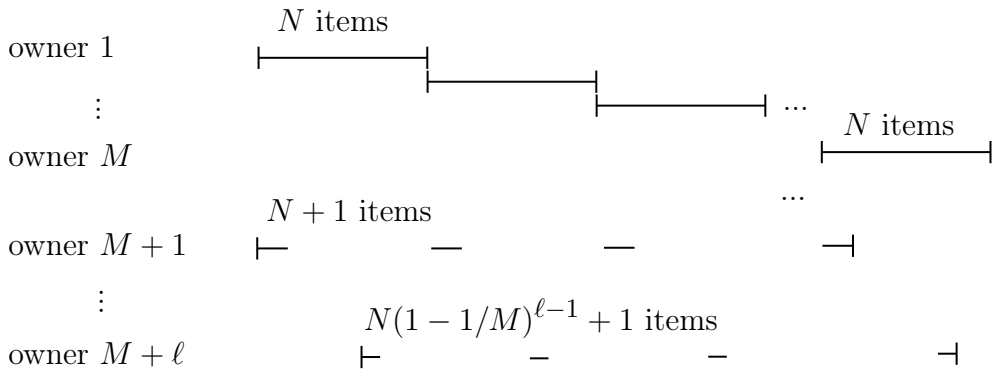


Figure C.1: Illustration of the worst case instance for the greedy algorithm. $l = 1, \cdots, L$.

On the other hand, consider a partition strategy that only picks the first M owners, which achieves the objective $\mathrm{OPT}(\mathcal{G}) = MN^\alpha$. Simply using it as a lower bound for the optimal objective, we can bound the approximation ratio as,

$$\begin{aligned}\frac{\mathrm{ALG}(\mathcal{G})}{\mathrm{OPT}(\mathcal{G})} &\leq \frac{\sum_{\ell=1}^{L}[(1-1/M)^{\ell-1}N+1]^\alpha + M(1-1/M)^{\alpha L}N^\alpha}{MN^\alpha} \\ &= \frac{\sum_{\ell=1}^{L}[(1-1/M)^{\ell-1}N+1]^\alpha}{MN^\alpha} + (1-1/M)^{\alpha L}.\end{aligned}$$

For any constant $M > 1$, we have as $N, L \to \infty$,

$$\begin{aligned}\lim_{N,L\to\infty} \frac{\sum_{\ell=1}^{L}[(1-1/M)^{\ell-1}N+1]^\alpha}{MN^\alpha} + (1-1/M)^{\alpha L} &= \lim_{N,L\to\infty} \sum_{\ell=1}^{L}(1-1/M)^{(\ell-1)\alpha}/M \\ &= \frac{1/M}{1-(1-1/M)^\alpha}.\end{aligned}$$

Finally, as $M \to \infty$, the ratio matches with the worst case approximation guarantee $1/\alpha$,

$$\lim_{M\to\infty} \frac{1/M}{1-(1-1/M)^\alpha} = \lim_{M\to\infty} \frac{1}{\alpha - 1/M} = 1/\alpha.$$

Meanwhile, despite the hardness result for the size-focused objective in Proposition 4.6, we would like to point out that the efficient algorithm for optimal partition under the α-th degree polynomial objective may exist. In particular, consider the following instance with

$m = n+1$, where the optimal partition cannot realize the set-cover problem for the hardness reduction. The owner i owns the items $\{2i-1, 2i\}$,$\forall i \in [n]$ and owner $n+1$ owns the items $\{1, 3, \cdots, 2n-1\}$. In this case, the optimal partition is not the minimum set cover, since

$$\text{obj}\left(\{\{1, 3, \cdots, 2n-1\}, \{2\}, \cdots \{2n\}\}\right) = n^{\alpha}+n > \text{obj}\left(\{\{1, 2\}, \cdots, \{2n-1, 2n\}\}\right) = n \times 2^{\alpha}.$$

Hence, it remains an open question on the hardness of finding the optimal partition under the α-th degree polynomial objective.

APPENDIX D

D.1 Supermodular Games

The formal definition of supermodular games is based on several basic concepts from the lattice theory: For a partial order set $\mathcal{L}$ with operator $\geq$, we define the "join" operator $x \vee y$ for every $x, y \in \mathcal{L}$, as the least upper bound of x, y, i.e., $(x \vee y) \geq x$, $(x \vee y) \geq y$ and $z \geq (x \vee y)$ for all $z \in \mathcal{L}, z \geq x$ and $z \geq y$. Similarly, the "meet" operator $x \wedge y$ is defined for every $x, y \in \mathcal{L}$, as the greatest lower bound of x, y, i.e., $x \geq (x \wedge y)$, $y \geq (x \wedge y)$ and $(x \vee y) \geq z$ for all $z \in \mathcal{L}, x \geq z$ and $y \geq z$. $\mathcal{L}$ is a lattice if for every $x, y \in \mathcal{L}$, $(x \wedge y), (x \vee y) \in \mathcal{L}$. We say $\mathcal{L}$ forms a complete lattice if every subset $\mathcal{L}' \subseteq \mathcal{L}$ forms a lattice. Therefore, any compact subset of $\mathbb{R}$ (with the usual order) is a complete lattice, as is any subset in $\mathbb{R}^d$ formed by the product of d compact sets in $\mathbb{R}$ (with the product order).

In addition, we say a function $f : \mathcal{L} \to \mathbb{R}$ is supermodular if for all $x, y \in \mathcal{L}$, it holds $f(x \wedge y) + f(x \vee y) \geq f(x) + f(y)$. A function $f : \mathcal{L}_1 \times \mathcal{L}_2 \to \mathbb{R}$ has increasing difference if for all $x' \geq x \in \mathcal{L}_1, y' \geq y \in \mathcal{L}_2$, it holds $f(x', y') - f(x, y') \geq f(x', y) - f(x, y)$. Now we can formally define supermodular games as follows.

Definition 12 (Supermodular Games)**.** *A strategic game* $\mathcal{G}(\mathcal{N}, \{\mathcal{A}_n\}_{n=1}^N, \{u_n\}_{n=1}^N)$ *is supermodular if for each* $n \in \mathcal{N}$*,*

1. *Each* $\mathcal{A}_n$ *is a complete lattice.*
2. $u_n(a_n, a_{-n})$ *is upper semi-continuous in* a_n *for each fixed* a_{-n}*, and it is continuous in* a_{-n} *for each fixed* a_n*, and has a finite upper bound.*
3. $u_n(a_n, a_{-n})$ *is supermodular in* a_n *for each fixed* a_{-n}*.*
4. $u_n(a_n, a_{-n})$ *has* increasing difference *in* a_n *and* a_{-n}*.*

The first condition is trivial when the $\mathcal{A}_n$ is a totally-ordered set (e.g., a compact set of real numbers). The second condition is trivial when the $\mathcal{A}$ is discrete. The third condition is

trivial when the $\mathcal{A}_n$ is single-dimensional. Hence, for the finite normal-form game of interest in this paper, we only need to focus on the condition of increasing difference.

Moreover, as pointed out by [Milgrom and Roberts, 1990, Etessami et al., 2020], the structural and algorithmic properties of supermodular game is preserved in a broader class of *games with strategic complementarities*, which relaxes the third and fourth condition to depend only on ordinal information on the utility functions, i.e. how the utilities compare to each other rather than their precise numerical values. In particular, the third condition can be relaxed to quasi-supermodularity, where a function $f : \mathcal{L} \to \mathbb{R}$ is quasi-supermodular if for all $x, y \in \mathcal{L}$, it holds $f(x) \geq f(x \wedge y) \implies f(x \vee y) \geq f(y)$. The fourth condition can be relaxed to the single-crossing condition, where a function $f : \mathcal{L}_1 \times \mathcal{L}_2 \to \mathbb{R}$ is single-crossing if for all $x' \geq x \in \mathcal{L}_1, y' \geq y \in \mathcal{L}_2$, it holds $f(x', y) \geq f(x, y) \implies f(x', y') \geq f(x', y)$.

D.2 Omitted Proofs in Section 5.5

D.2.1 Proof of Proposition 5.5

Proof of the first claim. This follows an induction argument. First, it is easy to see that A's action 1 is dominated by her action 2 since $u_A(1, j) = 1/\rho < 2/\rho = u_A(2, j)$ for all B's action $j \in [K]$. Second, we claim no other pure action of A or B can be eliminated by any pure or mixed strategy. This is due to two reasons: 1. for any $i > 1$, there exists $j = i - 1$ such that $u_1(i, j) = \frac{i}{\rho} \geq \max_{i'} u_1(i', j) = \frac{i}{\rho} \geq u_1(x, j)$ for any $x \in \mathcal{X}_A$; 2. for any $j \geq 1$, there exists $i = j$ such that $u_2(i, j) = \frac{j}{\rho} \geq \max_{j'} u_2(i, j') = \frac{j}{\rho} \geq u_2(i, x)$ for any $x \in \mathcal{X}_B$.

Now suppose that for some $i \geq 1$, A has eliminated action $1, \cdots, i$ and B has eliminated action $1, \cdots, i - 1$. We claim that B's action i will be the only iteratively dominated action currently. Specifically, it is now iteratively dominated by her action $i + 1$ because $u_B(j, i) = i/\rho < (i + 1)/\rho = u_B(j, i + 1)$ for all $j > i$ (note that A has eliminated any action $j \leq i$).

Therefore, B will next eliminate action i. On the other hand, it is easy to verify that any action $i' \geq i+1$ of B is not dominated currently by any non-eliminated mixed or pure strategy since whenever A plays i', B's action i' yields the largest utility for her. Specifically, we have $u_B(i', i') = i'/\rho > u_B(i', j)$, which is j/ρ when $j < i'$ and $-c/\rho$ when $j > i'$. Therefore, any $i' \geq i+1$ cannot be a iteratively dominated action of B. Similarly, one can also verify that A's any action $i' \geq i+1$ currently is not a iteratively dominated strategy neither. So B's action i is the only iteratively dominated action.

Similar to the analysis above, one can show that after agent B's action i is eliminated, agent A's action $i+1$ will become the only iteratively dominated action. By induction, we know the iterative elimination procedure will eliminate one action at each iteration until we reach the last action profile (K, K). So the elimination length is exactly $2K-2$.

Proof of the second claim. We show a stronger result: any finite dominance solvable game has a unique correlated equilibrium (CE). This also implies the uniqueness of Nash equilibrium (NE) since any NE must be a CE as well. The above conclusion should be standard. However, since we are unable to find any existing proof for the reference, we here include a formal argument for the readers with little background in game theory for completeness of the paper.

Given any game $\mathcal{G}$ with elimination length L_0 and elimination sets $E_1 \subset E_2 \cdots \subset E_{L_0}$, we claim that any CE must assign zero probability on actions in E_{L_0}. We complete the proof by contradiction. Suppose it is not the case; there must exist an action in E_{L_0} that has positive probability in some CE. Let l be the smallest index such that E_l contains an action a_n for agent n in this CE. By our choice of l, all actions in E_{l-1} have zero probability in the CE. This, however, implies that a_n must be iteratively dominated by some other action a_n' in this CE, which contradicts the definition of CE since whenever action a_n is recommended to agent-n, he would strictly prefer to deviating to action a_n'. Therefore, any action in E_{L_0} must have zero probability in any CE. This implies that there is only one

unique CE in any finite dominance solvable game, which is also the unique NE.

Finally, it is easy to see that $\frac{K}{\rho}$ is the largest possible utility in the payoff matrices. Therefore, at this equilibrium (K, K), both agents achieves the maximum possible utility $\frac{K}{\rho}$, and the game obtains the maximum social welfare $\frac{2K}{\rho}$.

D.2.2 Proof of Theorem 5.6

Proof. We explicitly construct a mixed strategy for the two agents and prove that it constitutes an ϵ-correlated-equilibrium and moreover satisfies the claimed properties in the stated theorem. Such construction is possible due to the special structure of the DIR games. Our construction is divided into two main steps: (1) constructing the support of the CE; (2) constructing the concrete probabilities.

Support of the constructed CE. Specifically, consider a distribution π over the joint action space $[K] \times [K]$ with support in the following format

$$\pi = \begin{bmatrix} \delta_1 & 0 & 0 & 0 & 0 \\ \delta_2 & \delta_3 & 0 & 0 & 0 \\ 0 & \delta_4 & \delta_5 & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \delta_{2K-2} & \delta_{2K-1} \end{bmatrix}.$$

We claim that π will be an ϵ-CE if its $\{\delta_i\}_{i=1}^{2K-1}$ satisfy the following inequality system:

$$\begin{cases} \delta_1 \cdot 1 \le \epsilon \cdot \rho \\ \delta_2 \cdot (-c-2) + \delta_3 \cdot 1 \le \epsilon \cdot \rho \\ \delta_4 \cdot (-c-3) + \delta_5 \cdot 1 \le \epsilon \cdot \rho \\ \quad \vdots \\ \delta_{2K-4} \cdot (-c-(K-1)) + \delta_{2K-3} \cdot 1 \le \epsilon \cdot \rho \end{cases} \quad \begin{cases} \delta_1 \cdot (-c-1) + \delta_2 \cdot 1 \le \epsilon \cdot \rho \\ \delta_3 \cdot (-c-2) + \delta_4 \cdot 1 \le \epsilon \cdot \rho \\ \quad \vdots \\ \delta_{2K-3} \cdot (-c-(K-1)) + \delta_{2K-2} \cdot 1 \le \epsilon \cdot \rho \end{cases} \tag{D.1}$$

and moreover $\sum_{i=1}^{2K-1} \delta_i = 1$. To see this, by the structure of π, we know that whenever agent A is recommended action i, agent B will be recommended either action $i-1$ or action i. Since $u_A(i, i-1) = u_A(i, i) = i/\rho$, agent A will get utility i/ρ for sure when recommended action i. It is thus easy to see that agent A will not be willing to deviate to any action $i' < i$ since his utility can only be at most i'/ρ for playing i'. Similarly, since $u_A(i', i) = u_A(i', i-1) = -c/\rho$ for any $i' > i+1$, so agent A will not deviate to any action $i' > i+1$ neither. In other words, the structure of π already guarantees that whenever agent A's action i is recommended, she would only be possibly having incentive to deviate action $i+1$, for which she gets utility $-c\delta_{2i-2}/\rho + (i+1)\delta_{2i-1}/\rho$. The ϵ-CE condition thus requires the following

$$i(\delta_{2i-2} + \delta_{2i-1})/\rho \le -c\delta_{2i-2}/\rho + (i+1)\delta_{2i-1}/\rho + \epsilon.$$

Simple algebraic calculation shows that the above is exactly the ith constraint in the left-hand-side of System (D.1). Similarly, the j'th constraint in the right-hand-side of System (D.1) says whenever the agent B is recommended action j, she would prefer j over $j+1$ for $j = 1, \cdots, K-1$, which is the only constraint needed to guarantee ϵ-CE.

Distribution of the constructed CE. We now explicitly construct $\{\delta_i\}_{i=1}^{2K-1}$ that satisfies linear system (D.1). Suppose $c > 1$. Pick any $k \in \{2, \ldots, 2K-1\}$, for any ϵ such that

$1/\epsilon \in (\rho \sum_{i=1}^{k-1} c^{i-1}, \rho \sum_{i=1}^{k} c^{i-1}]$, we can verify that

$$\delta_i = \begin{cases} \rho \cdot \epsilon c^{i-1} & i < k \\ 1 - \rho \cdot \epsilon \sum_{i=1}^{k-1} c^{i-1} & i = k \\ 0 & i > k \end{cases} \tag{D.2}$$

is a feasible solution to linear system (D.1), and therefore forms an $\epsilon-$CE. With $k \leq 2K-2$, we obtained an $\epsilon-$CE with $\delta_{2K-1} = 0$ for any ϵ satisfying $1/\epsilon \leq c^{2K-2} \leq \rho \sum_{i=1}^{2K-2} c^{i-1}$. This concludes our proof for the first part of Theorem 5.6.

Welfare property of the constructed CE. Finally, we derive the upper bound of the welfare at the above $\epsilon-$CE. Observe that by construction of π, the welfare (i.e., sum of agents' utilities) at the action profile for δ_i is precisely $(i+1)/\rho$. Therefore, we can compute and bound the total welfare as follows:

$$\sum_{i=1}^{2K-1} \delta_i \cdot \frac{i+1}{\rho} = \sum_{i=1}^{k} \delta_i \cdot \frac{i+1}{\rho} \leq \sum_{i=1}^{k} \delta_i \cdot \frac{k+1}{\rho} \leq \frac{k+1}{\rho}$$

where the first equality is by Eq. (D.2), and the first and second inequality is implied from the fact that $\frac{i+1}{\rho} \leq \frac{k+1}{\rho}, \forall i \leq k$ and $\sum_{i=1}^{k} \delta_i = 1$.

We know $\rho c^{k-2} \leq \rho \sum_{i=1}^{k-1} c^{i-1} < 1/\epsilon$. This implies $k - 2 < \frac{\log(1/\epsilon) - \log(\rho)}{\log(c)}$. Since $k \in \mathbb{N}$, we have $k - 1 \leq \lceil \frac{\log(1/\epsilon) - \log(\rho)}{\log(c)} \rceil$. Since $\rho \geq c$, $\lceil \frac{\log(1/\epsilon) - \log(\rho)}{\log(c)} \rceil \leq \lceil \frac{\log(1/\epsilon)}{\log(c)} - 1 \rceil \leq \lceil \frac{\log(1/\epsilon)}{\log(c)} \rceil - 1$. Therefore, the welfare is at most $\frac{k+1}{\rho}$, which is at most $\frac{1 + \lceil \log(1/\epsilon)/\log(c) \rceil}{2K}$ fraction of the equilibrium welfare $2K/\rho$.

□

D.2.3 DIR Games Are not Globally Variationally Stable

According to the definition of variational inequality (Eq (6) in [Cohen et al., 2017a]), to see why the global variational inequality fails to hold, we only need to show that there exists $(x, y) \in \Delta_{[K]} \times \Delta_{[K]}$ such that $v(x, y) \cdot ((x, y) - (x^*, y^*)) > -\frac{\Delta}{2}\|(x, y) - (x^*, y^*)\|$ for any $\Delta > 0$, where $(x^*, y^*) = (0, \cdots, 0, 1, 0, \cdots, 0, 1)$ is the unique NE of the game with payoff matrices defined in (5.2), $\|\cdot\|$ is the L_1 norm, and

$$v(x, y) = (v_{A,1}(y), v_{A,2}(y), \cdots, v_{A,K}(y), v_{B,1}(x), v_{B,2}(x), \cdots, v_{B,K}(x)),$$

where $v_{A,j}(y)$ is the payoff of when agent A plays pure strategy j and agent B plays the mixed strategy y.

Let $x = y = (0, \cdots, 0, 1, 0, \cdots, 0)$, i.e., the only non-zero term of x and j is the i-th element, where $1 \leq i \leq K - 1$. Then we have

$$v(x, y) \cdot ((x, y) - (x^*, y^*)) = 2i > 0 > -\frac{\Delta}{2}\|(x, y) - (x^*, y^*)\|,$$

meaning the global variational inequality is violated.

D.3 Additional Discussion for Merit-based Algorithms

In this section we demonstrate the broadness of merit-based algorithms by showing all Dual Averaging (DA) and Follow the Perturbed-Leader (FTPL) algorithms are merit-based. To complete the definition of merit-based algorithms, we formally introduce the concept of order-preserving functions:

Definition 13. *We call a function $F : \mathbb{R}^K \to \Delta_K$ order-preserving if for any $\boldsymbol{y} = (y_1, \cdots, y_k)$ and i, j such that $y_i < y_j$, we must have $F(\boldsymbol{y})_i \leq F(\boldsymbol{y})_j$.*

We call an online learning algorithm merit-based if it utilizes an order-preserving function

to determine the action distribution from accumulated rewards. The formal definition is shown in Algorithm 5.1.

By definition 8, Any algorithm 5.1 equipped with an order-preserving function F is merit-based. During each round, a merit-based algorithm first maps the accumulated score vector $\boldsymbol{y}_t$ to a distribution $\boldsymbol{p}_t = (p_1(t), \cdots, p_K(t)) \in \Delta_K$ with a pre-chosen order-preserving function F and then samples the current strategy from $\boldsymbol{p}_t$. A merit-based algorithm also needs to specify a learning rate sequence $\{\eta_t\}$ to accumulate the collected rewards $\widetilde{\boldsymbol{u}}_t$ from each round. We note that in general, the reward $\widetilde{\boldsymbol{u}}_t$ can be any unbiased estimation of the true reward $\boldsymbol{u}_t$. However, since we focus on demonstrating a negative side of merit-based algorithms in Theorem 5.7, we consider the most informative type of feedback, i.e., the noiseless full-information setting.

Next, we introduce the preliminaries of DA and FTPL algorithms, whose descriptions are shown in Algorithm D.1 and D.2.

ALGORITHM D.1: The DA Algorithm Framework

Input: Mirror map $Q : \mathcal{Y} \to \Delta_K$, learning rate sequence $\{\eta_t > 0\}$.
$\boldsymbol{y}_1 \leftarrow (0, \cdots, 0)$
for $t = 1 \dots T$ **do**
    Compute $\boldsymbol{p}_t = Q(\boldsymbol{y}_t)$
    Draw an action i_t from the distribution $\boldsymbol{p}_t$.
    Receive the expected payoff $\widetilde{\boldsymbol{u}}_t = (\widetilde{u}_1(t), \cdots, \widetilde{u}_K(t))$ for each action i from the first-order oracle.
    Update $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t \widetilde{\boldsymbol{u}}_t$.

In online learning literature, DA coincides with Follow-the-Regularized-Leader (FTRL) in the cases of linear losses, and is also known as the "lazy" version of online mirror descent (OMD) [Shalev-Shwartz et al., 2011]. Similar to the merit-based defined in Algorithm 5.1, the DA algorithm uses a "mirror map" Q to derive the mixed strategy $\boldsymbol{p}_t$ from the accumulated reward $\boldsymbol{y}_t$. By convention, the mirror map Q depends on a convex function (or regularizer)

h and is defined as

$$Q(\boldsymbol{y}) = \arg\max_{x \in \Delta_K} \{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - h(\boldsymbol{x})\}, \boldsymbol{y} \in \mathcal{Y}. \tag{D.3}$$

A regularizer is said to be *symmetric* if for any $1 \leq i < j \leq K$, $h(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_K) = h(x_1, \ldots, x_j, \ldots, x_i, \ldots, x_K)$. That is, the value of $h(x)$ will not change if we swap the values at any two coordinates of x. Most (if not all) known DA algorithms use symmetric and strictly convex regularizers. In this case, symmetry of h implies a natural property of the algorithm — i.e., if two actions have the same accumulated score, the algorithm should play either with equal probability. This exactly conforms to the definition of order-preserving function. We formalize this connection in Lemma D.2, which supports our main claim in Proposition D.1.

Proposition D.1. *Any DA algorithm equipped with a mirror map induced by a symmetric regularizer is merit-based.*

Proof. By the definition of merit-based algorithms, we only need to show the following Lemma:

Lemma D.2. *Any mirror map $\boldsymbol{p} = Q(\boldsymbol{y})$ induced by a symmetric regularizer h is order-preserving.*

Proof. We prove by contradiction. Suppose for some $\boldsymbol{y}$ and i, j where $y_i > y_j$, while $\boldsymbol{p} = Q(\boldsymbol{y})$ with $p_i < p_j$. By definition, $\boldsymbol{p} = Q(\boldsymbol{y}) = \arg\max_{x \in \Delta_K} \{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - h(\boldsymbol{x})\}$. However, consider the vector $\widetilde{\boldsymbol{p}} = (p_1, \ldots, p_j, \ldots, p_i, \ldots, p_K)$. By construction, $\widetilde{\boldsymbol{p}} \in \Delta_K$ and $h(\widetilde{\boldsymbol{p}}) = h(\boldsymbol{p})$. By rearrangement inequality, we have $\widetilde{p}_i y_j + \widetilde{p}_j y_j = p_j y_i + p_i y_j > p_i y_j + p_j y_j$ such that $\langle \boldsymbol{y}, \widetilde{\boldsymbol{p}} \rangle - h(\widetilde{\boldsymbol{p}}) > \langle \boldsymbol{y}, \boldsymbol{p} \rangle - h(\boldsymbol{p})$. This is a contradiction to $\boldsymbol{p} = \arg\max_{x \in \Delta_K} \{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - h(\boldsymbol{x})\}$. Therefore, it must be the case that whenever $y_i > y_j$, $p_i \geq p_j$. By definition 13, Lemma D.2 holds. □

□

The DA family includes many celebrated algorithms such as Exponential Weight (EW), lazy gradient descent (LGD) and fictitious play. Below we list a few widely used algorithms from the DA family which, unsurprisingly, all use symmetric and strictly convex regularizers:

1. when $h(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$ is the quadratic function, the mirror map $Q(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}\in\Delta_K} \|\boldsymbol{x}-\boldsymbol{y}\|^2$ takes the form of Euclidean projection and we obtain the lazy gradient descent (LGD) algorithm.

2. when $h(\boldsymbol{x}) = \sum_{i\in[K]} x_i \log x_i$ is the entropic regularizer, the mirror map $Q(\boldsymbol{y}) = \frac{(\exp(y_1),\cdots,\exp(y_K))}{\exp(y_1)+\cdots+\exp(y_1)}$ takes the form of logit choice map and we obtain entropic gradient descent algorithm, which is also known as the Hedge or Exponential Weight (EW).

3. when $h(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{x}\|^p$ is the normalized L_p norm and $p \to \infty$, $Q(\boldsymbol{y})$ always returns the pure best strategy to the opponent's average past mixed strategy, and the corresponding algorithm is known as fictitious play [Brown, 1951, Viossat and Zapechelnyuk, 2013].

Next, we show that FTPL algorithms are also merit-based. The outline of FTPL family is shown in Algorithm D.2.

ALGORITHM D.2: The FTPL Algorithm Framework

Input: A probability distribution $\mathcal{D}$, learning rate sequence $\{\eta_t > 0\}$.
$\boldsymbol{y}_1 \leftarrow (0, \cdots, 0)$
for $t = 1 \dots T$ **do**
  Sample $\{\epsilon_i\}_{i=1}^K$ independently from $\mathcal{D}$.
  Choose the action $i_t = \arg\max_{i\in[K]}\{y_i(t) + \epsilon_i\}$.
  Receive the expected payoff $\widetilde{\boldsymbol{u}}_t = (\widetilde{u}_1(t), \cdots, \widetilde{u}_K(t))$ for each action i from the first-order oracle.
  Update $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_t \widetilde{\boldsymbol{u}}_t$.

The distribution $\mathcal{D}$ can be any single variable distribution, e.g., Gaussian, Gumbel, exponential, etc. When $\mathcal{D}$ takes the Gumbel distribution with zero mean, the FTPL algorithm coincides with multiplicative weights update (MWU). For FTPL algorithms, the link function F that maps the accumulated rewards to action distribution is implicit. However, it is

straightforward to see that F is order-preserving and thus all FTPL algorithms are merit-based. We formalize the claim in the following proposition D.3:

Proposition D.3. *The class of FTPL algorithms are merit-based.*

Proof. Any FTPL algorithm can be equivalently represented in the form of Algorithm 5.1 with a link function F defined as

$$F(\boldsymbol{y})_i = \mathbb{P}\left[i = \arg\max_{k\in[K]}\{y_k + \epsilon_k\}\right], \forall k \in [K].$$

Because random variables ϵ_i are i.i.d., for any $y_i > y_j$ it must hold that

$$\mathbb{P}\left[i = \arg\max_{k\in[K]}\{y_k + \epsilon_k\}\right] \geq \mathbb{P}\left[j = \arg\max_{k\in[K]}\{y_k + \epsilon_k\}\right].$$

Therefore, F is order-preserving and the induced algorithm is merit-based. □

D.4 Omitted Proofs in Section 5.6

D.4.1 Proof of Corollary 5.12

From Eq (5.11), we know that a sufficient condition for t to *ε-essentially eliminate* a dominated action $a \in E_1$ is

$$\frac{t^{-b}}{K} + \exp\left(4\left(\sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b}}\right)\log^{\frac{1}{2}}\frac{2K}{\delta}\cdot t^{\frac{1+b}{2}} - \frac{\Delta t}{16(1+\beta)}\right) < \frac{\min\{\varepsilon, \Delta/2\}}{4KN}. \tag{D.4}$$

Note that a sufficient condition to satisfy Eq (D.4) above is the following

$$4\left(\sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b}}\right)\log^{\frac{1}{2}}\frac{2K}{\delta}\cdot t^{\frac{1+b}{2}} - \frac{\Delta t}{32(1+\beta)} < 0$$

$$\frac{t^{-b}}{K} < \frac{\min\{\varepsilon, \Delta/2\}}{8KN} \qquad \text{and} \qquad -\frac{\Delta t}{32(1+\beta)} < \log\frac{\min\{\varepsilon, \Delta/2\}}{8KN},$$

which can be simplified to

$$t > \max\{O(N^{\frac{1}{b}} \min\{\varepsilon, \Delta\}^{-\frac{1}{b}}), O(K^{\frac{1}{1-b}}(1+\sigma^2)^{\frac{1}{1-b}}\beta^{\frac{1}{1-b}}\Delta^{-\frac{2}{1-b}} \log^{\frac{1}{1-b}} \frac{1}{\delta}), O(\beta\Delta^{-1} \log \frac{KN}{\min\{\varepsilon, \Delta\}})\}.$$

To satisfy the above condition, we can simply take

$$t = O\left(\max\{N^{\frac{1}{b}}, K^{\frac{1}{1-b}}\} \max\{\varepsilon^{-\frac{1}{b}}, \Delta^{-\max\{\frac{1}{b}, \frac{2}{1-b}\}}\}(1+\sigma^2)^{\frac{1}{1-b}}\beta^{\frac{1}{1-b}} \log^{\frac{1}{1-b}} \frac{1}{\delta}\right).$$

In particular, when $b = \frac{1}{3}$, we may choose

$$T_1 = O\left(\max\{N^3, K^{1.5}\} \min\{\varepsilon, \Delta\}^{-3}(1+\sigma^2)^{1.5}\beta^{1.5} \log^{1.5} \frac{1}{\delta}\right).$$

Therefore, with probability at least $1 - |E_1|(T_1 + s) > 1 - 2|E_1|(T_1 + s)^2$, actions in E_1 will be essentially eliminated at round $T_1 + 1, \cdots, T_1 + s$.

From Eq (5.7), we can thus upper bound T_{L_0} by

$$\begin{aligned}
T_{L_0} &< (1 + \frac{8}{\Delta})^{\frac{L_0}{1+\beta}} \cdot \left(T_1 + L_0 \cdot \frac{(1+\beta)^2(4+\Delta)^2(8+\Delta)^2}{4(1+2\beta)\Delta^2} \log \frac{1}{\delta}\right) \\
&\sim O\left(\Delta^{-\frac{L_0}{1+\beta}}(T_1 + L_0\beta\Delta^{-2} \log \frac{1}{\delta})\right) \\
&= O\left(\Delta^{-\frac{L_0}{1+\beta}}(\max\{N^3, K^{1.5}\} \min\{\varepsilon, \Delta\}^{-3}(1+\sigma^2)^{1.5}\beta^{1.5} \log^{1.5} \frac{1}{\delta} + L_0\beta\Delta^{-2} \log \frac{1}{\delta})\right) \\
&= O\left(\max\{N^3, K^{1.5}\} \min\{\varepsilon, \Delta\}^{-3}(1+\sigma^2)^{1.5}\beta^{1.5} \log^{1.5} \frac{1}{\delta}\right),
\end{aligned}$$

which implies that E_{L_0} will be ε-essentially eliminated in $O\left(\max\{N^3, K^{1.5}\} \min\{\varepsilon, \Delta\}^{-3}(1+\sigma^2)^{1.5}\beta^{1.5} \log\right.$ iterations with probability at least $1 - 2KNT_{L_0}^2\delta$. Hence, we claim that E_{L_0} will be ε-essentially eliminated in

$$\widetilde{O}\left(\max\{N^3, K^{1.5}\} \min\{\varepsilon, \Delta\}^{-3}(1+\sigma^2)^{1.5}\beta^{1.5} \log^{1.5} \frac{1}{\delta}\right)$$

iterations with probability at least $1-\delta$ for any δ.

D.4.2 Proofs of Technical Lemmas

Proof. **of Lemma 5.13**

Fix $T > 0$ and any action a. Let $\Xi_t = \gamma_t^{(T)}(\widetilde{u}_a(t) - u_a(t))$ and $S_t = \sum_{s=1}^{t} \Xi_s$. Since T is fixed, $\mathbb{E}[S_t]$ is bounded. Moreover, we have

$$\mathbb{E}[S_t|S_{t-1}, \cdots, S_1] = S_{t-1} + \gamma_t^{(T)} \cdot \mathbb{E}[\widetilde{u}_a(t) - u_a(t)|\mathcal{F}_{t-1}] = S_{t-1}.$$

By definition, $\{S_t\}_{t=1}^{T}$ is a martingale.

Apply Azuma's inequality to $\{S_t\}$, for any $x > 0, 1 \le t \le T$, we have

$$\mathbb{P}[|S_t| \ge x] \le 2\exp\left(-\frac{x^2}{2W}\right). \tag{D.5}$$

W is a upper bound of $\sum_{i=1}^{t} \mathbb{E}[\Xi_i^2|\mathcal{F}_{i-1}]$. Note that

$$\begin{aligned}
\mathbb{E}[\Xi_t^2|\mathcal{F}_{t-1}] &= (\gamma_t^{(T)})^2\mathbb{E}\left[(0-u_a(t))^2 \cdot (1-p_a(t)) + \left(\frac{u_a(t)+\xi_t}{p_a(t)} - u_a(t)\right)^2 \cdot p_a(t)\middle|\mathcal{F}_{t-1}\right] \\
&= (\gamma_t^{(T)})^2\mathbb{E}\left[(u_a^2(t)+\xi_t^2) \cdot \frac{1}{p_a(t)} - u_a^2(t)\middle|\mathcal{F}_{t-1}\right] \\
&\le (\gamma_t^{(T)})^2\left(\frac{K(1+\sigma^2)}{\epsilon_t}\right),
\end{aligned} \tag{D.6}$$

we can take $n = T$ and $W = \sum_{i=1}^{T} \gamma_i^2(\frac{K(1+\sigma^2)}{\epsilon_i})$ in Eq (D.5) and obtain

$$x \ge \sqrt{2W\log\frac{2}{\delta}}, \tag{D.7}$$

which implies with probability at least $1-\delta$,

$$|\sum_{t=1}^{T}\gamma_t^{(T)}(\widetilde{u}_a(t)-u_a(t))|<2\left(\sqrt{\log\frac{2}{\delta}}\cdot\sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{(\gamma_t^{(T)})^2}{\epsilon_t}}\right).$$

□

Proof. **of Lemma 5.14**

By the definition of strict dominance, there exists $\Delta>0$ such that the mixed strategy $x=(x_1,\cdots,x_K)$ satisfies $u_x(t)-u_a(t)>\Delta$ for all $t>0$. From Lemma 5.13, we can take a union bound over all the actions $a\in[K]$ and obtain with probability $1-K\delta'$,

$$|\sum_{t=1}^{T}\gamma_t^{(T)}(\widetilde{u}_a(t)-u_a(t))|<2\left(\sqrt{\log\frac{2}{\delta'}}\cdot\sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{(\gamma_t^{(T)})^2}{\epsilon_t}}\right),\forall a\in[K].$$

As a result,

$$\begin{aligned}
&y_x(T+1)-y_a(T+1)\\
=&\sum_{t=1}^{T}\gamma_t[\widetilde{u}_x(t)-\widetilde{u}_a(t)]\\
=&\sum_{t=1}^{T}\gamma_t[u_x(t)-u_a(t)]+\sum_{t=1}^{T}\gamma_t[-u_x(t)+\widetilde{u}_x(t)]+\sum_{t=1}^{T}\gamma_t[-\widetilde{u}_a(t)+u_a(t)]\\
=&\sum_{t=1}^{T}\gamma_t[u_x(t)-u_a(t)]+\sum_{i\in\mathcal{A}}x_i\sum_{t=1}^{T}\gamma_t[-u_i(t)+\widetilde{u}_i(t)]+\sum_{t=1}^{T}\gamma_t[-\widetilde{u}_a(t)+u_a(t)]\\
\geq&\sum_{t=1}^{T}\gamma_t[u_x(t)-u_a(t)]-4\left(\sqrt{\log\frac{2}{\delta'}}\cdot\sqrt{K(1+\sigma^2)\sum_{t=1}^{T}\frac{\gamma_t^2}{\epsilon_t}}\right).
\end{aligned}$$

Letting $\delta' = \frac{\delta}{K}$ yields Eq (5.9). When $\gamma_t = (t/T)^\beta, \epsilon_t = t^{-b}$ and assume $T > \beta$, we have

$$\sum_{t=1}^{T} \gamma_t [u_x(t) - u_a(t)] \geq \Delta \sum_{t=1}^{T} \gamma_t = \Delta \sum_{s=1}^{T} (\frac{s}{T})^\beta > \Delta T \cdot \int_0^1 x^\beta dx = \frac{\Delta T}{1+\beta}, \tag{D.8}$$

$$\begin{aligned} \sqrt{K(1+\sigma^2) \sum_{t=1}^{T} \frac{\gamma_t^2}{\epsilon_t}} &< \sqrt{K(1+\sigma^2) T^{1+b} \int_{\frac{1}{T}}^{1+\frac{1}{T}} x^{2\beta+b} dx} \\ &< \sqrt{\frac{eK(1+\sigma^2)}{1+2\beta+b} \cdot T^{1+b}}, \end{aligned} \tag{D.9}$$

where Eq (D.9) holds because

$$\int_{\frac{1}{T}}^{1+\frac{1}{T}} x^\alpha dx = \frac{1}{1+\alpha} \cdot \left[\left(1 + \frac{1}{T}\right)^{1+\alpha} - \left(\frac{1}{T}\right)^{1+\alpha} \right] < \frac{1}{1+\alpha} \left(1 + \frac{1}{T}\right)^T < \frac{e}{1+\alpha}.$$

Therefore, Eq (5.10) holds. □

Proof. **of Lemma 5.15** Consider the sequence of events

$$A_t = \{x_{-i} \text{ contains essentially eliminated actions at round } T_k + t\}, \quad t = 1, \cdots, T'.$$

Specifically, let $T = T_k + T'$ be fixed and define random variables

$$Z_t = \sum_{s=T_k+1}^{T_k+t} \gamma_s^{(T)} [u_x(s) - u_a(s) - \frac{\Delta}{2}], \quad 1 \leq t \leq T'.$$

We further let $Z_0 = 0$ and now show that $\{Z_t\}_{t=0}^{T'}$ is a sub-martingale. Since E_k has been ε-essentially eliminated at any $t \geq T_k$ and since $\varepsilon < \frac{1}{2}$, we have

$$\mathbb{P}[A_t | A_{t-1}, \cdots, A_1] < |E_k| \cdot \frac{\min\{\varepsilon, \Delta/2\}}{4KN} \leq KN \cdot \frac{\Delta}{8KN} = \frac{\Delta}{8}.$$

Note that when A_t does not happen, no actions in E_k will be played by agent i's opponents, in which case we have $u_x(t) - u_a(t) \geq \Delta$; On the other hand, when A_t happens we have $u_x(t) - u_a(t) \geq -2$ due to bounded utilities. Hence,

$$\begin{aligned}
&\mathbb{E}[Z_t | Z_{t-1}, \cdots, Z_1] \\
&= Z_{t-1} + \gamma_t^{(T)} \cdot \mathbb{E}[u_x(t) - u_a(t)] - \gamma_t^{(T)} \cdot \frac{\Delta}{2} \\
&\geq Z_{t-1} + \gamma_t^{(T)} \cdot [\mathbb{P}[A_t | A_{t-1}, \cdots, A_1] \cdot (-2) + (1 - \mathbb{P}[A_t | A_{t-1}, \cdots, A_1]) \cdot \Delta] - \gamma_t^{(T)} \cdot \frac{\Delta}{2} \\
&> Z_{t-1} + \gamma_t^{(T)} \cdot \left[\frac{\Delta}{8} \cdot (-2) + (1 - \frac{\Delta}{8}) \cdot \Delta \right] - \gamma_t^{(T)} \cdot \frac{\Delta}{2} \\
&> Z_{t-1} + \gamma_t^{(T)} \cdot \left[\frac{\Delta}{8} \cdot (-2) + (1 - \frac{2}{8}) \cdot \Delta - \frac{\Delta}{2} \right] = Z_{t-1}.
\end{aligned}$$

Therefore, $\{Z_t\}_{t=1}^{T'}$ is a sub-martingale. Therefore, let $c_i = \gamma_{T_k+i}^{(T)}(2 + \frac{\Delta}{2})$ denote a upper bound of $|Z_i - Z_{i-1}|$. By Azuma's inequality, we have

$$\mathbb{P}[Z_t \leq -\epsilon] \leq \exp\left(\frac{-\epsilon^2}{2\sum_{i=1}^{t} c_i^2} \right), \forall \epsilon > 0. \tag{D.10}$$

Let

$$t = T', \epsilon = \frac{\Delta}{4} \sum_{t=T_k+1}^{T_k+T'} \gamma_t^{(T)}, \text{ and } \exp\left(\frac{-\epsilon^2}{2\sum_{i=1}^{t} c_i^2} \right) \leq \delta,$$

we obtain

$$\sum_{t=T_k+1}^{T_k+T'} \gamma_t [u_a(t) - u_{a'}(t)] > \frac{\Delta}{4} \sum_{t=T_k+1}^{T_k+T'} \gamma_t \tag{D.11}$$

with probability at least $1 - (|E_{k+1}| - |E_k|)T'\delta$ for any $a \in E_{k+1} \setminus E_k$, as long as T' satisfies

$$\frac{\Delta^2}{16} \left(\sum_{t=T_k+1}^{T_k+T'} \gamma_t \right)^2 > 2(2 + \frac{\Delta}{2})^2 \sum_{t=T_k+1}^{T_k+T'} \gamma_t^2 \log \frac{1}{\delta}. \tag{D.12}$$

We now derive a sufficient condition for Eq (D.12) to hold, after substituting $\gamma_t^{(T)} = (t/T)^\beta$

into Eq (D.12) and assuming $T \geq (1+\frac{8}{\Delta})^{\frac{1}{1+\beta}} T_k, T \geq T_k + 1 + 2\beta$. We first lower bound the LHS of Eq (D.12) via the following bound (recall $T = T_k + T'$):

$$\begin{aligned}\sum_{t=T_k+1}^{T_k+T'} \gamma_t^{(T)} &> T' \int_{\frac{T_k}{T}}^{1} x^\beta dx \\ &= \frac{T'}{1+\beta}\left[1 - \left(\frac{T_k}{T}\right)^{1+\beta}\right] \\ &\geq \frac{8T'}{(8+\Delta)(1+\beta)}, \end{aligned} \tag{D.13}$$

where Eq (D.13) holds because $T \geq (1+\frac{8}{\Delta})^{\frac{1}{1+\beta}} T_k$. We then upper bound the RHS of Eq (D.12) via the following bound:

$$\begin{aligned}\sum_{t=T_k+1}^{T_k+T'} (\gamma_t^{(T)})^2 &< -\left(\frac{T_k}{T}\right)^{2\beta} + 1 + (T - T_k)\int_{\frac{T_k}{T}}^{1} x^{2\beta} dx \\ &< 1 + \frac{T'}{1+2\beta} \leq \frac{2T'}{1+2\beta}, \end{aligned} \tag{D.14}$$

where Eq (D.14) holds because $T \geq T_k + 1 + 2\beta$.

Consequently, a sufficient condition for Eq (D.12) to hold is

$$\frac{\Delta^2}{16}\left(\frac{8T'}{(8+\Delta)(1+\beta)}\right)^2 > 2(2+\frac{\Delta}{2})^2 \frac{2T'}{1+2\beta}\log\frac{1}{\delta},$$

which yields

$$T' \geq \frac{(1+\beta)^2(4+\Delta)^2(8+\Delta)^2}{4(1+2\beta)\Delta^2}\log\frac{1}{\delta}. \tag{D.15}$$

Since Eq (D.15) implies $T' > 1 + 2\beta$, we can simply take

$$T_0 = \max\left\{\frac{(1+\beta)^2(4+\Delta)^2(8+\Delta)^2}{4(1+2\beta)\Delta^2}\log\frac{1}{\delta}, \left[(1+\frac{16}{\Delta})^{\frac{1}{1+\beta}} - 1\right]\cdot T_k\right\}. \tag{D.16}$$

Hence, we complete the proof. □

D.5 Elimination Length in Akerlof's Market for "Lemons"

Proposition D.4. *Suppose each seller observes his exact car quality, i.e., $\widetilde{q}_i = q_i$. With any $c_1 > 0, c_2 > 1$, the Market for "Lemons" game has elimination length L_0 at least $2\lceil \frac{N}{k} \rceil - 1$ if $q_i - q_{i-k} \geq c_1 > q_i - q_{i-k+1}, \forall i \in \{k+1, k+2, \cdots, N\}$ and $\mathcal{P} \supseteq \{q_1, q_2, \ldots, q_N\}$. The buyer offering any price $p \leq q_1$ and each seller i setting $a_i = 0$ are Nash equilibria of the game.*

Proof. In this proof, we say a seller i remains on the market if his $a_i = 1$ is not eliminated. We start with the following two claims about the dominance elimination by the buyer and sellers.

Claim D.5. *A seller with quality q_i should remain on the market if and only if the buyer have not eliminated all its action $p \geq q_i + c_1$. Meanwhile, no seller i should eliminate his $a_i = 0$, as long as the buyer has not eliminate its $p = q_1$.*

Proof. By construction, if a seller choose not to list, his utility is always 0 in regardless of the action of other sellers or buyer.

($\Rightarrow$): With some $p \geq q_i + c_1$, the seller would have non-negative utility, $u_i = p - q_i - c_1 \geq 0$ if he choose to list. This is no worse than the zero utility if he does not list.

($\Leftarrow$): For any $p < q_i + c_1$, the seller with quality q_i have utility $u_i = \min(p - q_i, 0) - c_1 < 0$, always negative if he choose to list. Since this utility is strictly dominated by than the zero utility if he does not list, thus $a_i = 1$ should be eliminated.

Meanwhile, if the buyer sets a price $p = q_1$, any seller who list his car would incur a negative utility $-c_i$ strictly worse than the zero utility of not listing. So in this case no seller should always list its car. □

Claim D.6. *The buyer should eliminate all its action $p > q$ if the highest car quality of the remaining sellers on the market is q. Conversely, as long as a seller of quality q_i remains on the market, the buyer should not eliminate its action $p = q_i$.*

Proof. ($\Leftarrow$): Among the remaining sellers on the market, let the highest car quality be q. Consider the buyer sets some price $p > q$, cf. $p = q$. The outcome from the two different price are the same in the sense that the buyer can get all the cars that the sellers choose to list with quality average quality $\overline{q}$. We can see that the buyer's revenue $c_2\overline{q}$ is the same, but she has strictly minimum cost at $p = q$, therefore all the price $p > q$ are actions dominated by $p = q$.

($\Rightarrow$): Consider the situation that only the seller i choose to list. In this case, the best response of the buyer is to set her price at q_i. This is because she does not want to set a price above q_i, according to the argument in paragraph above; setting a price below q_i, the seller will not sell and her utility is 0, which is strictly worse than her utility $(c_2 - 1)q_i > 0$ if the buyer sets her price $p = q_i$. □

We now provide an induction argument for the iterative elimination:

Base case: In the beginning of elimination, we know for any seller i, $i \leq N - k$, has his quality $q_i \leq q_N - c_1$. They cannot eliminate any of their actions as the buyer have not eliminate its action $p = q_1$, or $p = q_N \geq q_i + c_1$, according to Claim D.5. Meanwhile, we show that it takes at least 1 round to eliminate both buyer's action(s) $p > q_N$ and the sellers' action $a_i = 1, \forall i > N - k$. This is because the buyer may or may not need first eliminate all her action $p > q_N$ depending on the support of $\mathcal{P}$, according to Claim D.6; In the same or the following round (also depending on the support of $\mathcal{P}$), all seller $i > N - k$ must eliminate their action $a_i = 1$, according to Claim D.5.

Inductive case: We show the following inductive statement: for $i \in [1, \ldots, \lceil \frac{N}{k} \rceil - 1]$, at the beginning of round $2i+1$, given that we have eliminated $\{a_j = 1 | \forall j > N - ik\} \cup \{p > q_{N-(i-1)k}\}$, then it takes two rounds to first eliminate all of the buyer's action(s) $p > q_{N-ik}$ and subsequently the all seller j's action $a_j = 1$ with $j > N - (i+1)k$.

Given that we have eliminated $\{a_j = 1 | \forall j > N - ik\} \cup \{p > q_{N-(i-1)k}\}$, in the first round, we can eliminate of the buyer's actions $p > q_{N-ik}$, according to Claim D.6, as the highest car quality of the remaining sellers on the market is q_{N-ik}. We cannot eliminate the action of any seller j with quality $q_j \leq q_{N-ik} \leq q_{N-(i-1)k} - c_1$, according to Claim D.5, as the buyer could offer a price $p = q_1$, or $p = q_{N-(i-1)k} \geq q_j + c_1$.

In the second round, we can eliminate all seller j's action $a_j = 1$ with $j > N - (i+1)k$, according to Claim D.5, as the buyer eliminates all of her actions above q_{N-ik}. Yet we still cannot eliminate the action of any other seller, according to Claim D.5, as the buyer could offer a price $p = q_1$, or $p = q_{N-ik} \geq q_j + c_1, \forall i \leq N - (i+1)k$. We also cannot eliminate any action of the buyer, as no seller exit the market in the last round.

Therefore, at the beginning of round $2i+3$, we expand the elimination set to $\{a_j = 1 | \forall j > N-(i+1)k\} \cup \{p > q_{N-ik}\}$, so the induction follows. It is easy to see that in the last iteration when $N - \lceil \frac{N}{k} \rceil k < 0$, all sellers exits the market, and all buyer's price above q_ν are eliminated, where we can compute $\nu = N + k - \lceil \frac{N}{k} \rceil k = \begin{cases} N \mod k, & k \nmid N \\ 1, & k \mid N \end{cases} \geq 1$. This terminates the iterative process of dominance elimination. The remaining action profiles must include all Nash equilibrium. Since all sellers have only one action left, their remaining action profile, i.e., not to list their cars, is the Nash equilibrium strategy. The buyer has utility zero for any of her actions $p \leq q_1 \leq q_\nu$, which are best responses to the sellers' equilibrium strategies and therefore form Nash equilibria along with the sellers' unique action profile.

As the base case takes at least one round to reach, and the induction stage takes $2\lceil \frac{N}{k} \rceil - 2$ round, in total, the elimination length is at least $2\lceil \frac{N}{k} \rceil - 1$. □