University of California
Los Angeles

Optimal Sequential Resource Sharing and Exchange in Multi-Agent Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Yuanzhang Xiao

2014

Abstract of the Dissertation

Optimal Sequential Resource Sharing and Exchange in Multi-Agent Systems

by

Yuanzhang Xiao

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2014

Professor Mihaela van der Schaar, Chair

Central to the design of many engineering systems and social networks is to solve the underlying resource sharing and exchange problems, in which *multiple* decentralized agents make *sequential* decisions over time to optimize some long-term performance metrics. It is challenging for the decentralized agents to make optimal sequential decisions because of the complicated coupling among the agents and across time. In this dissertation, we mainly focus on three important classes of multi-agent sequential resource sharing and exchange problems and derive optimal solutions to them.

First, we study multi-agent resource sharing with imperfect monitoring, in which *self-interested* agents have *imperfect* monitoring of the resource usage and inflict *strong negative* externality (i.e. strong interference and congestion) among each other. Despite of the imperfect monitoring, the strong negative externality, and the self-interested agents, we propose an optimal, distributed, easy-to-implement resource sharing policy that achieves Pareto optimal outcomes at the equilibrium. A key feature of the optimal resource sharing policy is that it is *nonstationary*, namely it makes decisions based on the history of past (imperfect) monitoring of the resource usages. The applications of our proposed design in wireless spectrum sharing problems enable us to improve the spectrum efficiency by up to 200% and achieve up to 90% energy saving, compared to state-of-the-art (stationary) spectrum sharing policies.

Second, we study multi-agent resource sharing with decentralized information, in which each agent has a *private*, independently and stochastically changing state (whose transition may depend on the agent's action), and the agents' actions are coupled through resource sharing constraints. Despite of the dentralized information (i.e. private states), we propose distributed resource sharing policies that achieve the social optimum, and apply the proposed policies to demand-side management in smart grids, and joint resource allocation and packet scheduling in wireless video transmissions. The proposed policies demonstrate significant performance gains over existing myopic policies that do not take into account the state dynamics and the policies based on Lyapunov optimization that were proposed for single-agent problems.

Finally, we study multi-agent resource exchange with imperfect monitoring, in which *self-interested*, *anonymous* agents exchange services (e.g. task solving in crowdsourcing platforms, file sharing in peer-to-peer networks, answering in question-and-answer forums). Due to the anonymity of the agents and the lack of fixed partners, free-riding is prevalent, and can be addressed by rating protocols. We propose the *first* rating protocol that can achieve the social optimum at the equilibrium under *imperfect monitoring* of the service quality. A key feature of the optimal rating protocol is again that it is *nonstationary*, namely it recommends desirable behaviors based on the history of past rating distributions of the agents.

The dissertation of Yuanzhang Xiao is approved.

Paulo Tabuada

Lieven Vandenberghe

William Zame

Mihaela van der Schaar, Committee Chair

University of California, Los Angeles

2014

To my parents and my wife

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Roadmap 2

1.2.1 Resource Sharing With Imperfect Monitoring 2

1.2.2 Resource Sharing With Decentralized Information 3

1.2.3 Resource Exchange With Imperfect Monitoring 4

2 Resource Sharing With Imperfect Monitoring **5**

2.1 Introduction 5

2.1.1 Related Works in Repeated Game Theory Literature . . . 6

2.2 Model 9

2.2.1 Stage Game 9

2.2.2 The Repeated Game with Imperfect Public Monitoring . . 10

2.2.3 Assumptions on the Stage Game 11

2.2.4 Assumptions on the Monitoring Structure 12

2.3 Main Results 13

2.3.1 Ruling out Some Efficient PPE Payoffs 13

2.3.2 Characterizing Efficient Equilibrium Payoffs 16

2.3.3 Constructing Efficient Perfect Public Equilibria 21

2.3.4 Robustness 22

2.3.5 Two Players 24

2.4 Applications to Throughput-Maximizing Spectrum Sharing 25

2.4.1 Motivation 25
2.4.2 Related Works 27
2.4.3 Model and Problem Formulation 30
2.4.4 Link to The General Model in Section 2.2 36
2.4.5 Simulation Results 38
2.5 Applications and Extensions to Energy-Efficient Spectrum Sharing 44
2.5.1 Motivation 44
2.5.2 Related Works 46
2.5.3 Model and Problem Formulation 47
2.5.4 Simulation Results 48
2.6 Conclusion 51
2.7 Appendix 52

3 Resource Sharing With Decentralized Information 68
3.1 Introduction 68
3.2 Related Works 72
3.2.1 Related Works on Demand-Side Management 72
3.2.2 Related Theoretical Frameworks 74
3.3 System Model 75
3.3.1 The System Setup 75
3.3.2 The DSM Strategy 81
3.4 The Design Problem 82
3.5 Optimal Foresighted Demand Side Management 83
3.5.1 The aggregator's Decision Problem and Its Conjectured Price 83
3.5.2 The Optimal Decentralized DSM Strategy 86

3.5.3 Learning Unknown Dynamics 89

3.5.4 Detailed Comparisons with Existing Frameworks 90

3.6 Simulation Results 91

3.6.1 Learning and Tracking The Optimal Policies Without Knowledge of State Transition Probabilities 93

3.6.2 Performance Evaluation 95

3.6.3 The Conjectured Prices 99

3.6.4 Comparisons of the DSM strategies 100

3.7 Conclusion 102

3.8 Appendix 103

4 Resource Exchange With Imperfect Monitoring 106

4.1 Introduction 106

4.2 Related Works 110

4.2.1 Related Works on Rating Protocols 110

4.2.2 Related Works in Game Theory Literature 112

4.2.3 Related Mathematical Frameworks 113

4.3 System Model and Problem Formulation 115

4.3.1 System Model 115

4.3.2 The Rating Protocol Design Problem 123

4.4 Sources of Inefficiency 124

4.4.1 Stationary Recommended Strategies 125

4.4.2 Lack of Differential Punishments 128

4.5 Socially Optimal Design 129

4.5.1 The APS Technique 129

4.5.2 Socially Optimal Design 130

4.5.3 Whitewashing-Proofness 140

4.6 Simulation Results 141

4.7 Conclusion 145

4.8 Appendix 146

4.8.1 Proof of Proposition 4 146

4.8.2 Analytical Expression of The Lower-Bound Discount Factor 152

4.8.3 Proof of Theorem 7 153

4.8.4 Complete Description of the Algorithm 166

5 Concluding Remarks 168

References 170

List of Figures

2.1 $\mu = (0, 1/4, 0)$; V_μ is regular 18

2.2 $\mu = (1/2, 1/2, 1/2)$; V_μ is regular 19

2.3 $\mu = (1/4, 0, 0)$; V_μ is not regular 19

2.4 An illustration of the best operating points achievable by different policies in a two-SU spectrum sharing system. 29

2.5 An example system model with two secondary users. The solid line represents a link for data transmission, and the dashed line indicate a link for control signals. The channel gains for the corresponding data link are written in the figure. The primary user (PU) specifies the interference temperature (IT) limit to the local spectrum server (LSS). The LSS sets the intermediate IT limit to the secondary users and send distress signals if the estimated interference power exceeds the IT limit. 31

2.6 The procedure of solving the design problem. 38

2.7 Performance comparison of the proposed policy and the optimal policy with constant power levels ('stationary' in the legend) under different numbers of users and different cross interference levels. A zero average throughput indicates that there exists no feasible policy that satisfies all the constraints in the policy design problem. 39

2.8 Performance comparison of the proposed policy and the punish-forgive policy with the optimal punishment length under different error variances and different false alarm probabilities. 41

2.9 The impact of the variance of the measurement error on the performance of the proposed policy and the minimum discount factor required under which the proposed policy is deviation-proof. . . . 43

2.10 The impact of the false alarm probability on the performance of the proposed policy and the minimum discount factor required under which the proposed policy is deviation-proof. 43

2.11 50

2.12 Dynamics of average energy consumption with users entering and leaving the network. At $t = 0$, there are 10 PUs and 2 SUs. SU 2 leaves at $t = 100$. SU 3 enters at $t = 150$. PU 11 enters at $t = 200$. SUs 4–8 enter at $t = 250$. We only show PUs 1, 5, 9, 11 (solid lines) and SUs 1, 2, 3, 4 (dashed lines) in the figure. 50

2.13 Dynamics of average throughput under the same dynamics of the entry and exit of users as in Fig. 2.12. 51

2.14 Illustrations of the feasible continuation payoffs when $\kappa_{ij}^{+} \leq 0$. $\bar{\gamma}_j = \frac{1}{\lambda_j}\left(1 - \sum_{k \neq j} \lambda_k \mu_k\right)$. 60

3.1 The system model of the smart grid. The information flow to the ISO is denoted by red dashed lines, the information flow to the aggregators is denoted by black dotted lines, and the information flow sent from the ISO is denoted by blue dash-dot lines. 76

3.2 Illustration of the interaction between the ISO and aggregator i (i.e. their decision making and information exchange) in one period. 78

3.3 Illustration of the entities' decision making and information exchange in the design framework based on conjectured prices. . . . 85

3.4 Convergence of the PDS-based learning algorithm. The two aggregators' average long-term costs converge to the optimal one under the PDS-based learning algorithm. 94

3.5 The PDS-based learning algorithm tracks the optimal solution when the underlying distribution of the energy generation is time-varying. 94

3.6 The normalized total cost per hour versus the capacity of the energy storage in the IEEE 14-bus system. 96

3.7 The normalized total cost per hour versus the capacity of the energy storage in the IEEE 30-bus system. 96

3.8 The normalized total cost per hour versus the capacity of the energy storage in the IEEE 118-bus system. 96

3.9 The normalized total cost per hour versus the uncertainty in renewable energy generation in the IEEE 14-bus system. The aggregators have energy storage of capacity 25 MW and 50 MW, respectively. 98

3.10 The average individual costs of the aggregators with different energy storage in the IEEE 118-bus system. 99

3.11 The conjectured prices and the real expected locational marginal prices (LMPs) in IEEE 14-bus and 118-bus systems. 100

4.1 Illustration of the rating mechanism in one period. 115

4.2 Illustration of the rating update rule. The circle denotes the rating, and the arrow denotes the rating update with corresponding probabilities. 119

4.3 Illustration of how to build the self-generating set. The left figure shows the set of feasible payoffs in one state (i.e. under the rating distribution (s_0, s_1)). The right figure shows the sets of feasible payoffs in different states (i.e. rating distributions) and their intersection, namely the set of common feasible payoffs in all the states (i.e. under all the rating distributions). 134

4.4 Illustration of the self-generating set, which is a triangle within the set of common feasible payoffs in Fig. 4.3. 134

4.5 The decomposition of payoffs. The left figure shows the decomposition in period 0, when the payoff to decompose is the target payoff $(b - c - \epsilon_0, b - c - \epsilon_1)$; the right figure shows the decomposition in period 1, when the payoff to decompose is the continuation payoff starting from period 1 and when the users have different ratings. . 136

4.6 Illustration of how to choose the plan in order to decompose a given payoff. Each self-generating set is partitioned into two parts. In each period, a recommended plan (the altruistic plan "a", the fair plan "f", or the selfish plan "s") is chosen, depending on which part of the self-generating set the expected payoffs fall into. 137

4.7 Illustration of the target payoff of a rating-1 user and the lowest continuation payoff of a rating-0 user. 141

4.8 Evolution of states and recommended plans taken in different rating mechanisms. 143

4.9 Price of stationarity of different stationary rating mechanisms under different rating update errors. 143

4.10 Lower-bound discount factors under different performance loss tolerances and rating update errors. 144

4.11 Performance gain/loss (in percentage) under different inaccuracy of estimation (in percentage). 145

List of Tables

2.1 The algorithm used by each player. 23

2.2 Comparison With Related Works In Dynamic Spectrum Sharing. 29

2.3 Comparisons against stationary policies. 46

2.4 Comparisons against nonstationary policies. 47

3.1 Comparisons With Related Works on Demand-Side Management. 73

3.2 Comparisons With Related Mathematical Frameworks. 74

3.3 Each entity's knowledge, and the corresponding results. 77

3.4 Distributed algorithm to compute the optimal decentralized DSM strategy. 87

3.5 Information needed by each entity to implement the algorithm. . . 88

3.6 Relationship between the proposed and existing theoretical frameworks. 91

3.7 Comparisons of different strategies. 102

4.1 Related Works on Rating Protocols. 110

4.2 Related Mathematical Frameworks. 114

4.3 Gift-Giving Game Between A Client and A Server. 120

4.4 Normalized social welfare of stationary strategies restricted on $A^{\mathrm{a}fs}$. 127

4.5 Minimum punishment probabilities of rating mechanisms restricted on $A^{\mathrm{a}fs}$ when $\varepsilon = 0.1$. 128

4.6 Algorithm to construct recommended strategies. 139

4.7 The algorithm of constructing the equilibrium strategy by the rating mechanism. 167

Acknowledgments

First of all, I would like to thank my advisor, Prof. Mihaela van der Schaar, without whom this thesis definitely does not exist. I benefited and learned tremendously from her passion and enthusiasm for high-quality research, her unique taste of excellent research topics, and her perpetual energy. She gave me a lot of freedom to pursue the research that interests me, and tolerated me for my mistakes and my detours.

I would also like to thank Prof. William Zame, without whom the first part of this thesis does not exist. I am amazed by his sheer intelligence, and the way he think about a research problem and ask inspiring questions.

I thank the other committee members, Prof. Paolo Tabuada for giving me tough time during the qualifying exam, which prepares me for "stepping up my game", and Prof. Lieven Vandenberghe for his great class on convex optimization.

I thank the colleagues, labmates, and friends in UCLA: Dr. Yi Su, Dr. Fangwen Fu, Dr. Nick Mastronarde, Dr. Shaolei Ren, Dr. Yu Zhang, Khoa Phan, Dr. Luca Canzian, Jie Xu, Siming Song, Jianyu Wang, Linqi Song, Kartik Ahuja, Xiaochuan Zhao, Jianshu Chen, Ming Zhang. They help me in many aspects in my research and my life, and make my PhD life much more colorful.

Finally, I would like to thank my lovely, beautiful, and intelligent wife, Wenjing, and my parents. Their love, support, and encouragement get me through my PhD career. No word in the world can express my gratitude for them.

Vita

2006	B.E. (Electronic Engineering), Tsinghua University.
2009	M.E. (Electronic Engineering), Tsinghua University.
2009–present	Research Assistant, Electrical Engineering Department, UCLA.

Publications

[9] Linqi Song, **Yuanzhang Xiao**, and Mihaela van der Schaar, "Demand Side Management in Smart Grids using a Repeated Game Framework," To appear in *IEEE J. Sel. Areas in Commun., Special Issue on Smart Grid Communications.* Available at: http://arxiv.org/abs/1311.1887

[8] **Yuanzhang Xiao** and Mihaela van der Schaar, "Optimal Foresighted Multi-User Wireless Video," Accepted with minor revision by *IEEE J. Sel. Topics Signal Process., Special issue on Visual Signal Processing for Wireless Networks.* Available at: http://arxiv.org/abs/1311.4227

[7] **Yuanzhang Xiao** and Mihaela van der Schaar, "Energy-efficient Nonstationary Spectrum Sharing," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 810-821, Mar. 2014. Available at: http://arxiv.org/abs/1211.4174

[6] Jie Xu, Yiannis Andreopoulos, **Yuanzhang Xiao** and M. van der Schaar, "Non-stationary Resource Allocation Policies for Delay-constrained Video Stream-

ing: Application to Video over Internet-of-Things-enabled Networks," emphIEEE J. Sel. Areas in Commun., Special Issue on Adaptive Media Streaming, vol. 32, no. 4, pp. 782-794, Apr. 2014.

[5] Luca Canzian, **Yuanzhang Xiao**, William Zame, Michele Zorzi, Mihaela van der Schaar, "Intervention with Private Information, Imperfect Monitoring and Costly Communication: Design Framework," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3192–3205, Aug. 2013.

[4] Luca Canzian, **Yuanzhang Xiao**, William Zame, Michele Zorzi, Mihaela van der Schaar, "Intervention with Complete and Incomplete Information: Application to Flow Control," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3206–3218, Aug. 2013.

[3] **Yuanzhang Xiao** and Mihaela van der Schaar, "Dynamic Spectrum Sharing Among Repeatedly Interacting Selfish Users With Imperfect Monitoring," *IEEE J. Sel. Areas Commun., Special issue on Cognitive Radio Series*, vol. 30, no. 10, pp. 1890–1899, Nov. 2012.

[2] **Yuanzhang Xiao**, Jaeok Park, and Mihaela van der Schaar, "Repeated Games With Intervention: Theory and Applications in Communications," *IEEE Trans. Commun.*, vol. 60, no. 10, pp. 3123–3132, Oct. 2012.

[1] **Yuanzhang Xiao**, Jaeok Park, and Mihaela van der Schaar, "Intervention in Power Control Games with Selfish Users," *IEEE J. Sel. Topics Signal Process., Special issue on Game Theory In Signal Processing*, vol. 6, no. 2, pp. 165–179, Apr. 2012.

CHAPTER 1

Introduction

1.1 Motivation

This thesis focuses on the optimal design of multi-agent systems, in which decentralized autonomous agents with conflicting objectives and coupled constraints either share a common resource or exchange resources among each other. Examples of resource sharing systems range from classic resource sharing problems in Electrical Engineering such as power control, medium access control, flow control, workload and task scheduling etc., to emerging new problems such as demand-side management in smart electric power grids and resource allocation in cloud data centers. Examples of resource exchange systems include peer-to-peer (P2P) networks such as BitTorrent, in which agents exchange data, and social crowdsourcing platforms such as Amazon Mechanical Turk, in which agents exchange labor or services.

These seemingly-different systems share (some of) the following important common features: the negative externalities among the agents caused by the interference and congestion, the imperfect monitoring abilities of the agents (i.e. they cannot perfectly observe the resource usage status in the system or the quality of the resources provided by the other agents), the information decentralization among the agents (i.e. each agent may have private local information unknown to the others) and their selfish behaviors. Because of these important features, the optimal resource sharing/exchange policy should be nonstationary, namely the action should depend not only on the current state of the system, but also on the history of past states. However, the theory of designing nonstationary policies is still very under-developed, due to the difficulty involved in analyzing and designing the highly-complicated nonstationary policies. Nevertheless, it is of great importance to develop the theory of designing nonstationary policies, because of the inefficiency of stationary policies (i.e. policies in which the action depends only on the current state and this dependence is time-invariant). In our recent works on cognitive radio networks [2][3], social crowdsourcing and P2P system [7], smart grids [4], we have shown that nonstationary policies significantly outperform state-of-the-art stationary policies.

1.2 Roadmap

In this thesis, we focus on three important classes of multi-agent sequential resource sharing and exchange problems and derive optimal solutions to them.

1.2.1 Resource Sharing With Imperfect Monitoring

In the first part of this thesis (based on works [1][2][3]), we study multi-agent resource sharing with imperfect monitoring, in which *self-interested* agents have *imperfect* monitoring of the resource usage and inflict *strong negative* externality (i.e. strong interference and congestion) among each other. Despite of the imperfect monitoring, the strong negative externality, and the self-interested agents, we propose an optimal, distributed, easy-to-implement resource sharing policy that achieves Pareto optimal outcomes at the equilibrium. A key feature of the optimal resource sharing policy is that it is *nonstationary*, namely it makes decisions based on the history of past (imperfect) monitoring of the resource usages.

Our framework can be applied to a variety of important engineering applications such as power control, flow control, and demand-side management in smart grids. In wireless communication networks, our proposed nonstationary spectrum sharing policies can improve the spectrum efficiency by up to 200% [2], and reduce the energy consumption by up to 90% while achieving the same spectrum efficiency [3], compared to the state-of-the-art spectrum sharingpolicies. In smart electric power grids [4], our proposed optimal design can reduce the operational costs of the grids by up to 60% by optimal demand-side management and optimal usage of energy storage.

1.2.2 Resource Sharing With Decentralized Information

In the second part of this thesis (based on works [5][6]), we study multi-agent resource sharing with decentralized information, in which each agent has a *private*, independently and stochastically changing state (whose transition may depend on the agent's action), and the agents' actions are coupled through resource sharing constraints. Despite of the dentralized informtion (i.e. private states), we propose distributed resource sharing policies that achieve the social optimum, and apply the proposed policies to demand-side management in smart grids, and joint resource allocation and packet scheduling in wireless video transmissions. The proposed policies demonstrate significant performance gains over existing myopic policies that do not take into account the state dynamics and the policies based on Lyapunov optimization that were proposed for single-agent problems.

We apply our framework in demand side management and economic dispatch in smart grids with energy storage. Our proposed optimal policy achieves significant reduction in the total system operating cost, compared to the optimal myopic demand side management and economic dispatch (up to 60% reduction), and the foresighted demand side management and economic dispatch based on the Lyapunov optimization framework (up to 30% reduction).

1.2.3 Resource Exchange With Imperfect Monitoring

In the third part of this thesis (based on works [7]), we study multi-agent resource exchange with imperfect monitoring, in which a large population of *self-interested*, *anonymous* agents exchange services (e.g. task solving in crowdsourcing platforms, file sharing in peer-to-peer networks, answering in question-and-answer forums). In such systems, the absence of a fixed partner and the anonymity of the agents create an incentive problem, namely the agents tend to free-ride (for instance, in P2P systems, agents tend to download without uploading). In addition, a agent generally may not be able to perfectly monitor its partners action, either due to the agents inaccurate assessment of its partners action (e.g., in a crowdsourcing system, the client, who wants to translate something into a foreign language, cannot accurately evaluate the servers translation), or due to some system errors (e.g., in a P2P system, the client does not receive the serves data due to network errors). Most existing resource exchange policies are designed under the assumption of perfect monitoring, and will lead to a system collapse if monitoring is even slightly imperfect. A few recent works proposed stationary policies under imperfect monitoring, which have over 50% performance loss compared to the social optimum. In [7], we proposed the first nonstationary resource exchange policy that can achieve the social optimum even under imperfect monitoring.

CHAPTER 2

Resource Sharing With Imperfect Monitoring

2.1 Introduction

The problem of efficient sharing of a resource is nearly ubiquitous. Unless the resource is a pure public good, each agent's use of the resource imposes a negative externality on other users. Hence, self-interested, strategic agents will find it difficult to share the resource efficiently, at least in the short run. In some circumstances – those we focus on in this chapter – the negative externality is so strong – competition for the resource is so destructive – that it will be *impossible* for users to share the resource efficiently, at least in the short run. The purpose of this chapter is to propose resource sharing policies that are efficient in the long run – even when outcomes depend stochastically on actions, monitoring is very limited and players are not very patient.

We formulate the resource sharing scenario using the framework of repeated games with imperfect public monitoring. Within our framework, we abstract what we see as the essential features of the resource allocation problems by two assumptions about the stage game. The first is that for each player i there is a unique action profile $\tilde{\boldsymbol{a}}^i$ that i most prefers. (In many resource sharing scenarios, $\tilde{\boldsymbol{a}}^i$ would be the profile in which only player i accesses the resource.) The second is that for every action profile $\boldsymbol{a}$ that is *not* in the set $\{\tilde{\boldsymbol{a}}^i\}$ of preferred action profiles the corresponding utility profile $u(\boldsymbol{a})$ lies *below* the hyperplane H spanned by the utility profiles $\{u(\tilde{\boldsymbol{a}}^i)\}$. (In many resource sharing scenarios, this corresponds to the assumption that allowing access to the resource by more than one individual strictly lowers (weighted) social welfare.) We capture the notion that monitoring is very limited by assuming that players do not observe the profile $\boldsymbol{a}$ of actions but rather only some signal $y \in Y$ whose distribution $\rho(y|\boldsymbol{a})$ depends on the true profile $\boldsymbol{a}$, and that (profitable) single-player deviations from i's preferred action profile $\tilde{\boldsymbol{a}}^i$ can be statistically distinguished from conformity with $\tilde{\boldsymbol{a}}^i$ in the same way. (But we do not assume that different deviations from $\tilde{\boldsymbol{a}}^i$ can be distinguished from *each other*.) We emphasize the setting in which there are only two signals – "good" and "bad" – because this setting offers the sharpest results and the clearest intuition and, as we shall see, because two signals are often enough.

Our results are different from existing results in the repeated game theory literature in three important aspects: we do not assume a rich signal structure (rather, we require only two signals), we do not assume players are arbitrarily patient (rather, we find an explicit lower bound on the requisite discount factor), and we provide an *explicit (distributed) algorithm* that takes as inputs the parameters – stage game payoffs, discount factor, target payoff – and computes the strategy – the action to be chosen by each player following each public history. This algorithm can be carried out by each player separately and in real time – there is no need for the designer to specify/describe the strategies to be played. A consequence of our constructive algorithm is that the strategies we identify enjoy a useful robustness property: generically, the equilibrium strategies are, for many periods, locally constant in the parameters of the environment and of the problem.

2.1.1 Related Works in Repeated Game Theory Literature

As discusses before, we model the resource sharing scenarios as repeated games with imperfect public monitoring. The literature on repeated games with imperfect public monitoring is quite large – much too large to survey here; we refer instead to [8] and the references therein. However, explicit comparisons with two papers in this literature may be especially helpful. The first and most obvious comparison is with [9] by Fudenberg, Levine, and Maskin (hereafter FLM) on the Folk Theorem for repeated games with imperfect public monitoring. As do we, FLM consider a situation in which a single stage game G with action space $\boldsymbol{A}$ and utility function $u : \boldsymbol{A} \to \mathbb{R}^n$ is played repeatedly over an infinite horizon; monitoring is public but imperfect, so players do not observe actions but only a public signal of those actions. In this setting, $\mathrm{co}[u(\boldsymbol{A})]$ is the set of payoff profiles that can be achieved as long run average utilities for *some* discount factor and *some* infinite set of plays of the stage game G. Under certain assumptions, FLM prove that any payoff vector in the interior of $\mathrm{co}[u(\boldsymbol{A})]$ that is strictly individually rational can be achieved in a PPE of the infinitely repeated game. However, the assumptions FLM maintain are very different from ours in two very important dimensions (and some other dimensions that seem less important, at least for the present discussion). The first is that the signal structure is rich and informative; in particular, that the number of signals is at least one less than the number of actions of any two players. The second is that players are arbitrarily patient: that is, the discount factor δ is as close to 1 as we like. (More precisely: given a target utility profile v, there is some $\delta(v)$ such that if the discount factor $\delta > \delta(v)$ then there is a PPE of the repeated game that yields the target utility profile v.) In particular, FLM do not identify any PPE for any *given* discount factor $\delta < 1$. By contrast, we require only two signals *even if action spaces are infinite* and we do *not* assume players are patient: all target payoffs can be achieved for some *fixed* discount factor – which may be very far from 1. Moreover, because FLM consider only payoffs in the interior of $\mathrm{co}[u(\boldsymbol{A})]$, they have nothing to say about achieving *efficient* payoffs. Their results do imply that efficient payoffs can be arbitrarily well approximated by payoffs that can be achieved in PPE, but only if the corresponding discount factors are arbitrarily close to 1.

By contrast, Fudenberg, Levine, and Takahashi [21] (hereafter FLT) *do* show

how (some) efficient payoffs can be achieved in PPE. Given Pareto weights $\lambda_1, \ldots, \lambda_n$ set $\Lambda = \sup\{\sum \lambda_i u_i(\boldsymbol{a}) : \boldsymbol{a} \in \boldsymbol{A}\}$ and consider the hyperplane $H = \{x \in \mathbb{R}^n : \sum \lambda_i x_i = \Lambda\}$. The intersection $H \cap \mathrm{co}[u(\boldsymbol{A})]$ is a part of the Pareto boundary of $\mathrm{co}[u(\boldsymbol{A})]$. As do we, FLT ask what vectors in $H \cap \mathrm{co}[u(\boldsymbol{A})]$ can be achieved in PPE of the infinitely repeated game. They identify the largest (compact convex) set $Q \subset H \cap \mathrm{co}[u(\boldsymbol{A})]$ with the property that every target vector $v \in \mathrm{int}Q$ (the relative interior of Q with respect to H) can be achieved in a PPE of the infinitely repeated game for *some* discount factor $\delta(v) < 1$. However, because FLT consider arbitrary stage games and arbitrary monitoring structures, the set Q identified by FLT may be empty, and FLT do not provide any conditions that guarantee that Q is not empty. Moreover, as in FLM, FLT assume that players are arbitrarily patient, so do not identify any PPE for any *given* discount factor $\delta < 1$. Having said this, we should also point out that FLT identify the closure of the set of *all* payoff vectors in the interior of $H \cap \mathrm{co}[u(\boldsymbol{A})]$ that can be achieved in a PPE for some discount factor, while we identify only some. So there is a trade-off: FLT find more PPE payoffs but provide much less information about the ones they find; we find fewer PPE payoffs but provide much more information about the ones we find.

At the risk of repetition, we want to emphasize the most important features of our results. The first is that we do not assume discount factors are arbitrarily close to 1. The importance of this seems obvious in all environments – especially since the discount factor encodes both the innate patience of players *and* the probability that the interaction continues. The second is that we impose different – and in many ways weaker – requirements on the monitoring structure; indeed, we require only two signals, even if action spaces are infinite. Again, the importance of this seems obvious in all environments, but especially in those in which signals are not generated by some exogenous process but must be provided by a designer. In the latter case it seems obvious – and in practice may be of supreme importance – that the designer may wish or need to choose a simple information structure that employs a small number of signals, saving on the cost of observing the outcome of play and on the cost of communicating to the agents. More generally, the designer may face a trade-off between the efficiency obtainable with a finer information structure and the cost of using that information structure. Finally, because we provide a distributed algorithm for calculating equilibrium play, neither the agents nor a designer need to work out the equilibrium strategies in advance; all calculations can be done online, in real time.

The rest of this chapter is organized as follows. Section 2.2 presents the formal model. Section 2.3 presents our main results: we first give some preliminary results, presenting conditions under which *no* efficient payoffs can be achieved in PPE for *any* discount factor; we then presents the main technical result (Theorem 1); we finally presents the implications for PPE (Theorems 2,3). In Section 2.4, we apply our theoretical results to maximize the throughput in spectrum sharing scenarios. In Section 2.5, we apply and extend our theoretical results to minimize the energy consumption while fulfilling minimum throughput requirements in spectrum sharing scenarios. Section **??** concludes this chapter. We relegate all proofs to the Appendix.

2.2 Model

2.2.1 Stage Game

The stage game consists of

- a set $N = \{1, \ldots, n\}$ of players
- for each player i
  - a compact set A_i of actions

  - a continuous utility function $u_i : \boldsymbol{A} = A_1 \times \cdots \times A_n \to \mathbb{R}$

- a compact set of public signals Y

- a measurable map $\boldsymbol{a} \mapsto \rho(\cdot|\boldsymbol{a}) : \boldsymbol{A} \to \Delta(Y)$

We interpret $u_i(\boldsymbol{a})$ as i's *ex ante* (expected) utility when $\boldsymbol{a}$ is played and $\rho(y|\boldsymbol{a})$ as the probability that the signal y is observed when $\boldsymbol{a}$ is played.

2.2.2 The Repeated Game with Imperfect Public Monitoring

In the repeated game, the stage game G is played in every period $t = 0, 1, 2, \ldots$. In each period t, the moves are made in the following order:

- Each player i chooses its action a_i^t.

- The public signal y^t is realized and observed by all the players.

- Each player i receives its utility u_i^t.

A *public history* of length t is a sequence $(y^0, y^1, \ldots, y^{t-1}) \in Y^t$. We write $\mathcal{H}(t)$ for the set of public histories of length t, $\mathcal{H}^T = \bigcup_{t=0}^{T} \mathcal{H}(t)$ for the set of public histories of length at most T and $\mathcal{H} = \bigcup_{t=0}^{\infty} \mathcal{H}(t)$ for the set of all public histories of all finite lengths. A *private history* for player i includes the public history, the actions taken by player i, and the realized utilities observed by player i, so a *private history* of length t is a a sequence $(a_i^0, \ldots, a_i^{t-1}; u_i^{*,0}, \ldots, u_i^{*,t-1}; y^0, \ldots, y^{t-1}) \in A_i^t \times \mathbb{R}^t \times Y^t$. We write $\mathcal{H}_i(t)$ for the set of i's private histories of length t, $\mathcal{H}_i^T = \bigcup_{t=0}^{T} \mathcal{H}_i(t)$ for the set of i's private histories of length at most T and $\mathcal{H}_i = \bigcup_{t=0}^{\infty} \mathcal{H}_i(t)$ for the set of i's private histories of all finite lengths.

A *pure strategy* for player i is a mapping from all private histories into the set of pure actions $\pi_i : \mathcal{H}_i \to A_i$. A *public strategy* for player i is a pure strategy that is independent of i's own action/utility history; equivalently, a mapping from public histories to i's pure actions $\pi_i : \mathcal{H} \to A_i$.

We assume all players discount future utilities using the same discount factor $\delta \in (0,1)$ and we use long-run averages, so if the stream of expected utilities is $\{u^t\}$ the vector of long-run average utilities is $(1-\delta)\sum_{t=0}^{\infty} \delta^t u^t$. A strategy profile $\boldsymbol{\pi} : \mathcal{H}_1 \times \ldots \times \mathcal{H}_n \to \boldsymbol{A}$ induces a probability distribution over public and private histories and hence over *ex ante* utilities. We write $U(\boldsymbol{\pi})$ for the vector of expected (with respect to this distribution) long-run average *ex ante* utilities when players follow the strategy profile $\boldsymbol{\pi}$.

As usual a strategy profile $\boldsymbol{\pi}$ is an *equilibrium* if each player's strategy is optimal given the strategies of others. A strategy profile is a *public equilibrium* if it is an equilibrium and each player uses a public strategy; it is a *perfect public equilibrium (PPE)* if it is a public equilibrium following every public history.

2.2.3 Assumptions on the Stage Game

To this point we have described a very general setting; we now impose additional assumptions – first on the stage game and then on the information structure – that we exploit in our results.

Set $u(\boldsymbol{A}) = \{u(\boldsymbol{a}) \in \mathbb{R}^n : \boldsymbol{a} \in \boldsymbol{A}\}$ and let $\mathrm{co}[u(\boldsymbol{A})]$ be the convex hull of $u(\boldsymbol{A})$. For each i set

$$\begin{aligned} \tilde{v}_i^i &= \max_{\boldsymbol{a} \in \boldsymbol{A}} u_i(\boldsymbol{a}) \\ \tilde{\boldsymbol{a}}^i &= \arg\max_{\boldsymbol{a} \in \boldsymbol{A}} u_i(\boldsymbol{a}) \end{aligned}$$

Compactness of the action space $\boldsymbol{A}$ and continuity of utility functions u_i guarantee that $u(\boldsymbol{A})$ and $\mathrm{co}[u(\boldsymbol{A})]$ are compact, that $\tilde{v}_i^i$ is well-defined and that the arg max is not empty. For convenience, we assume that the arg max is a singleton; i.e., the maximum utility $\tilde{v}_i^i$ for player i is attained at a *unique* strategy profile $\tilde{\boldsymbol{a}}^i$.[1] We refer to $\tilde{\boldsymbol{a}}^i$ as i's *preferred action profile* and to $\tilde{v}^i = u(\tilde{\boldsymbol{a}}^i)$ as i's *preferred utility profile*. In the context of resource sharing, $\tilde{\boldsymbol{a}}^i$ will typically be the (unique) action

[1] This assumption could be avoided, at the expense of some technical complication.

profile at which agent i has optimal access to the resource and other agents have none. For this reason, we will often say that i is *active* at the profile $\tilde{\boldsymbol{a}}^i$ and other players are *inactive.* Set $\tilde{\boldsymbol{A}} = \{\tilde{\boldsymbol{a}}^i : i \in N\}$ and $\tilde{V} = \{\tilde{v}^i : i \in N\}$ and write $V = \mathrm{co}\,(\tilde{V})$ for the convex hull of $\tilde{V}$. Note that $\mathrm{co}(u(\boldsymbol{A}))$ is the convex hull of the set of vectors that can be achieved – for *some* discount factor – as long-run average *ex ante* utilities of repeated plays of the game G (not necessarily equilibrium plays of course) and that V is the convex hull of the set of vectors that can be achieved – for *some* discount factor – as long-run average *ex ante* utilities of repeated plays of the game G in which only actions in $\tilde{\boldsymbol{A}}$ are used. We refer to $\mathrm{co}[u(\boldsymbol{A})]$ as the set of *feasible payoffs* and to V as the set of *efficient payoffs*.[2]

We abstract the motivating class of resource allocation problems by imposing conditions on the set of preferred utility profiles, which abstract the idea that there are strong negative externalities.

Assumption 1 The affine span of $\tilde{V}$ is a hyperplane H and all *ex ante* utility vectors of the game other than the those in $\tilde{V}$ lie below H. That is, there are weights $\lambda_1, \ldots, \lambda_n > 0$ such that $\sum \lambda_j u_j(\tilde{\boldsymbol{a}}^i) = 1$ for each i and $\sum \lambda_j u_j(\boldsymbol{a}) < 1$ for each $\boldsymbol{a} \in \boldsymbol{A}, \boldsymbol{a} \notin \tilde{\boldsymbol{A}}$.[3]

2.2.4 Assumptions on the Monitoring Structure

As noted in the Introduction, we focus on the case in which there are only two signals.

Assumption 2 The set Y contains precisely two signals and $\rho(y|\boldsymbol{a}) > 0$ for every $y \in Y$ and $\boldsymbol{a} \in \boldsymbol{A}$. (The monitoring structure has *full support.*)

We assume that profitable deviations from the profiles $\tilde{\boldsymbol{a}}^i$ exist and be statis-

[2] The latter is a slight abuse of terminology: because V is the intersection of the set of feasible payoffs with a bounding hyperplane, every payoff vector in V is Pareto efficient and yields maximal weighted social welfare and other feasible payoffs yield lower weighted social welfare – but other feasible payoffs might also be Pareto efficient.

[3] That the sum is 1 is just a normalization.

tically detected in a particularly simple way.

Assumption 3 For each $i \in N$ and each $j \neq i$ there is an action $a_j \in A_j$ such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$. Moreover, there is a labeling $Y = \{y^i_g, y^i_b\}$ with the property that for all $j \neq i$,

$$a_j \in A_j, u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i) \Rightarrow \rho(y^i_g | a_j, \tilde{\boldsymbol{a}}^i_{-j}) < \rho(y^i_g |, \tilde{\boldsymbol{a}}^i).$$

That is, given that other players are following $\tilde{\boldsymbol{a}}^i$, any strictly profitable deviation by player j strictly reduces the probability that the "good" signal y^i_g is observed (equivalently: strictly increases the probability that the "bad" signal y^i_b is observed).

The import of Assumption 3 is that all profitable single player deviations from $\tilde{\boldsymbol{a}}^i$ alter the signal distribution in the *same direction* although perhaps not to the same extent. We allow for the possibility that non-profitable deviations may not be detectable in the same way – perhaps not detectable at all – and for the possibility that which signal is "good" and which is "bad" depend on the identity of the *active* player i.

2.3 Main Results

In this section, we find conditions – on the discount factor among other things – that enable us to construct PPE that achieve payoffs in V (efficient payoffs). We also propose simple distributed algorithms to construct PPE. We end this section by presenting the robustness property of the constructed PPE.

2.3.1 Ruling out Some Efficient PPE Payoffs

We first show that under certain conditions, certain efficient payoffs *cannot* be achieved in PPE no matter what the discount factor is. To this end, we identify two measures of benefits from deviation. (These same measures will play a prominent role in the next Section as well.) Given $i, j \in N$ with $i \neq j$ set:

$$\begin{aligned}
\alpha(i,j) &= \sup\left\{ \frac{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - u_j(\tilde{\boldsymbol{a}}^i)}{\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)} : \right. \\
&\qquad \left. a_j \in A_j, u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i) \right\} \qquad (2.1) \\
\beta(i,j) &= \inf\left\{ \frac{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - u_j(\tilde{\boldsymbol{a}}^i)}{\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)} : \right. \\
&\qquad \left. a_j \in A_j, u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) < u_j(\tilde{\boldsymbol{a}}^i), \rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) < \rho(y^i_b|\tilde{\boldsymbol{a}}^i) \right\} \qquad (2.2)
\end{aligned}$$

(We follow the usual convention that the supremum of the empty set is $-\infty$ and the infimum of the empty set is $+\infty$.)

Note that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - u_j(\tilde{\boldsymbol{a}}^i)$ is the gain or loss to player j from deviating from i's preferred action profile $\tilde{\boldsymbol{a}}^i$ and $\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)$ is the increase or decrease in the probability that the bad signal occurs (equivalently, the decrease or increase in the probability that the good signal occurs) following the same deviation. In the definition of $\alpha(i,j)$ we consider only deviations that are strictly profitable; by Assumption 4, such deviations exist and strictly increase the probability that the bad signal occurs, so $\alpha(i,j)$ is strictly positive. In the definition of $\beta(i,j)$ we consider only deviations that are strictly unprofitable *and* strictly decrease the probability that the bad signal occurs, so $\beta(i,j)$ is the infimum of strictly positive numbers and so is necessarily $+\infty$ or finite and non-negative.[4]

To understand the significance of these numbers, think about how player j could gain by deviating from $\tilde{\boldsymbol{a}}^i$. Most obviously, j could gain by deviating to an action that *increases* its current payoff. By assumption, such a deviation will *increase* the probability of a bad signal; assuming that a bad signal leads to a lower continuation utility, whether such a deviation will be profitable will depend on the

[4] Note that if we strengthened Assumption 4 so that *any* deviation – profitable or not – increased the probability of a bad signal (as is the case in Examples 1-3 and would be the case in most resource allocation scenarios), then $\beta(i,j)$ would be the infimum of the empty set whence $\beta(i,j) = +\infty$.

current gain and on the change in probability; $\alpha(i,j)$ represents a measure of net profitability from such deviations. However, player j could also gain by deviating to an action that *decreases* its current payoff but also *decreases* the probability of a bad signal, and hence leads to a higher continuation utility. $\beta(i,j)$ represents a measure of net profitability from such deviations.

Because $\tilde{V}$ lies in the supporting hyperplane H and the utilities for action profiles not in $\tilde{\boldsymbol{A}}$ lie strictly below H, in order that the strategy profile $\boldsymbol{\pi}$ achieves an efficient payoff it is necessary and sufficient that $\boldsymbol{\pi}$ use only preferred action profiles: $U(\boldsymbol{\pi}) \in V$ if and only if $\boldsymbol{\pi}(h) \in \tilde{\boldsymbol{A}}$ for every public history h (independently of the discount factor δ). For PPE strategies we can say a lot more. The first Proposition is almost obvious; the second and third seem far from obvious. (All proofs are in the Appendix.)

Proposition 1 *In order that $\tilde{v}^i$ be achievable in a PPE equilibrium (for any discount factor δ) it is necessary and sufficient that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) \le u_j(\tilde{\boldsymbol{a}}^i)$ for every $j \neq i$ and every $a_j \in A_j$.*

Proposition 2 *If $\boldsymbol{\pi}$ is an efficient PPE (for any discount factor δ) and i is active following some history (i.e., $\boldsymbol{\pi}(h) = \tilde{\boldsymbol{a}}^i$ for some h) then*

$$\alpha(i,j) \le \beta(i,j) \tag{2.3}$$

for every $j \in N, j \neq i$.

Proposition 3 *If $\boldsymbol{\pi}$ is an efficient PPE (for any discount factor δ) and i is active following some history (i.e., $\boldsymbol{\pi}(h) = \tilde{\boldsymbol{a}}^i$ for some h) then for every $a_i \in A_i$*

$$\tilde{v}^i_i - u_i(a_i, \tilde{\boldsymbol{a}}^i_{-i}) \ge \frac{1}{\lambda_i} \sum_{j \neq i} \lambda_j \, \alpha(i,j) \left[\rho(y^i_b | a_i, \tilde{\boldsymbol{a}}^i_{-i}) - \rho(y^i_b | \tilde{\boldsymbol{a}}^i)\right] \tag{2.4}$$

The import of Propositions 2 and 3 is that if any of these inequalities fail then certain efficient payoff vectors can *never* be achieved in PPE, no matter what the discount factor is. In the next Sections, we show how these inequalities and other conditions yield necessary and sufficient conditions that certain sets be self-generating and hence yield sufficient conditions for efficient PPE.

Proposition 2 might seem quite mysterious: α is a measure of the current gain to deviation and β is a measure of the future gain to deviation; there seems no obvious reason why PPE should necessitate any particular relationship between α and β. As the proof will show, however, the assumption of two signals and the efficiency of payoffs in V imply that α is bounded above and β is bounded below by the same quantity, which is a weighted difference of continuation values – a quantity that does have an obvious connection to PPE.

2.3.2 Characterizing Efficient Equilibrium Payoffs

In order to find efficient PPE payoffs we use the technique developed in [14] by Abreu, Pearce, and Stacchetti (hereafter APS) and look for self-generating sets of efficient payoffs.

Fix a subset $W \subset \mathrm{co}[u(\boldsymbol{A})]$ and a *target payoff* $v \in \mathrm{co}[u(\boldsymbol{A})]$. Recall from APS that v can be *decomposed with respect to* W (for a given discount factor $\delta < 1$) if there exist an action profile $\boldsymbol{a} \in \boldsymbol{A}$ and continuation payoffs $\gamma : Y \to W$ such that

- v is the (weighted) average of current and continuation payoffs when players follow $\boldsymbol{a}$
$$v = (1-\delta)u(\boldsymbol{a}) + \delta \sum_{y \in Y} \rho(y|\boldsymbol{a})\gamma(y)$$
- continuation payoffs provide no incentive to deviate: for each j and each $a_j \in A_j$
$$v_j \geq (1-\delta)u_j(a_j, \boldsymbol{a}_{-j}) + \delta \sum_{y \in Y} \rho(y|a_j, \boldsymbol{a}_{-j})\gamma_j(y)$$

Write $\mathcal{B}(W, \delta)$ for the set of target payoffs $v \in \mathrm{co}[u(\boldsymbol{A})]$ that can be decomposed with respect to W for the discount factor δ. Recall that W is *self-generating* if

$W \subset \mathcal{B}(W, \delta)$; i.e., every target vector in W can be decomposed with respect to W.

Because V lies in the hyperplane H, if $v \in V$ and it is possible to decompose $v \in V$ with respect to *any* set and for *any* discount factor, then the associated action profile $\boldsymbol{a}$ must lie in $\tilde{\boldsymbol{A}}$ and the continuation payoffs must lie in V. Because we are interested in efficient payoffs we can therefore restrict our search for self-generating sets to subsets $W \subset V$. In order to understand which sets $W \subset V$ can be self-generating, we need to understand how players might profitably gain from deviating from the current recommended action profile. Because we are interested in subsets $W \subset V$, the current recommended action profile will always be $\tilde{\boldsymbol{a}}^i$ for some i, so we need to ask how a player j might profitably gain from deviating from $\tilde{\boldsymbol{a}}^i$. For player $j \neq i$, a profitable deviation might occur in one of two ways: j might gain by choosing an action $a_j \neq \tilde{a}^i_j$ that increases j's *current* payoff or by choosing an action $a_j \neq \tilde{a}^i_j$ that alters the signal distribution in such a way as to increase j's *future* payoff. Because $\tilde{\boldsymbol{a}}^i$ yields i its best current payoff, a profitable deviation by i might occur only by choosing an action that that alters the signal distribution in such a way as to increase i's *future* payoff. In all cases, the issue will be the net of the current gain/loss against the future loss/gain.

We focus attention on sets of the form

$$V_\mu = \{v \in V : v_i \geq \mu_i \text{ for each } i\}$$

where $\mu \in \mathbb{R}^n$; we assume without further comment that $V_\mu \neq \emptyset$. For lack of a better term, we say that V_μ is *regular* if for each $i \in N$ there is a vector $\hat{v}^i \in V_\mu$ such that $\hat{v}^i_j = \mu_j$ for each $j \neq i$. Whether or not V_μ is regular depends both on the shape of V and on the magnitude of μ: see Figures 2.1, 2.2, 2.3 for instance. A few simple facts are useful to note:

- If $\tilde{v}^i_j = 0$ for all $i, j \in N$ with $i \neq j$ (as is the case in many resource sharing scenarios such as Examples 2, 3) then V_μ is regular for every $\mu \geq 0$.

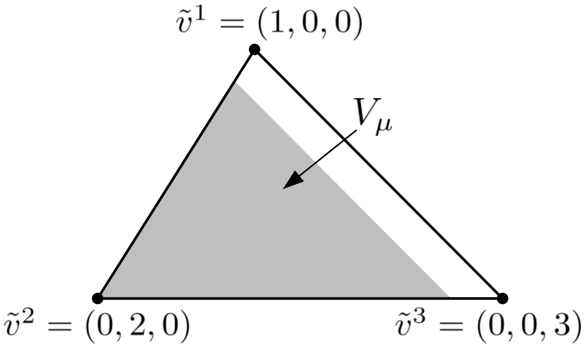


Figure 2.1: $\mu = (0, 1/4, 0)$; V_μ is regular

- If $V_\mu \neq \emptyset$ and V_μ is a subset of the interior of V (relative to the hyperplane H) then V_μ is regular.

- If v lies in the interior of V (relative to the hyperplane H) and $\mu = v - \epsilon \cdot \mathbf{1}$ for $\epsilon > 0$ sufficiently small, then $v \in V_\mu$ and V_μ is regular.

- If V_μ is not a singleton then it must contain a point of the interior of V (relative to the hyperplane H).

If V_μ is a singleton, it can only be a self-generating set (and hence achievable in a PPE) if $V_\mu = \tilde{v}^i$ for i; because we have already characterized this possibility in Proposition 1, we focus on the non-degenerate case in which V_μ is not a singleton and hence contains a point of the interior of V. Note that a point in the interior of V can only be achieved by a repeated game strategy in which *all* players are active following some history.

The following result provides necessary and sufficient conditions on μ, the payoff structure, the information structure and the discount factor that a regular V_μ be a self-generating set.

Theorem 1 *Fix μ; assume that V_μ is regular and not an extreme point of V. In order that V_μ be a self-generating set, it is necessary and sufficient that the following conditions be satisfied:*

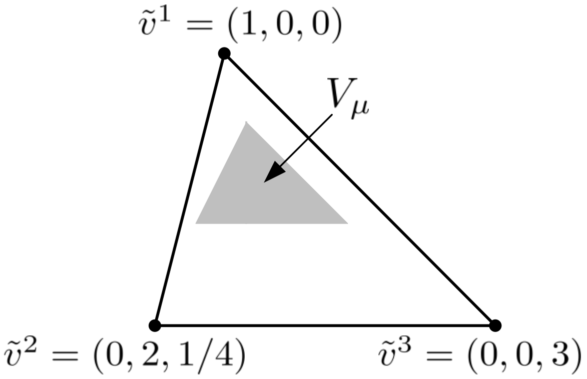

$\tilde{v}^1 = (1, 0, 0)$
V_μ
$\tilde{v}^2 = (0, 2, 1/4)$
$\tilde{v}^3 = (0, 0, 3)$

Figure 2.2: $\mu = (1/2, 1/2, 1/2)$; V_μ is regular

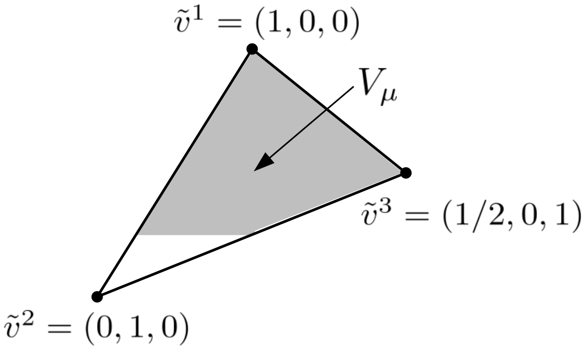

$\tilde{v}^1 = (1, 0, 0)$
V_μ
$\tilde{v}^3 = (1/2, 0, 1)$
$\tilde{v}^2 = (0, 1, 0)$

Figure 2.3: $\mu = (1/4, 0, 0)$; V_μ is not regular

C*ondition 1 for all $i, j \in N$ with $i \neq j$:*

$$\alpha(i,j) \leq \beta(i,j) \tag{2.5}$$

C*ondition 2 for all $i \in N$ and all $a_i \in A_i$:*

$$\tilde{v}_i^i - u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i) \geq \frac{1}{\lambda_i} \sum_{j \neq i} \lambda_j \, \alpha(i,j) \left[\rho(y_b^i | a_i, \tilde{\boldsymbol{a}}_{-i}^i) - \rho(y_b^i | \tilde{\boldsymbol{a}}^i)\right] \tag{2.6}$$

C*ondition 3 for all $i \in N$:*

$$\mu_i \geq \max_{j \neq i} \left(\tilde{v}_i^j + \alpha(j,i)[1 - \rho(y_b^j | \tilde{\boldsymbol{a}}^j)]\right) \tag{2.7}$$

C*ondition 4 the discount factor δ satisfies:*

$$\delta \geq \underline{\delta}_\mu \triangleq \left(1 + \frac{1 - \sum_i \lambda_i \mu_i}{\sum_i \left[\lambda_i \tilde{v}_i^i + \sum_{j \neq i} \lambda_j \, \alpha(i,j) \, \rho(y_b^i | \tilde{\boldsymbol{a}}^i)\right] - 1}\right)^{-1} \tag{2.8}$$

One way to contrast our approach with that of FLM (and FLT) is to think about the constraints that need to be satisfied to decompose a given target payoff v with respect to a given set V_μ. By definition we must find a current action profile $\boldsymbol{a}$ and continuation payoffs γ. The achievability condition (that v is the weighted combination of the utility of the current action profile and the expected continuation values) yields a family of linear equalities. The incentive compatibility conditions (that players must be deterred from deviating from $\boldsymbol{a}$) yield a family of linear inequalities. In the context of FLM, satisfying all these linear inequalities simultaneously requires a large and rich collection of signals so that many different continuation payoffs can be assigned to different deviations. Because we have only two signals, we are only able to choose two continuation payoffs but still must satisfy the same family of inequalities – so our task is much more difficult. It is this difficulty that leads to the Conditions in Theorem 1.

Note that $\underline{\delta}_\mu$ is *decreasing* in μ. Since Condition 3 puts an absolute lower bound on μ and Condition 4 puts an absolute lower bound on $\underline{\delta}_\mu$ this means that (subject to the regularity constraint) there is a μ^* such that V_{μ^*} is the largest self-generating set (of this form) and $\underline{\delta}_{\mu^*}$ is the smallest discount factor (for which any set of this form can be self-generating). This may seem puzzling – increasing the discount factor beyond a point makes no difference – but remember that we are providing a characterization of self-generating sets and not of PPE payoffs. However, as we shall see in Theorem 4, for the two-player case, we do obtain a complete characterization of (efficient) PPE payoffs and we demonstrate the same phenomenon.

2.3.3 Constructing Efficient Perfect Public Equilibria

Because every payoff in a self-generating set can be achieved in a PPE, Theorem 1 immediately provides sufficient conditions achieving (some) given target payoffs in perfect public equilibrium. In fact, we can provide an explicit algorithm for *computing* PPE strategies. A consequence of this algorithm is that (at least when action spaces are finite), the constructed PPE enjoys an interesting and potentially useful robustness property.

Given the various parameters of the environment (game payoffs, information structure, discount factor) and of the problem (lower bound, target vector), the algorithm takes as input in period t the current continuation vector $v(t)$ and computes, for each player j, an indicator $d_j(v(t))$ defined as follows:

$$d_j(v(t)) = \frac{\lambda_j[v_j(t) - \mu_j]}{\lambda_j[\tilde{v}_j^j - v_j(t)] + \sum_{k\neq j} \lambda_k\, \alpha(j,k)\rho(y_b^j|\tilde{\boldsymbol{a}}^j)}$$

(Note that each player can compute every d_j from the current continuation vector $v(t)$ and the various parameters.) Having computed $d_j(v(t))$ for each j, the algorithm finds the player i^* whose indicator is greatest. (In case of ties, we arbitrarily choose the player with the largest index.) The current action profile is i^*'s preferred action profile $\tilde{\boldsymbol{a}}^{i^*}$. The algorithm then uses the labeling $Y = \{y_g^{i^*}, y_b^{i^*}\}$ to compute continuation values for each signal in Y.

Theorem 2 *If the conditions in Theorem 1 are satisfied, then every payoff $v \in V_\mu$ can be achieved in a PPE. For $v \in V_\mu$, a PPE strategy profile that achieves v can be computed by the algorithm in Table 4.6.*

2.3.4 Robustness

A consequence of our constructive algorithm is that, for generic values of the parameters of the environment and of the problem and for as many periods as we specify, the strategies we identify are locally constant in these parameters. To make this precise, we assume for this subsection that action spaces A_i are finite. The parameters of the model are the utility mapping $U : \boldsymbol{A} \to \mathbb{R}^n$ and the probabilities $\rho(\cdot|\cdot) : Y \times \boldsymbol{A} \to [0, 1]$. Because the probabilities must sum to 1 and we require full support, the parameter space of the model is

$$\Omega = (R^n \times [0, 1])^{\boldsymbol{A}}$$

The parameters of the problem are the discount factor δ, the constraint vector μ and the target profile v^*; because the target profile lies in a hyperplane, the parameter space for the particular problem is

$$\Theta = (0, 1) \times \mathbb{R}^n \times \mathbb{R}^{n-1}$$

Let $\Xi \subset \Omega \times \Theta$ be the subset of parameters that satisfy the Conditions of Theorem 1. For $\xi \in \Xi$, the algorithm generates an strategy profile

$$\boldsymbol{\pi}_\xi : \mathcal{H} \to \mathbf{A}$$

For $T \geq 0$ we write $\boldsymbol{\pi}_\xi^T$ for the restriction of $\boldsymbol{\pi}_\xi$ to the set $\mathcal{H}^T$ of histories of length at most T.

Table 2.1: The algorithm used by each player.

```
Input: The current continuation payoff $v(t) \in V_\mu$

For each $j$
        Calculate the indicator $d_j(v(t))$
Find the player $i$ with largest indicator (if a tie, choose largest $i$)
    $i = \max_j \{\arg\max_{j \in N} d_j(v(t))\}$
Player $i$ is active; chooses action $\tilde{\boldsymbol{a}}_i^i$
Players $j \neq i$ are inactive; choose action $\tilde{\boldsymbol{a}}_j^i$
Update $v(t+1)$ as follows:
    if $y^t = y_g^i$ then
        $v_i(t+1) = \tilde{v}_i^i + (1/\delta)(v_i(t) - \tilde{v}_i^i) - (1/\delta - 1)(1/\lambda_i) \sum_{j \neq i} \lambda_j \alpha(i,j) \rho(y_b^i | \tilde{\boldsymbol{a}}^i)$
        $v_j(t+1) = \tilde{v}_j^i + (1/\delta)(v_j(t) - \tilde{v}_j^i) + (1/\delta - 1) \alpha(i,j) \rho(y_b^i | \tilde{\boldsymbol{a}}^i)$
            for all $j \neq i$
    if $y^t = y_b^i$ then
        $v_i(t+1) = \tilde{v}_i^i + (1/\delta)(v_i(t) - \tilde{v}_i^i) + (1/\delta - 1)(1/\lambda_i) \sum_{j \neq i} \lambda_j \alpha(i,j) \rho(y_g^i | \tilde{\boldsymbol{a}}^i)$
        $v_j(t+1) = \tilde{v}_j^i + (1/\delta)(v_j(t) - \tilde{v}_j^i) - (1/\delta - 1) \alpha(i,j) \rho(y_g^i | \tilde{\boldsymbol{a}}^i)$
            for all $j \neq i$
```

Theorem 3 *For each $T \geq 0$ there is a subset $\Xi_T \subset \Xi$ that is closed and has measure 0 with the property that the mapping $\xi \to \boldsymbol{\pi}_\xi^T : \Xi \to \mathcal{H}^T$ is locally constant on the complement of Ξ_T.*

In words: if ξ, ξ' are close together and neither lies in the proscribed small set of parameters Ξ_T, then the strategies $\boldsymbol{\pi}_\xi, \boldsymbol{\pi}_{\xi'}$ *coincide* for at least the first T periods.

2.3.5 Two Players

Theorem 1 provides a complete characterization of self-generating sets that have a special form. If there are only two players then maximal self-generating sets – the set of all PPE – have this form and so it is possible to provide a complete characterization of PPE. We focus on what seems to be the most striking finding: either there are no efficient PPE outcomes at all (for any discount factor $\delta < 1$) or there is a discount factor $\delta^* < 1$ with the property that any target payoff in V that can be achieved as a PPE for *some* δ can already be achieved for *every* $\delta \geq \delta^*$.

Theorem 4 *Assume $N = 2$ (two players). Either*

- *no target profile in V can be supported in a PPE for any $\delta < 1$* or
- *there exist μ_1^*, μ_2^* and a discount factor $\delta^* < 1$ such that if δ is any discount factor with $\delta^* \leq \delta < 1$ then the set of payoff vectors that can be supported in a PPE when the discount factor is δ is precisely*

$$E = \{v \in V : v_i \geq \mu_i^* \textit{ for } i = 1, 2\}$$

The proof yields the following explicit expressions for μ_1^*, μ_2^* and δ^*:

$$\mu_1^* = \tilde{v}_1^2 + \alpha(2,1)[1 - \rho(y_b^2|\tilde{\boldsymbol{a}}^2)],\ \mu_2^* = \tilde{v}_2^1 + \alpha(1,2)[1 - \rho(y_b^1|\tilde{\boldsymbol{a}}^1)],$$

$$\delta^* = \left(1 + \frac{1 - \lambda_1 \mu_1^* - \lambda_2 \mu_2^*}{\sum_i \left[\lambda_i \tilde{v}_i^i + \lambda_{-i}\, \alpha(i,-i)\, \rho(y_b^i | \tilde{\boldsymbol{a}}^i) \right] - 1} \right)^{-1}.$$

2.4 Applications to Throughput-Maximizing Spectrum Sharing

2.4.1 Motivation

Cognitive radios have increased in popularity in recent years, because they have the potential to significantly improve the spectrum efficiency. Specifically, cognitive radios enable the secondary users (SUs), who initially have no rights to use the spectrum, to share the spectrum with primary users (PUs), who are licensed to use the spectrum, as long as the PUs' quality of service (QoS), such as the throughput, is not affected by the SUs [25]. A common approach to guarantee PUs' QoS requirements is to impose *interference temperature* (IT) constraints [25]-[27][29]–[37]; that is, the SUs cannot generate an interference level higher than the interference temperature limit set by the PUs. One of the major challenges in designing cognitive radio systems is to construct a spectrum sharing policy that achieves high spectrum efficiency while maintaining the IT constraints set by PUs. The spectrum sharing policy, which specifies the SUs' transmit power levels, is essential to improve spectrum efficiency and protect the PUs' QoS.

Since SUs can use the spectrum as long as they do not degrade the PUs' QoS, they coexist and interact with each other in the system for long periods of time. It is then natural to model the interaction among the SUs as a repeated game. Moreover, due to strong multi-user interference and the imperfect estimation of the interference temperature, we model the interaction as a repeated game with strong negative externality and imperfect monitoring defined in Section 2.2. A repeated-game strategy prescribes what action to take given past observations, and therefore, can be considered as a spectrum sharing policy. If a repeated game strategy constitutes an equilibrium, then no user can gain from deviation at any occasion. Hence, an equilibrium strategy is a deviation-proof spectrum sharing policy.

Based on the results in Section 2.3, we propose a design framework for constructing efficient deviation-proof spectrum sharing policies. Our design framework first characterizes the set of Pareto optimal operating points achievable by deviation-proof policies, and then for any operating point in this set, constructs a deviation-proof policy to achieve it. The proposed policy can be easily implemented in a distributed manner. Moreover, the proposed spectrum sharing policies exhibit the following key advantages over state-of-the-art policies:

- The proposed spectrum sharing policies allow the users to choose *time-varying* power levels (e.g. the users can transmit in a time-division multiple-access (TDMA) fashion). Under strong multi-user interference, this is much more efficient than most existing policies which require the SUs to transmit at *constant* power levels.

- The proposed policies achieve Pareto optimal operating points, even when the SUs are *impatient*, namely they discount future payoffs, and their discount factor are strictly smaller than one.

- Under the proposed policies, the requirement on the users' monitoring ability is significantly relaxed compared to existing works based on repeated games, which require either perfect monitoring of all the users' individual transmit power levels [39]–[42] or sufficiently good monitoring to distinguish sufficiently many IT levels [9]. Specifically, their monitoring ability can be limited in that they only need to distinguish *two* IT levels regardless of the number of power levels each user can choose from, and their monitoring can be imperfect due to the *erroneous* measurements of the interference temperature.[5]

- The proposed policies are deviation-proof, namely self-interested users annot improve their QoS by deviating from the policy, and hence find it in their self-interests to follow the proposed policies.

2.4.2 Related Works

In general, the optimal spectrum sharing policy should allow SUs to transmit at different power levels temporally even when the environment (e.g. the number of SUs, the channel gains) remains unchanged. However, most existing spectrum sharing policies require the SUs to transmit at *constant* power levels over the time horizon in which they interact[6] [26]–[38]. These policies with constant power levels are inefficient in many spectrum sharing scenarios where the interference among the SUs is strong. Under strong multi-user interference, increasing one user's power level significantly degrades the other users' QoS. Hence, when the cross channel gains are large, the feasible QoS region is nonconvex [43]. In this case of nonconvex feasible QoS region, a spectrum sharing policy with constant power levels is inferior to a policy with *time-varying* power levels in which the users transmit in a time-division multiple-access (TDMA) fashion, because the latter can achieve the Pareto boundary of the convex hull of the nonconvex feasible QoS region.

The spectrum sharing policy in a repeated game framework was studied in [39]–[42], under the assumption of *perfect* monitoring, namely the assumption that each SU can perfectly monitor the individual transmit power levels of all the other SUs. In the policies in [39]–[42], when a deviation from the prescribed policy

[5] As will be described later, there is an entity that regulates the interference temperature in the system, who measures the interference temperature imperfectly and feedbacks to the users a binary signal indicating whether the constraints are violated.

[6] Although some spectrum sharing policies go through a transient period of adjusting the power levels before the convergence to the optimal power levels, the users maintain constant power levels after the convergence.

by any user is detected, a perpetual punishment phase [39] or a punishment phase of certain duration [40][42] will be triggered. In the punishment phase, all the users transmit at the maximum power levels to create strong interference to each other, resulting in low QoS of all the users as a punishment. Due to the threat of this punishment, all the users will follow the policy in their self-interests. However, since the monitoring can never be perfect, the punishment phase, in which all the users receive low throughput, will be triggered even if no one deviates. Thus, the users' repeated-game payoffs, averaged over all the stage-game payoffs, cannot be Pareto optimal because of the low payoffs received in the punishment phases. Hence, the policies in [39]–[42] must have performance loss in practice where the monitoring is always imperfect.

We illustrate the performance gain of the proposed policies over the existing policies in Fig. 2.4. We show the best operating points achievable by different classes of policies in a spectrum sharing system with two SUs. Due to the strong multi-user interference, the best operating points achievable by policies with constant power levels [26]–[38] (the dashed curve) are Pareto dominated by the best operating points achieved by policies with time-varying power levels (the straight line). The proposed policy, which are deviation-proof, can achieve a portion of the Pareto optimal operating points (the thick line). Under imperfect monitoring, the policies designed under the assumption of perfect monitoring [39]–[42] (the solid curve) have large performance loss compared to the proposed policy.

Finally, we summarize the comparison of our work with the existing works in dynamic spectrum sharing in Table 4.1. We distinguish our work from existing works in the following categories: the power levels prescribed by the spectrum sharing policy are constant or time-varying, whether the policy can be implemented in a distributed fashion or not, whether the policy is deviation-proof or not, and what are the requirements on the SUs' monitoring ability. The "monitoring" category is only discussed within the works based on repeated games.

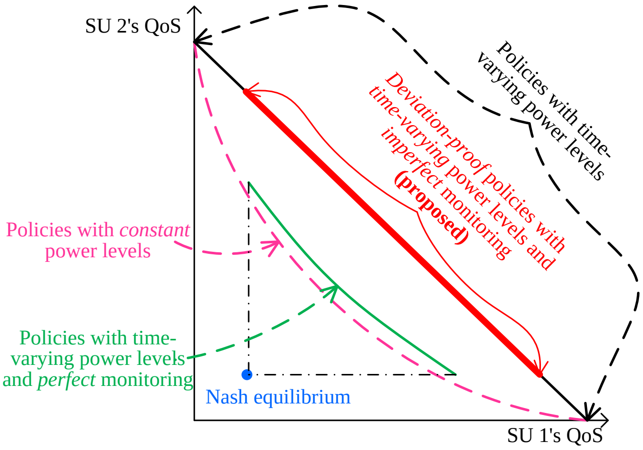

SU 2's QoS
Policies with time-varying power levels
Deviation-proof policies with time-varying power levels and imperfect monitoring (proposed)
Policies with constant power levels
Policies with time-varying power levels and perfect monitoring
Nash equilibrium
SU 1's QoS

Figure 2.4: An illustration of the best operating points achievable by different policies in a two-SU spectrum sharing system.

Table 2.2: Comparison With Related Works In Dynamic Spectrum Sharing.

	Power levels	Distributed	Deviation-proof	Monitoring
[26][27]	Constant	No	No	N/A
[28]–[35]	Constant	Yes	No	N/A
[36]–[38]	Constant	Yes	Yes	N/A
[39]–[42]	Time-varying	Yes	Yes	Perfect
Proposed	Time-varying	Yes	Yes	Imperfect

2.4.3 Model and Problem Formulation

2.4.3.1 Stage Game

We consider a system with one primary user and N secondary users (see Fig 3.1 for an illustrating example of a system with two secondary users). The set of SUs is denoted by $N \triangleq \{1, 2, \ldots, n\}$. Each SU has a transmitter and a receiver. The channel gain from SU i's transmitter to SU j's receiver is g_{ij}. Each SU i chooses a power level a_i from a compact set A_i. We assume that $0 \in A_i$, namely SU i can choose not to transmit. We write SU i's maximum transmit power as $A_i^{\max}$. The set of joint power profiles is denoted by $\boldsymbol{A} = \prod_{i \in N} \boldsymbol{A}_i$, and the joint power profile of all the SUs is denoted by $\boldsymbol{a} = (a_1, \ldots, a_n) \in \boldsymbol{A}$. Let $\boldsymbol{a}_{-i}$ be the power profile of all the SUs other than SU i. Each SU i's instantaneous payoff (QoS) is a function of the joint power profile, namely $u_i : \boldsymbol{A} \to \mathbb{R}^+$. Each SU i's payoff $u_i(\boldsymbol{a})$ is decreasing in the other SUs' power levels a_j, $\forall j \neq i$. Note that we do *not* assume that $u_i(\boldsymbol{a})$ is increasing in a_i.[7] But we do assume that $u_i(\boldsymbol{a}) = 0$ if $a_i = 0$, because a SU's payoff should be zero when it does not transmit. One example of many possible payoff functions is the SU's throughput:

$$u_i(\boldsymbol{a}) = \log_2 \left(1 + \frac{a_i g_{ii}}{\sum_{j \in \mathcal{N}, j \neq i} a_j g_{ji} + \sigma_i} \right), \tag{2.9}$$

where σ_i is the noise power at SU i's receiver.

2.4.3.2 Imperfect Monitoring

As in [32]–[35], there is a local spectrum server (LSS) serving as a mediating entity among the SUs. The LSS has a receiver to measure the interference temperature and a transmitter to broadcast signals, but it cannot control the actions of the autonomous SUs. The LSS could be a device deployed by the PU or simply the

[7] In some scenarios with energy efficiency considerations, the payoff is defined as the ratio of throughput to transmit power, which may not monotonically increase with the transmit power.

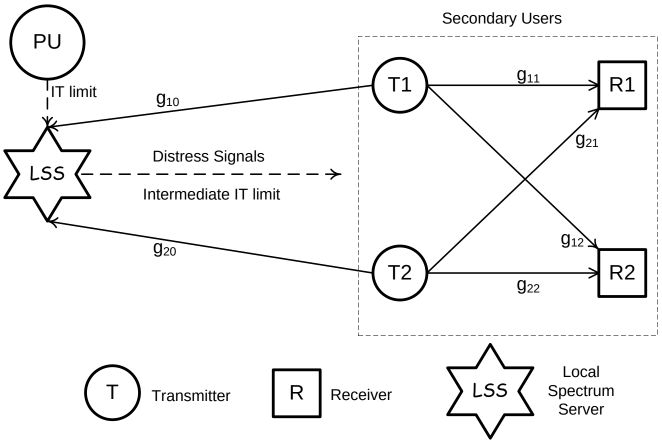


Figure 2.5: An example system model with two secondary users. The solid line represents a link for data transmission, and the dashed line indicate a link for control signals. The channel gains for the corresponding data link are written in the figure. The primary user (PU) specifies the interference temperature (IT) limit to the local spectrum server (LSS). The LSS sets the intermediate IT limit to the secondary users and send distress signals if the estimated interference power exceeds the IT limit.

PU itself, if the PU manages by itself the spectrum leased to the SUs. Even when the PU is the LSS, it is beneficial to consider the LSS as a separate logical entity that performs the functionality of spectrum management. The LSS could also be a device deployed by some regulatory agency such as Federal Communications Commission (FCC), who uses it for spectrum management in that local geographic area. In both cases, the LSS aims to improve the spectrum efficiency (e.g. the sum throughput of all the SUs) and the fairness, while ensuring that the IT limit set by the PU is not violated. Note that the PU may also want to maximize the spectrum efficiency to maximize its revenue obtained from spectrum leasing, since its revenue may be proportional to the sum throughput of the SUs.

The LSS measures the interference temperature at its receiver imperfectly. The measurement can be written as $\sum_{i \in N} a_i g_{i0} + \varepsilon$, where g_{i0} is the channel gain from SU i's transmitter to the LSS's receiver, and ε is the additive measurement error. We assume that the measurement error has zero mean and a probability distribution function f_ε known to the LSS. We assume as in most existing works (e.g. [26]–[36]) that the IT limit $\bar{I}$ set by the PU is known perfectly by the LSS. Although the LSS aims to keep the interference temperature below the IT limit $\bar{I}$, it will set a lower intermediate IT limit $I \leq \bar{I}$ to be conservative because of measurement errors. Hence, the IT constraint imposed by the LSS is

$$\textstyle\sum_{i \in \mathcal{N}} a_i g_{i0} \leq I. \tag{2.10}$$

Even if the actual interference temperature $\sum_{i \in \mathcal{N}} a_i g_{i0}$ does not exceed the intermediate IT limit I, the erroneous measurement $\sum_{i \in \mathcal{N}} a_i g_{i0} + \varepsilon$ may still exceed the IT limit $\bar{I}$ set by the PU. In this case, the LSS will broadcast a distress signal to all the SUs. Given the joint power profile $\boldsymbol{a}$, this false alarm probability is

$$\Gamma(\boldsymbol{a}) = \Pr\left(\textstyle\sum_{i \in \mathcal{N}} a_i g_{i0} + \varepsilon > \bar{I} \mid \sum_{i \in \mathcal{N}} a_i g_{i0} \leq I\right), \tag{2.11}$$

where $\Pr(A)$ is the probability that the event A happens. We can see that a larger intermediate IT limit I enables the SUs to transmit at higher power levels, but results in a larger false alarm probability and a higher frequency of sending distress signals. Hence, there is an interesting tradeoff between the spectrum efficiency and the cost of sending distress signals.

We denote the set of events of whether the IT limit $\bar{I}$ is violated by $Y = \{y_0, y_1\}$. The (measurement) outcome y is determined by

$$y = \begin{cases} y_0, & \text{if } \sum_{i \in N} a_i g_{i0} + \varepsilon > \bar{I} \\ y_1, & \text{otherwise} \end{cases}. \tag{2.12}$$

We write the conditional probability distribution of the outcome y given the joint power profile $\boldsymbol{a}$ as $\rho(y|\boldsymbol{a})$, which can be calculated as

$$\begin{aligned} \rho(y_1|\boldsymbol{a}) &= \int_{x \leq \bar{I} - \sum_{i \in N} a_i g_{i0}} f_\varepsilon(x)\ dx, \\ \rho(y_0|\boldsymbol{a}) &= 1 - \rho(y_1|\boldsymbol{a}). \end{aligned} \tag{2.13}$$

At the end of time slot t, the LSS sends a distress signal if the outcome $y^t = y_0$. Note that the LSS does not send signals when the outcome is y_1, and the SUs know that the outcome is y_1 by default when they do not receive the distress signal.

2.4.3.3 Spectrum Sharing Policies

The system is time slotted at $t = 0, 1, \ldots$. We assume that the users are synchronized as in [26]–[38]. At the beginning of time slot t, each SU i chooses its power level a_i^t, and receives a payoff $u_i(\boldsymbol{a}^t)$. The LSS obtains the measurement $\sum_{i \in \mathcal{N}} a_i^t g_{i0} + \varepsilon^t$, where ε^t is the realization of the error ε at time slot t, and compare the measurement with the IT limit $\bar{I}$. At the end of time slot t, the LSS sends a distress signal if the outcome $y^t = y_0$. Note that the LSS does not send signals when the outcome is y_1, and the SUs know that the outcome is y_1 by default when they do not receive the distress signal.

Note that in repeated games with perfect monitoring [39]–[42], the outcome available to each SU at time slot t is precisely the joint power profile chosen by the SUs, i.e. $y^t = \boldsymbol{a}^t$. We say the monitoring is imperfect if $y^t \neq \boldsymbol{a}^t$. In a general repeated game with imperfect monitoring, in order to achieve Pareto optimality, the set of outcomes Y should have a large cardinality, namely $|Y| \geq |\boldsymbol{A}_i| + |\boldsymbol{A}_j| - 1$ for all $i \in \mathcal{N}$ and all $j \neq i$ [9]. In contrast, our proposed policy can achieve Pareto optimality even when $|Y| = 2$ regardless of the cardinality of the SU's action set $\boldsymbol{A}_i$.

At each time slot t, each SU i determines its transmit power a_i^t based on its history, which is a collection of all the past power levels it has chosen and all the past measurement outcomes. Formally, the history of SU i up to time slot $t \geq 1$ is $h_i^t = \{a_i^0, y^0; \ldots; a_i^{t-1}, y^{t-1}\} \in (\boldsymbol{A}_i \times Y)^t$, and that at time slot 0 is $h_i^0 = \varnothing$. The history of SU i contains private information about SU i's power levels that is unknown to the other SUs; in contrast, we define the *public history* as $h^t = \{y^0; \ldots; y^{t-1}\} \in Y^t$ for $t \geq 1$ and $h^0 = \varnothing$. The public history h^t only contains the measurement outcomes that are known to all the SUs.

We focus on *public strategies*, in which each SU's decision depends on the public history only. Hence, each SU i's strategy π_i is a mapping from the set of all possible public histories to its action set, namely $\pi_i : \sqcup_{t=0}^{\infty} Y^t \rightarrow \boldsymbol{A}_i$. The spectrum sharing policy is the joint strategy profile of all the SUs, defined as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$.

The SUs are selfish and maximize their own long-term discounted payoffs. Assuming, as in [39]–[**?**], the same discount factor $\delta \in [0, 1)$ for all the SUs, each SU i's (long-term discounted) payoff can be written as

$$U_i(\boldsymbol{\pi}) = (1-\delta)\left[u_i(\boldsymbol{a}^0) + \sum_{t=1}^{\infty} \delta^t \cdot \sum_{y^{t-1} \in Y} \rho(y^{t-1} | \boldsymbol{a}^{t-1}) u_i(\boldsymbol{a}^t) \right],$$

where $\boldsymbol{a}^0$ is determined by $\boldsymbol{a}^0 = \boldsymbol{\pi}(\varnothing)$, and $\boldsymbol{a}^t$ for $t \geq 1$ is determined by $\boldsymbol{a}^t = \boldsymbol{\pi}(h^t) = \boldsymbol{\pi}(h^{t-1}; y^{t-1})$. The discount factor represents the "patience" of the SUs; a larger discount factor indicates that a SU is more patient. The discount factor is determined by the delay sensitivity of the SUs' applications.

We define the deviation-proof policy as the perfect public equilibrium (PPE) of the game. The PPE prescribes a strategy profile $\boldsymbol{\pi}$ from which no SU has incentive to deviate after any given history at any time slot, and thus can be considered as a deviation-proof policy. It is normally more strict than Nash equilibrium, because it requires that the SUs have no incentive to deviate at any given history, while Nash equilibrium only guarantees this at the histories that possibly arise from the equilibrium strategy. We can also consider PPE in repeated games with imperfect monitoring as the counterpart of subgame perfect equilibrium defined in repeated games with perfect monitoring [8].

Before the definition of PPE, we introduce the concept of continuation strategy: SU i's continuation strategy induced by any history $h^t \in Y^t$, denoted $\pi_i|_{h^t}$, is defined by $\pi_i|_{h^t}(h^\tau) = \pi_i(h^t h^\tau), \forall h^\tau \in Y^\tau$, where $h^t h^\tau$ is the concatenation of the history h^t followed by the history h^τ. By convention, we denote $\boldsymbol{\pi}|_{h^t}$ and $\boldsymbol{\pi}_{-i}|_{h^t}$ the continuation strategy profile induced by h^t of all the SUs and that of all the SUs other than SU i, respectively. Then the PPE is defined as follows [8, Definition 7.1.2]

Definition 1 (Perfect Public Equilibrium) *A strategy profile $\boldsymbol{\pi}$ is a perfect public equilibrium if for any public history $h^t \in Y^t$, the induced continuation strategy $\boldsymbol{\pi}|_{h^t}$ is a Nash equilibrium of the continuation game, namely for all $i \in \mathcal{N}$,*

$$U_i(\boldsymbol{\pi}|_{h^t}) \geq U_i(\pi_i'|_{h^t}, \boldsymbol{\pi}_{-i}|_{h^t}), \text{ for all } \pi_i'. \tag{2.14}$$

We define the equilibrium payoff as a vector of payoffs $\mathbf{v} = (U_1(\boldsymbol{\pi}), \ldots, U_N(\boldsymbol{\pi}))$ achieved at the equilibrium.

2.4.3.4 Problem Formulation

The primary user or the regulatory agency aims to maximize an objective function defined on the SUs' payoffs, $W(U_1(\boldsymbol{\pi}), \ldots, U_N(\boldsymbol{\pi}))$. This definition of the objective function is general enough to include the objective functions deployed in many existing works, such as [26]–[48][39][40]. An example of the objective function is the weighted sum payoff $\sum_{i=1}^{N} w_i U_i$, where $\{w_i\}_{i=1}^{N}$ are the weights satisfying $w_i \in [0,1], \forall i$ and $\sum_{i=1}^{N} w_i = 1$. The PU (respectively, the regulatory agency) maximizes the objective function for the revenue (the spectrum efficiency), while maintaining the IT constraint (2.10). To reduce the cost of sending distress signals, a constraint on the false alarm probability is also imposed as $\Gamma(\boldsymbol{a}) \leq \bar{\Gamma}$, where $\bar{\Gamma}$ is the maximum false alarm probability allowed. At the maximum of the welfare function, some SUs may have extremely low payoffs. To avoid this, a minimum payoff guarantee $\gamma_i \geq 0$ is imposed for each SU i. To sum up, we can formally define the policy design problem as follows

$$\begin{aligned} \max_{\boldsymbol{\pi}} \quad & W(U_1(\boldsymbol{\pi}), \ldots, U_N(\boldsymbol{\pi})) \\ s.t. \quad & \boldsymbol{\pi} \text{ is public perfect equilibrium}, \\ & \sum_{i \in \mathcal{N}} \pi_i(h^t) \cdot g_{i0} \leq I, \ \forall t, \ \forall h^t \in Y^t, \\ & \Gamma(\boldsymbol{\pi}(h^t)) \leq \bar{\Gamma}, \ \forall t, \ \forall h^t \in Y^t, \\ & U_i(\boldsymbol{\pi}) \geq \gamma_i, \ \forall i \in \mathcal{N}. \end{aligned} \tag{2.15}$$

2.4.4 Link to The General Model in Section 2.2

The repeated game model for the spectrum sharing scenario with strong multi-user interference is a special case of the general repeated game model proposed in Section 2.2.

First, each SU i's preferred power profile, written as $\tilde{\boldsymbol{a}}^i = (\tilde{a}_1^i, \ldots, \tilde{a}_n^i)$, is the joint power profile that maximizes SU i's payoff subject to the IT constraint, namely

$$\tilde{\boldsymbol{a}}^i = \arg\max_{\boldsymbol{a} \in \boldsymbol{A}} u_i(\boldsymbol{a}), \text{ subject to } \textstyle\sum_{i \in n} a_i g_{i0} \leq I. \tag{2.16}$$

Since u_i is decreasing in $a_j, \forall j \neq i$, we have $\tilde{a}_j^i = 0, \ \forall j \neq i$. For notational simplicity, we define the maximum payoff achievable by SU i as $\bar{v}_i \triangleq u_i(\tilde{\boldsymbol{a}}^i)$.

We can check easily that Assumptions 2-3 in Section 2.2 are satisfied.

Assumption 1 in Section 2.2 is satisfied when the multi-user interference is strong, which is the scenario that we are most interested in. We are interested in the scenario with strong multi-user interference, because when the multi-user interference is weak, power control becomes less important. We say a spectrum sharing scenario has strong multi-user interference if the following property is satisfied.

Definition 2 (Strong Multi-user Interference) *A spectrum sharing scenario has strong multi-user interference, if the set of feasible payoffs* $\mathcal{V} = \mathrm{conv}\{u(\boldsymbol{a}) = (u_1(\boldsymbol{a}), \ldots, u_n(\boldsymbol{a})) : \boldsymbol{a} \in \boldsymbol{A}, \sum_{i \in n} a_i g_{i0} \leq I\}$*, where* $\mathrm{conv}(X)$ *is the convex hull of* X*, has* $N+1$ *extremal points*[8]*:* $(0, \ldots, 0) \in \mathbb{R}^N$*,* $u(\tilde{\boldsymbol{a}}^1), \ldots, u(\tilde{\boldsymbol{a}}^N)$*.*

This definition characterizes the strong interference among the SUs: the increase of one SU's payoff comes at such an expense of the other SUs' payoffs that the set of feasible payoffs without time sharing is nonconvex. A spectrum sharing scenario satisfies this property when the cross channel gains among users are large [43]. In the extreme case of strong multi-user interference, simultaneous transmissions from different SUs result in packet loss, as captured in the collision model [44]. According to this definition, the set of feasible payoffs can be written as $\mathcal{V} = \mathrm{conv}\{(0, \ldots, 0), u(\tilde{\boldsymbol{a}}^1), \ldots, u(\tilde{\boldsymbol{a}}^n)\}$. Moreover, its Pareto boundary is $\boldsymbol{\mathcal{B}} = \{\mathbf{v} \in \mathcal{V} : \sum_{i=1}^n v_i/\bar{v}_i = 1, \ v_i \geq 0, \forall i\}$ as part of a hyperplane, which can be achieved only by SUs transmitting in a TDMA fashion.

Since Assumptions 1-3 in Section 2.2 are satisfied, we can apply the results in Section 2.3 to characterize the Pareto optimal equilibrium payoffs and construct Pareto optimal spectrum sharing policies. We illustrate the design framework in

[8] *The extremal points of a convex set are those that are not convex combinations of other points in the set.*

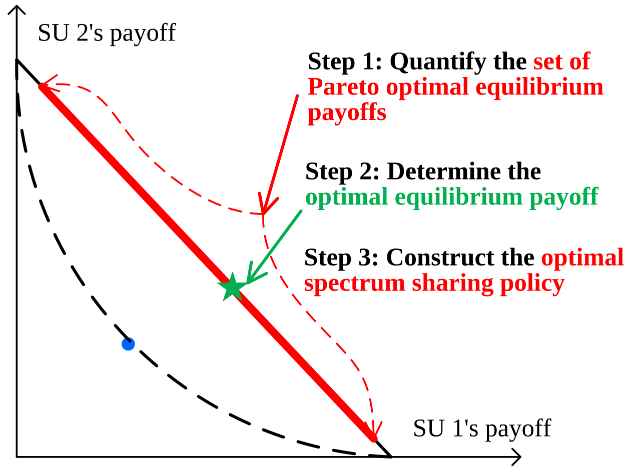

SU 2's payoff
Step 1: Quantify the set of Pareto optimal equilibrium payoffs
Step 2: Determine the optimal equilibrium payoff
Step 3: Construct the optimal spectrum sharing policy
SU 1's payoff

Figure 2.6: The procedure of solving the design problem.

Fig. 2.6. Details are omitted due to space limitation; interesting readers can refer to [2].

2.4.5 Simulation Results

We demonstrate the performance gain of our spectrum sharing policy over existing policies, and validate our theoretical analysis through numerical results. Throughout this section, we use the following system parameters by default unless we change some of them explicitly. The noise powers at all the SUs' receivers are normalized as 0 dB. The maximum transmit powers of all the SUs are 10 dB, $\forall i$. For simplicity, we assume that the direct channel gains have the same distribution $g_{ii} \sim \mathcal{CN}(0,1), \forall i$, and the cross channel gains have the same distribution $g_{ij} \sim \mathcal{CN}(0,\beta), \forall i \neq j$, where β is defined as the *cross interference level*. The channel gain from each SU to the LSS also satisfies $g_{i0} \sim \mathcal{CN}(0,1), \forall i$. The IT limit set by the PU is $\bar{I} = 10$ dB. The measurement error ε is Gaussian distributed with zeros mean and variance 0.1. The maximum false alarm probability is $\bar{\Gamma} = 10\%$. The SUs' payoffs are their throughput as in (2.9). The welfare function is the average payoff, i.e. $W = \sum_{i=1}^{N} \frac{1}{N} U_i$. The minimum payoff guarantee is 10% of the maximum achievable payoff, i.e. $\gamma_i = 0.1 \cdot \bar{v}_i, \forall i$.

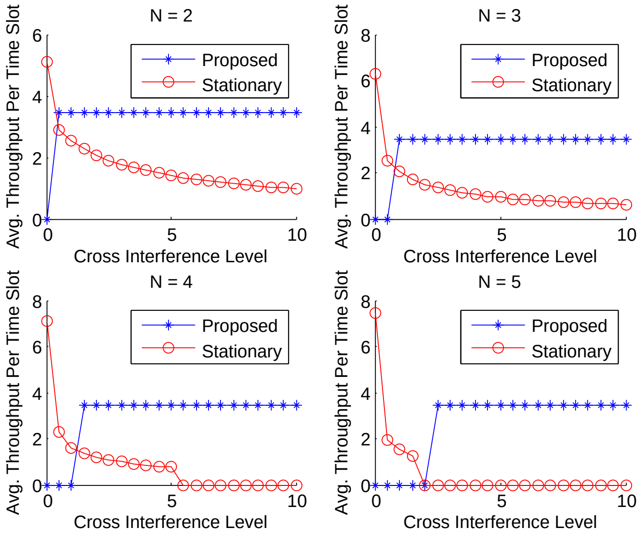


Figure 2.7: Performance comparison of the proposed policy and the optimal policy with constant power levels ('stationary' in the legend) under different numbers of users and different cross interference levels. A zero average throughput indicates that there exists no feasible policy that satisfies all the constraints in the policy design problem.

2.4.5.1 Performance Evaluation

We first compare the performance of the proposed policy with that of the optimal policy with constant power levels. The optimal policy with constant power levels (or "the optimal stationary policy") is the solution to the modified version of the design problem (2.15). First, we add an additional constraint that the power profile is constant, namely $\boldsymbol{\pi}(h^t) = \boldsymbol{a}^\star$ for all $t \geq 0$ and for all $h^t \in Y^t$. Second, we drop the incentive constraint that $\boldsymbol{\pi}$ is PPE from (2.15). Hence, the performance of the optimal stationary policy is the best that can be achieved by existing stationary policies [28]–[35], and is an upper bound for the deviation-proof stationary policies [36]–[38].

In Fig. 2.7, we compare the performance of the proposed policy and that of the optimal stationary policy under different cross interference levels and different numbers of SUs. As expected, the proposed policy outperforms the optimal stationary policy in medium to high cross interference levels (approximately when $\beta \geq 1$). In the cases of high cross interference levels ($\beta \geq 2$) and many users ($N = 5$), the stationary policy fails to meet the minimum payoff guarantees due to strong interference (indicated by zero average throughput in the figure). On the other hand, the desirable feature of the proposed policy is that the average throughput does not decrease with the increase of the cross interference level, because SUs transmit in a TDMA fashion. For the same reason, the average throughput does not change with the number of SUs.

Note that the proposed policy is infeasible (zero average throughput) when the cross interference level is very small. This is because it cannot be deviation-proof in this scenario. When the interference level is very small, SU j can deviate from $\tilde{\boldsymbol{a}}^i$ and receives a high reward $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j})$ because the interference from SU i, $\tilde{a}^i_i g_{ij}$, is small. Hence, the benefit of deviation b_{ij} is large, and the deviation is inevitable. This observation leads to an efficient way for the LSS to check the cross interference level without knowing the channel gains. If the proposed policy is infeasible, the LSS knows that the cross interference level is low, and can switch to stationary policies.

2.4.5.2 Comparison with "punish-forgive" policies proposed under perfect monitoring

We also compare the proposed policy with existing policies designed under the assumption of perfect monitoring [39]–[42]. Specifically, we consider the "punish-forgive" policy in [39]–[42], which requires SUs to switch to the punishment phase of L time slots once a deviation is detected. In the punishment phase, all the SUs transmit at the maximum power levels to create high interference to the

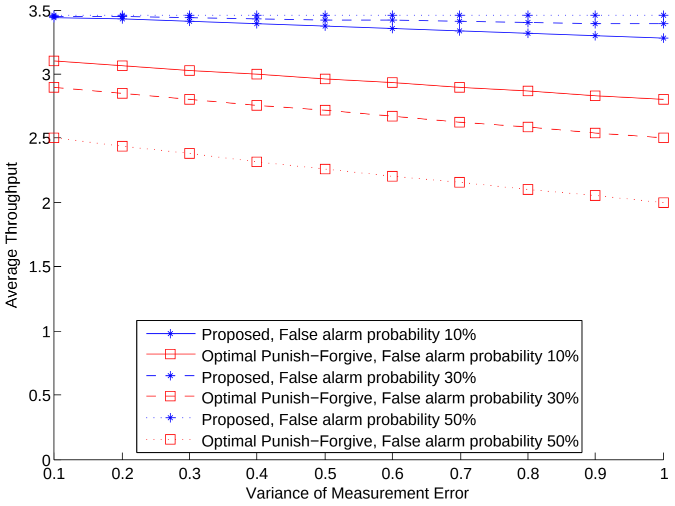


Figure 2.8: Performance comparison of the proposed policy and the punish-forgive policy with the optimal punishment length under different error variances and different false alarm probabilities.

deviator[9]. A special case of the punish-forgive policy when the punishment length $L = \infty$ [39] is the celebrated "grim-trigger" strategy in game theory literature [8]. As discussed before, the punish-forgive policy works well if the SUs can perfectly monitor the individual power levels of all the SUs, because in this case, the punishment serves as a threat and will never be carried out in the equilibrium. However, when the SUs have imperfect monitoring ability, the punishment will be carried out with some positive probability, which decreases all the SUs' average payoffs.

Fig. 2.8 shows that the proposed policy outperforms the punish-forgive policies under different variances of measurement errors and different false alarm probabilities. For each combination of the error variance and the false alarm probability, we choose the punish-forgive policy with the optimal punishment length. The performance of punish-forgive polices degrades with the increase of the error variance

[9] Note that all the SUs transmitting at the maximum power levels. For the punish-forgive policy [39]–[42], we allow the violation of the IT constraint in the punishment phase. Note that the IT constraint is never violated in the proposed policy.

and the false alarm probability, because of the increasing probability of mistakenly triggered punishments. Some interesting observation on how the performance of the proposed policy changes with the error variance and the false alarm probability is explained in details in the following subsections.

2.4.5.3 Impacts of Variances of Measurement Errors

Fig. 2.9 shows that with the increase of the variance of measurement errors, the average throughput decreases, and the SUs' patience (the discount factor) required to achieve Pareto optimal equilibrium payoffs increases. First, when the error variance increases, the intermediate IT limit I must decrease to maintain the constraint on the false alarm probability. The decrease of I leads to the decrease of SUs' maximum transmit power levels allowed, which results in the decrease of the average throughput. Another impact of the increase in the error variance is that $\rho(y_0|a_j, \tilde{\boldsymbol{a}}^i_{-j}) = \int_{x>\bar{I}-a_j g_{j0}-\tilde{a}^i_i g_{i0}} f_\varepsilon(x)dx$ increases, which leads to the increase of benefit of deviation b_{ij}. Hence, the minimum discount factor $\underline{\delta}$ increases according to Theorem 1.

2.4.5.4 Impacts of Constraints on The False Alarm Probability

Fig. 2.10 shows that with the increase of the false alarm probability limit $\bar{\Gamma}$, both the average throughput and the users' patience (the discount factor) required to achieve Pareto optimal equilibrium payoffs increase. First, with an increased false alarm probability limit, the intermediate IT limit I can increase, which leads to an increase of the SUs' maximum transmit power levels and thus an increase of the users' throughput. Meanwhile, since

$$\rho(y_0|\tilde{\boldsymbol{a}}^i) - \rho(y_0|p_j, \tilde{\boldsymbol{a}}^i_{-j}) = -\int_{\bar{I}-I-g_{j0}a_j}^{\bar{I}-I} f_\varepsilon(x)dx$$

increases when I increases, the benefit of deviation b_{ij} increases. This leads to an increase of the minimum discount factor.

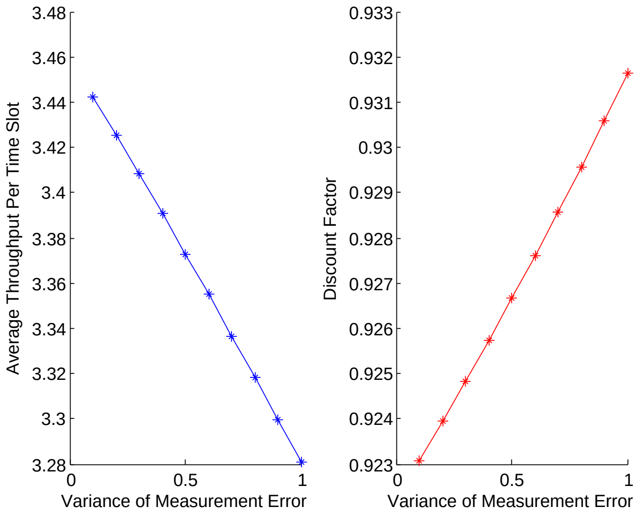


Figure 2.9: The impact of the variance of the measurement error on the performance of the proposed policy and the minimum discount factor required under which the proposed policy is deviation-proof.

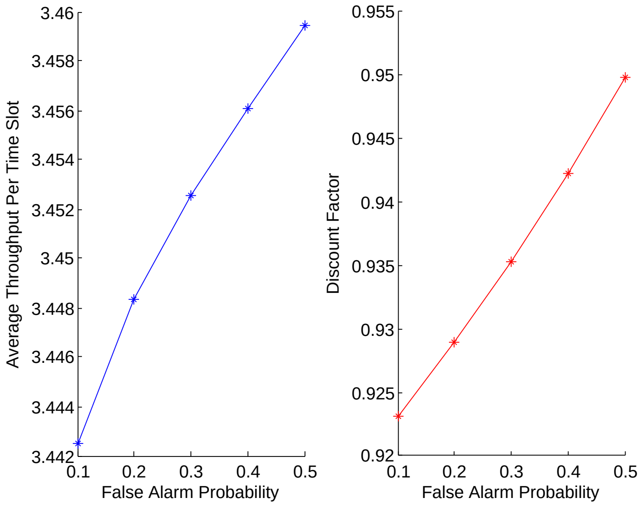


Figure 2.10: The impact of the false alarm probability on the performance of the proposed policy and the minimum discount factor required under which the proposed policy is deviation-proof.

This observation indicates an interesting design tradeoff. On one hand, a smaller false alarm probability can reduce the overhead of sending distress signals, and can also relax the requirement on SUs' patience. On the other hand, a larger false alarm probability can increase the average throughput, such that the spectrum efficiency or the revenue can increase. Our theoretical results characterize such a tradeoff, which can be used to choose the optimal intermediate IT limit I.

2.5 Applications and Extensions to Energy-Efficient Spectrum Sharing

2.5.1 Motivation

In this section, we develop a novel design framework for energy-efficient spectrum sharing among autonomous users who aim to minimize their energy consumptions subject to minimum throughput requirements. This problem is much more challenging than the throughput maximization problem studied in Section 2.4. This is because in the energy minimization problem, the users are coupled through the minimum throughput constraints, but not through the payoffs (i.e. energy expenditure). However, we can utilize and extend the ideas described before this section to solve the energy minimization problem.

We briefly discuss the difficulty in energy minimization problem and its differences from the throughput maximization problem. In the throughput maximization problem, we aim to design TDMA spectrum sharing policies that maximize the users' total throughput without considering energy efficiency. Under this design objective, each user will transmit at the maximum power level in its slot, as long as the interference temperature constraint is not violated. Hence, what we optimized was *only the transmission schedule of the users.* In energy minimization, since we aim to minimize the energy consumption subject to the minimum throughput requirements, we need to optimize *both the transmission schedule and the users' transmit power levels*, which makes the design problem more challenging.

We explain the differences in the design frameworks in details. Both design frameworks include three steps: characterization of the set of feasible operating points, selection of the optimal operating point, and the distributed implementation of the policy. The fundamental difference is in the first step, which is the most important step in the design. In the throughput maximization problem, since each user transmits at the maximum power level in its slot, we know that the set of feasible operating points lies in the hyperplane determined by each user's maximum achievable throughput. Hence, we only need to determine which portion of this particular hyperplane is achievable. On the contrary, in the energy minimization problem, since the users may not transmit at the maximum power levels in their slots, the feasible operating points lie in a *collection of hyperplanes*, each of which goes through the vector of minimum throughput requirements. Hence, it is more difficult to characterize the set of feasible operating points in the energy minimization problem. Due to the more complicated characterization of the feasible operating points, the selection of the optimal operating point (the second step) also becomes a more complicated optimization problem in the energy minimization problem (although we can prove that it can be converted to a convex optimization problem under reasonable assumptions). In summary, in the energy minimization problem, the first two steps in the design framework are fundamentally different from those in the throughput maximization problem, and are more challenging.

Both design frameworks have similar third steps: given the optimal operating point obtained in the second step, each user runs a simple and intuitive algorithm that achieves the optimal operating point in a decentralized manner. However, in

Table 2.3: Comparisons against stationary policies.

	Energy -efficient	Feedback (Overhead)	User number	Deviation -proof
[27][28][48]–[50]	No	Error-free, unquantized	Fixed	No
[36][38]	No	Error-free, unquantized (Large)	Fixed	Yes
[51]–[59]	Yes	Error-free, unquantized (Large)	Fixed	Yes
[35][60]	Yes	Error-free, unquantized (Large)	Varying	Yes
[39]–[41]	No	Error-free, unquantized (Large)	Fixed	Yes
Proposed	Yes	Erroneous, binary (One-bit)	Varying	Yes

this section, we further take the advantage of the simplicity and intuition of the algorithm, and extend it to the scenario in which PUs/SUs enter and leave the network. This makes the design framework in this section more robust to the user dynamics compared to the framework in Section 2.4.

2.5.2 Related Works

2.5.2.1 Stationary Spectrum Sharing Policies

Most existing works propose stationary spectrum sharing policies. We compare against them in Table 2.3. Note that throughout this section, the feedback is the information on interference and noise power levels sent from a user's receiver to its transmitter. The proposed nonstationary polices significantly outperform stationary policies in terms of spectrum and energy efficiencies. In addition, most existing policies require error-free and unquantized feedback, which incurs a large overhead.

2.5.2.2 Nonstationary Spectrum Sharing Policies

There have been some works that develop nonstationary policies using repeated games [42], Markov decision processes (MDPs) [61], and multi-art bandit [62]–[64]. We summarize the major differences between the existing nonstationary policies and our proposed policy in Table 2.4.

Table 2.4: Comparisons against nonstationary policies.

	[42]	[61]	[62]–[64]	Proposed
Energy -efficient	No	No	No	Yes
Power control	Yes	No	No	Yes
Users	Heterogenous	Homogenous	Homogenous	Heterogenous
Feedback (Overhead)	Error-free unquantized (Large)	Erroneous binary (One-bit)	Error-free binary (One-bit)	Erroneous binary (One-bit)
User number	Fixed	Fixed	Fixed	Varying
Deviation-proof	Yes	No	No	Yes

2.5.3 Model and Problem Formulation

2.5.3.1 Model

The model is very similar to the spectrum sharing model in Section 2.4, with a few important differences.

First, we allow the existence of multiple PUs, instead of a single PU as in Section 2.4. Specifically, we consider a cognitive radio network that consists of m primary users and n secondary users transmitting in a single frequency channel. The set of PUs and that of SUs are denoted by $M \triangleq \{1, 2, \ldots, m\}$ and $N \triangleq \{m+1, m+2, \ldots, m+n\}$, respectively.

Second, we include the PUs' power control problem in the design framework, in order to improve the energy efficiency of the PUs. Specifically, we model the PUs' actions as their transmit power levels $a_i \in A_i$ for $i \in N$. In contrast, in Section 2.4, we abstracted the PU as an interference temperature constraint and did not optimize its power control problem. The optimization of PUs' power control is extremely important when there are multiple PUs, because multiple PUs may cause large interference to each other if their power control is not optimized.

Finally, we consider a different design problem, namely the energy minimization problem. Before writing down the design problem, we define the users' long-term discounted average energy consumption as

$$P_i(\boldsymbol{\pi}) = \mathbb{E}_{h^0,h^1,\ldots} \left\{ (1-\delta) \sum_{t=0}^{\infty} \delta^t \cdot \pi_i(h^t)) \right\}.$$

The energy efficiency criterion can be represented by a function defined on all the users' average energy consumptions, $E(P_1(\boldsymbol{\pi}), \ldots, P_{m+n}(\boldsymbol{\pi}))$. Note, importantly, that the energy efficiency criterion can also reflect the priority of the PUs over the SUs. For example, the energy efficiency criterion can be the weighted sum of all the users' energy consumptions, i.e. $E(P_1(\boldsymbol{\pi}), \ldots, P_{m+n}(\boldsymbol{\pi})) = \sum_{i \in M \cup N} w_i \cdot P_i(\boldsymbol{\pi})$ with $w_i \geq 0$ and $\sum_{i \in M \cup N} w_i = 1$. Each user i's weight w_i indicates the importance of this user. We can set higher weights for PUs and lower weights for SUs.

Then the energy minimization problem can be formalized as follows:

$$\begin{aligned} \min_{\boldsymbol{\pi}} \quad & E(P_1(\boldsymbol{\pi}), \ldots, P_{m+n}(\boldsymbol{\pi})) \\ s.t. \quad & U_i(\boldsymbol{\pi}) \geq U_i^{\min}, \ \forall i \in M \cup N, \end{aligned} \tag{2.17}$$

where $U_i^{\min}$ is the minimum throughput requirement of user i.

As we have discussed before, it is much more challenging to sovle the energy minimization problem than the throughput maximization problem. We have also derived convergence results under dynamic entry and exit of users. Due to space limitation, we refer interested readers to [3] for more details.

2.5.4 Simulation Results

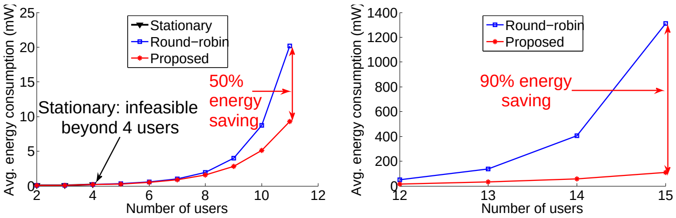
We demonstrate the performance gain of our spectrum sharing policy over existing policies, and validate our theoretical analysis through numerical results. Throughout this section, we use the following system parameters by default unless we change some of them explicitly. The noise powers at all the users' receivers are 0.05 W. For simplicity, we assume that the direct channel gains have the same distribution $g_{ii} \sim \mathcal{CN}(0, 1), \forall i$, and the cross channel gains have the same distribution $g_{ij} \sim \mathcal{CN}(0, 0.25), \forall i \neq j$. The users have the same minimum throughput requirement of 1 bits/s/Hz. The discount factor is 0.95. The interference temperature threshold is $\theta = 1$ W. The measurement error ε is Gaussian distributed with zeros mean and variance 0.1. The energy efficiency criterion is the average energy consumption across users.

2.5.4.1 Comparisons Against Existing Policies

First, assuming that the population is fixed, we compare the proposed policy against the optimal stationary policy in [54]–[35], and the optimal round-robin policy with cycle length $L = M + N$ (i.e. each user gets one slot in a cycle). We compare the energy efficiency of the policies as the number of users increase in Fig. 2.11. Each data point plotted is the average of 1000 channel realizations. First, we can see that the stationary policy becomes infeasible when the number of users is more than 4. In contrast, the round-robin and proposed policies remain feasible when the number of users increases. Second, the proposed policy achieves significant energy saving compared to the round-robin policy, especially when the number of users is large. Specifically, it achieves 50% and 90% energy saving compared to the round-robin policy when the number of users is 11 and 15, respectively. These are exactly the deployment scenarios where improvements in spectrum and energy efficiency are much needed.

2.5.4.2 Adapting to Users Entering and Leaving the Network

We demonstrate how the proposed policy can seamlessly adapt to the entry and exit of PUs/SUs. We consider a network with 10 PUs and 2 SUs initially. The PUs' minimum throughput requirements range from 0.2 bits/s/Hz to 0.38 bits/s/Hz with 0.02 bits/s/Hz increments, namely PU n has a minimum throughput re-



(a) Small numbers of users. (b) Large numbers of users.

Figure 2.11: Energy efficiency of the stationary, round-robin, and proposed policies under different numbers of users.

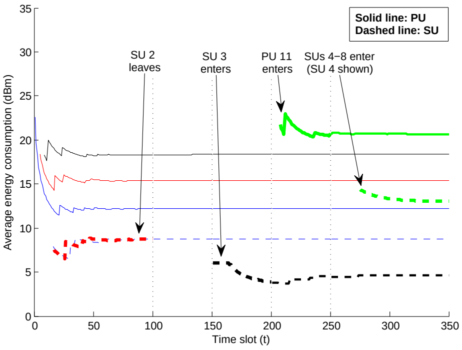


Figure 2.12: Dynamics of average energy consumption with users entering and leaving the network. At $t = 0$, there are 10 PUs and 2 SUs. SU 2 leaves at $t = 100$. SU 3 enters at $t = 150$. PU 11 enters at $t = 200$. SUs 4–8 enter at $t = 250$. We only show PUs 1, 5, 9, 11 (solid lines) and SUs 1, 2, 3, 4 (dashed lines) in the figure.

quirement of $0.2 + (n - 1) * 0.02$ bits/s/Hz. The SUs' have the same minimum throughput requirement of 0.1 bits/s/Hz. We show the dynamics of average energy consumptions and throughput of several PUs and all the SUs in Fig. 2.12 and Fig. 2.13, respectively.

In the first 100 time slots, we can see that all the users quickly achieve the minimum throughput requirements at around $t = 50$. PUs have different energy consumptions because of their different minimum throughput requirements. The two SUs converge to the same average energy consumption and average through-

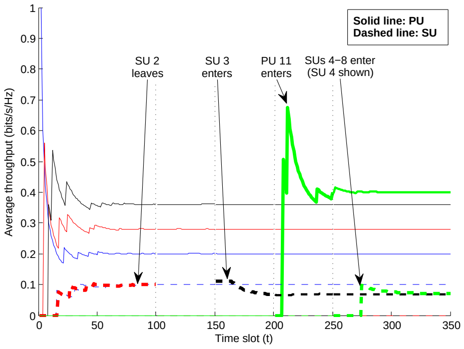


Figure 2.13: Dynamics of average throughput under the same dynamics of the entry and exit of users as in Fig. 2.12.

put. There are SUs leaving ($t = 100$) and entering ($t = 150, 250$), and a PU entering ($t = 200$). We can see that during the entire process, the PUs/SUs that are initially in the system maintain the same throughput and energy consumption. The new PU (PU 11) has a higher energy consumption, because of its higher minimum throughput requirement (0.4 bits/s/Hz), and because of the limited transmission opportunities left for it. SU 3, however, does not need a higher energy consumption because it occupies the time slots originally assigned to SU 2, who left the network at $t = 100$. But SU 4 does need a higher energy consumption, because there are more SUs and less transmission opportunities in the network after $t = 250$.

2.6 Conclusion

In this chapter, we studied a large class of repeated games with imperfect monitoring, in which the players have strong negative externality among each other and have very limited (i.e. binary) and imperfect monitoring. Our theoretical results diverge from much of the familiar literature on repeated games with imperfect public monitoring. We obtain stronger conclusions about efficient PPE (bounds on the discount factor, explicitly constructive strategies).

We applied our theoretical framework to throughput maximization problems in spectrum sharing, and extend our framework to energy minimization problems. In both settings, our proposed framework significantly outperform the state-of-the-art spectrum sharing policies.

2.7 Appendix

The proof of Proposition 1 is immediate and omitted.

Proof of Proposition 2 Fix an active player i and an inactive player j. Set

$$\begin{aligned} A(i,j) &= \Big\{a_j \in A_j : u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)\Big\} \\ B(i,j) &= \Big\{a_j \in A_j : u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) < u_j(\tilde{\boldsymbol{a}}^i), \rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) < \rho(y^i_b|\tilde{\boldsymbol{a}}^i)\Big\} \end{aligned}$$

If either of $A(i,j)$ or $B(i,j)$ is empty then $\alpha(i,j) \leq \beta(i,j)$ by default, so assume in what follows that neither of $A(i,j)$, $B(i,j)$ is empty.

Fix a discount factor $\delta \in (0,1)$ and let $\boldsymbol{\pi}$ be PPE that achieves an efficient payoff. Assume that i is active following some history: $\boldsymbol{\pi}(h) = \tilde{\boldsymbol{a}}^i$ for some h. Because $\boldsymbol{\pi}$ achieves an efficient payoff, we can decompose the payoff v following h as the weighted sum of the current payoff from $\tilde{\boldsymbol{a}}^i$ and the continuation payoff assuming that players follow $\boldsymbol{\pi}$; because $\boldsymbol{\pi}$ is a PPE, the incentive compatibility condition for all players j must obtain. Hence for all $a_j \in A_j$ we have

$$\begin{aligned} v_j &= (1-\delta)u_j(\tilde{\boldsymbol{a}}^i) + \delta \sum_{y \in Y} \rho(y|\tilde{\boldsymbol{a}}^i)\gamma_j(y) \\ &\geq (1-\delta)u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) + \delta \sum_{y \in Y} \rho(y|a_j, \tilde{\boldsymbol{a}}^i_{-j})\gamma_j(y). \end{aligned}$$

Substituting probabilities for the good and bad signals yields

$$\begin{aligned} v_j &= (1-\delta)u_j(\tilde{\boldsymbol{a}}^i) + \delta\Big[\rho(y^i_g|\tilde{\boldsymbol{a}}^i)\gamma_j(y^i_g) + \rho(y^i_b|\tilde{\boldsymbol{a}}^i)\gamma_j(y^i_b)\Big] \qquad (2.18) \\ &\geq (1-\delta)u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) + \delta\Big[\rho(y^i_g|a_j, \tilde{\boldsymbol{a}}^i_{-j})\gamma_j(y^i_g) + \rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j})\gamma_j(y^i_b)\Big]. \end{aligned}$$

Rearranging yields

$$\Big[\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)\Big]\Big[\gamma_j(y_g^i) - \gamma_j(y_b^i)\Big]\Big[\frac{\delta}{1-\delta}\Big] \geq \Big[u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - u_j(\tilde{\boldsymbol{a}}^i)\Big].$$

Now suppose $j \neq i$ is an inactive player. If $a_j \in A(i,j)$ then $\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i) > 0$ (by Assumption 4) so

$$\Big[\gamma_j(y_g^i) - \gamma_j(y_b^i)\Big]\Big[\frac{\delta}{1-\delta}\Big] \geq \frac{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - u_j(\tilde{\boldsymbol{a}}^i)}{\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)}. \tag{2.19}$$

If $a_j \in B(i,j)$ then $\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i) < 0$ (by definition) so

$$\Big[\gamma_j(y_g^i) - \gamma_j(y_b^i)\Big]\Big[\frac{\delta}{1-\delta}\Big] \leq \frac{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - u_j(\tilde{\boldsymbol{a}}^i)}{\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)}. \tag{2.20}$$

Taking the sup over $a_j \in A(i,j)$ in (2.19) and the inf over $a_j \in B(i,j)$ in (2.20) yields $\alpha(i,j) \leq \beta(i,j)$ as desired. □

Proof of Proposition 3 As above, we assume i is active following the history h and that v is the payoff following h. Fix $a_i \in A_i$. By definition, $u_i(\tilde{\boldsymbol{a}}^i) > u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i)$. With respect to probabilities, there are two possibilities. If $\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) \leq \rho(y_b^i|\tilde{\boldsymbol{a}}^i)$ then we immediately have

$$\tilde{v}_i^i - u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i) \geq \frac{1}{\lambda_i}\sum_{j\neq i}\lambda_j\alpha(i,j)[\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)],$$

because the left-hand side is positive and the right-hand side is non-positive ($\alpha(i,j)$ is positive due to Assumption 4). If $\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) > \rho(y_b^i|\tilde{\boldsymbol{a}}^i)$ we proceed as follows.

We begin with (2.18) but now we apply it to the active user i, so that for all $a_i \in A_i$ we have

$$\begin{aligned} v_i &= (1-\delta)u_i(\tilde{\boldsymbol{a}}^i) + \delta\Big[\rho(y_g^i|\tilde{\boldsymbol{a}}^i)\gamma_i(y_g^i) + \rho(y_b^i|\tilde{\boldsymbol{a}}^i)\gamma_i(y_b^i)\Big] \\ &\geq (1-\delta)u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i) + \delta\Big[(\rho(y_g^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i)\gamma_i(y_g^i) + \rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i)\gamma_i(y_b^i)\Big]. \end{aligned}$$

Rearranging yields

$$\gamma_i(y_g^i) - \gamma_i(y_b^i) \ \geq \left[\frac{1-\delta}{\delta}\right]\left[\frac{u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i) - u_i(\tilde{\boldsymbol{a}}^i)}{\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)}\right].$$

Because continuation payoffs are in V, which lies in the hyperplane H, the continuation payoffs for the active user can be expressed in terms of the continuation payoffs for the inactive users as

$$\gamma_i(y) = \frac{1}{\lambda_i}\left[1 - \sum_{j \neq i} \lambda_j \gamma_j(y)\right].$$

Hence

$$\gamma_i(y_g^i) - \gamma_i(y_b^i) = -\frac{1}{\lambda_i}\sum_{j \neq i} \lambda_j [\gamma_j(y_g^i) - \gamma_j(y_b^i)].$$

Applying the incentive compatibility constraints for the inactive users implies that for each $a_j \in A(i,j)$ we have

$$\gamma_j(y_g^i) - \gamma_j(y_b^i) \geq \left[\frac{1-\delta}{\delta}\right]\left[\frac{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - u_j(\tilde{\boldsymbol{a}}^i)}{\rho(y_b^i|a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)}\right].$$

In particular

$$\gamma_j(y_g^i) - \gamma_j(y_b^i) \geq \left[\frac{1-\delta}{\delta}\right]\alpha(i,j),$$

and hence

$$\gamma_i(y_g^i) - \gamma_i(y_b^i) \leq -\frac{1}{\lambda_i}\left[\frac{1-\delta}{\delta}\right]\left[\sum_{j \neq i} \lambda_j \alpha(i,j)\right].$$

Putting these all together, canceling the factor $[1-\delta]/\delta$ and remembering that we are in the case $\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) > \rho(y_b^i|\tilde{\boldsymbol{a}}^i)$ yields

$$\tilde{v}_i^i - u_i(a_i, \tilde{\boldsymbol{a}}_{-i}^i) \geq \frac{1}{\lambda_i}\sum_{j \neq i} \lambda_j \alpha(i,j)[\rho(y_b^i|a_i, \tilde{\boldsymbol{a}}_{-i}^i) - \rho(y_b^i|\tilde{\boldsymbol{a}}^i)]$$

which is the desired result. □

Proof of Theorem 1 Assume that V_μ is regular and not an extreme point, and is a self-generating set; we verify Conditions 1-4 in turn.

Since V_μ is regular, for each $i \in N$ there is a payoff profile $\hat{v}^i \in V_\mu$ with the property that $\hat{v}_j^i = \mu_j$ for each $j \neq i$. Necessarily, $\hat{v}^i$ is the unique such point and $\hat{v}^i = \arg\max\{v_i : v \in V_\mu\}$. Because V lies in the hyperplane H we have

$$\hat{v}_j^i = \begin{cases} \mu_j & \text{if} \quad j \neq i \\ \frac{1}{\lambda_i}\left(1 - \sum_{k \neq i} \lambda_k \mu_k\right) & \text{if} \quad j = i \end{cases}.$$

Because V_μ is self-generating, we can decompose $\hat{v}^i$:

$$\hat{v}^i = (1-\delta)u(\tilde{\boldsymbol{a}}^k) + \delta \sum_y \rho(y|\tilde{\boldsymbol{a}}^k)\gamma(y) \tag{2.21}$$

for some $\tilde{\boldsymbol{a}}^k$. If $k \neq i$ then (because $V_\mu \neq \{\tilde{v}^k\}$) we must have $\mu_k < \tilde{v}^k_k$ which implies that $\gamma_k(y) < \mu_k$ for some y; since continuation payoffs must lie in V_μ this is a contradiction. Hence in the decomposition (2.21) we must have $\tilde{\boldsymbol{a}}^k = \tilde{\boldsymbol{a}}^i$. In other words, each player i must be active, in order to decompose $\hat{v}^i$. So Propositions 2 and 3 yield Conditions 1 and 2.

It is convenient to first establish the following inequality on μ_j on the way to establishing the bounds in Condition 3.

$$\mu_j > \max_{i \neq j} \tilde{v}^i_j \text{ for all } j \in N$$

To see this, suppose to the contrary that there exists a i, j such that $\mu_j \leq \tilde{v}^i_j$. Consider i's preferred payoff profile $\hat{v}^i$ in V_μ. Because decomposing $\hat{v}^i$ requires that we use $\tilde{\boldsymbol{a}}^i$, it follows that

$$\mu_j = (1-\delta) \cdot \tilde{v}^i_j + \delta \cdot \sum_y \rho(y|\tilde{\boldsymbol{a}}^i)\gamma_j(y)$$

If $\mu_j < \tilde{v}^i_j$ then $\sum_{y \in Y} \rho(y|\tilde{\boldsymbol{a}}^i)\gamma_i(y) < \mu_j$ and so $\gamma_j(y) < \mu_j$ for some y. This contradicts that fact that $\gamma(y) \in V_\mu$. If $\mu_j = \tilde{v}^i_j$, we must have $\sum_y \rho(y|\tilde{\boldsymbol{a}}^i)\gamma_j(y) = \mu_j$. Since $\gamma_j(y) \geq \mu_j$ for all y, we must have $\gamma_j(y^i_g) = \gamma_j(y^i_b) = \mu_j$. By assumption, player j has a currently profitable deviation a_j so that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$, which implies that the continuation payoff $\gamma_j(y^i_g) = \gamma_j(y^i_b) = \mu_j$ cannot satisfy the incentive compatibility constraints. Hence, we must have $\mu_j > \tilde{v}^i_j$ as asserted.

With all this in hand we derive Condition 3. To do this, we suppose i is active and examine the decomposition of the inactive player j's payoff in greater detail. Because $\mu_j > \tilde{v}^i_j$ and $v_j \geq \mu_j$ for every $v \in V_\mu$ we certainly have $v_j > \tilde{v}^i_j$. We can write j's incentive compatibility condition as

$$\begin{aligned} v_j &= (1-\delta)\cdot \tilde{v}_j^i + \delta \cdot \sum_{y\in Y} \rho(y|\tilde{\boldsymbol{a}}^i)\cdot \gamma_j(y) \\ &\geq (1-\delta)\cdot u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) + \delta \cdot \sum_{y\in Y} \rho(y|a_j, \tilde{\boldsymbol{a}}_{-j}^i)\cdot \gamma_j(y). \end{aligned} \tag{2.22}$$

From the equality constraint in (2.22), we can solve for the discount factor δ as

$$\delta = \frac{v_j - \tilde{v}_j^i}{\sum_{y\in Y} \gamma_j(y)\rho(y|\tilde{\boldsymbol{a}}^i) - \tilde{v}_j^i}$$

(Note that the denominator can never be zero and the above equation is well defined, because $v_j > \tilde{v}_j^i$ implies that $\sum_{y\in Y} \gamma_j(y)\rho(y|\tilde{\boldsymbol{a}}^i) > \tilde{v}_j^i$.) We can then eliminate the discount factor δ in the inequality of (2.22). Since $v_j > \tilde{v}_j^i$, we can obtain equivalent inequalities, depending on whether a_j is a profitable or unprofitable current deviation:

- If $u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) > \tilde{v}_j^i$ then

$$\begin{aligned} v_j \leq \sum_{y\in Y} \gamma_j(y) \Bigg[& \left(1 - \frac{v_j - \tilde{v}_j^i}{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \tilde{v}_j^i}\right) \rho(y|\tilde{\boldsymbol{a}}^i) \\ & + \frac{v_j - \tilde{v}_j^i}{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \tilde{v}_j^i} \rho(y|a_j, \tilde{\boldsymbol{a}}_{-j}^i) \Bigg] \end{aligned} \tag{2.23}$$

- If $u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) < \tilde{v}_j^i$ then

$$\begin{aligned} v_j \geq \sum_{y\in Y} \gamma_j(y) \Bigg[& \left(1 - \frac{v_j - \tilde{v}_j^i}{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \tilde{v}_j^i}\right) \rho(y|\tilde{\boldsymbol{a}}^i) \\ & + \frac{v_j - \tilde{v}_j^i}{u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) - \tilde{v}_j^i} \rho(y|a_j, \tilde{\boldsymbol{a}}_{-j}^i) \Bigg] \end{aligned} \tag{2.24}$$

For notational convenience, write the coefficient of $\gamma_j(y_g^i)$ in the above inequalities as

$$\begin{aligned} c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j}) &\triangleq \left(1 - \frac{v_j - \tilde{v}^i_j}{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j}\right) \rho(y^i_g|\tilde{\boldsymbol{a}}^i) \\ &\quad + \left(\frac{v_j - \tilde{v}^i_j}{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j}\right) \rho(y^i_g|a_j, \tilde{\boldsymbol{a}}^i_{-j}) \\ &= \rho(y^i_g|\tilde{\boldsymbol{a}}^i) + (v_j - \tilde{v}^i_j) \left(\frac{\rho(y^i_g|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_g|\tilde{\boldsymbol{a}}^i)}{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j}\right) \\ &= \rho(y^i_g|\tilde{\boldsymbol{a}}^i) - (v_j - \tilde{v}^i_j) \left(\frac{\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)}{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j}\right) \end{aligned}$$

According to (2.23), if $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > \tilde{v}^i_j$ then

$$c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j}) \cdot \gamma_j(y^i_g) + \Big[1 - c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j})\Big]\gamma_j(y^i_b) \leq v_j \tag{2.25}$$

Since $\gamma_j(y^i_g) > \gamma_j(y^i_b)$, this is true if and only if

$$\kappa^+_{ij} \cdot \gamma_j(y^i_g) + (1 - \kappa^+_{ij}) \cdot \gamma_j(y^i_b) \leq v_j, \tag{2.26}$$

where $\kappa^+_{ij} \triangleq \sup\{c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j}) : a_j \in A_j : u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > \tilde{v}^i_j\}$. (Fulfilling the inequalities (2.25) for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$ is equivalent to fulfilling the single inequality (2.26). If (2.26) is satisfied, then the inequalities (2.25) are satisfied for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$ because $\gamma_j(y^i_g) > \gamma_j(y^i_b)$ and $\kappa^+_{ij} \geq c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j})$ for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$. Conversely, if the inequalities (2.25) are satisfied for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$ and (2.26) were violated, so that $\kappa^+_{ij} \cdot \gamma_j(y^i_g) + (1 - \kappa^+_{ij}) \cdot \gamma_j(y^i_b) > v_j$, then we can find a $\kappa'_{ij} < \kappa^+_{ij}$ such that $\kappa'_{ij} \cdot \gamma_j(y^i_g) + (1 - \kappa'_{ij}) \cdot \gamma_j(y^i_b) > v_j$. Based on the definition of the supremum, there exists at least a a'_j such that $u_j(a'_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$ and $c_{ij}(a'_j, \tilde{\boldsymbol{a}}^i_{-j}) > c'_{ij}$, which means that $c_{ij}(a'_j, \tilde{\boldsymbol{a}}^i_{-j}) \cdot \gamma_j(y^i_g) + (1 - c_{ij}(a'_j, \tilde{\boldsymbol{a}}^i_{-j})) \cdot \gamma_j(y^i_b) > v_j$. This contradicts the fact that the inequalities (2.26) are fulfilled for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) > u_j(\tilde{\boldsymbol{a}}^i)$.)

Similarly, according to (2.24), for all a_j such that $u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) < \tilde{v}^i_j$, we must have

$$c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j})\gamma_j(y^i_g) + [1 - c_{ij}(a_j, \tilde{\boldsymbol{a}}^i_{-j})]\gamma_j(y^i_b) \geq v_j.$$

Since $\gamma_j(y_g^i) > \gamma_j(y_b^i)$, the above requirement is fulfilled if and only if

$$\kappa_{ij}^- \cdot \gamma_j(y_g^i) + (1 - \kappa_{ij}^-) \cdot \gamma_j(y_b^i) \geq v_j,$$

where $\kappa_{ij}^- \triangleq \inf \left\{ c_{ij}(a_j, \tilde{\boldsymbol{a}}_{-j}^i) : a_j \in A_j, u_j(a_j, \tilde{\boldsymbol{a}}_{-j}^i) < \tilde{v}_j^i \right\}$. Hence, the decomposition (2.22) for user $j \neq i$ can be simplified as:

$$\begin{aligned} \rho(y_g^i|\tilde{\boldsymbol{a}}^i) \cdot \gamma_j(y_g^i) + [1 - \rho(y_g^i|\tilde{\boldsymbol{a}}^i)]\gamma_j(y_b^i) &= \tilde{v}_j^i + \frac{v_j - \tilde{v}_j^i}{\delta} \\ \kappa_{ij}^+ \gamma_j(y_g^i) + (1 - \kappa_{ij}^+) \cdot \gamma_j(y_b^i) &\leq v_j \\ \kappa_{ij}^- \gamma_j(y_g^i) + (1 - \kappa_{ij}^-) \cdot \gamma_j(y_b^i) &\geq v_j \end{aligned} \tag{2.27}$$

Keep in mind that the various continuation values γ and the expressions $\kappa_{ij}^+, \kappa_{ij}^-$ depend on v_j; where necessary we write the dependence explicitly. Note that there could be many $\gamma_j(y_g^i)$ and $\gamma_j(y_b^i)$ that satisfy (2.27). For a given discount factor δ, we call all the continuation payoffs that satisfy (2.27) *feasible* – but whether particular continuation values lie in V_μ depends on the discount factor.

We assert that $\kappa_{ij}^+(\mu_j) \leq 0$ for all $i \in N$ and for all $j \neq i$. To see this, we look again at player i's preferred payoff profile $\hat{v}^i$ in V_μ, which is necessarily decomposed by $\tilde{\boldsymbol{a}}^i$. We look at the following constraint for player $j \neq i$ in (2.27):

$$\kappa_{ij}^+ \ \gamma_j(y_g^i) + (1 - \kappa_{ij}^+) \ \gamma_j(y_b^i) \leq \mu_j.$$

Suppose that $\kappa_{ij}^+(\mu_j) > 0$. Since player j has a currently profitable deviation from $\tilde{\boldsymbol{a}}^i$, we must set $\gamma_j(y_g^i) > \gamma_j(y_b^i)$. Then to satisfy the above inequality, we must have $\gamma_j(y_b^i) < \mu_j$. In other words, when $\kappa_{ij}^+(\mu_j) > 0$, all the feasible continuation payoffs of player j must be outside V_μ. This contradicts the fact that V_μ is self-generating so the assertion follows.

The definition of $\kappa^{+}_{ij}(\mu_j)$ and the fact that $\kappa^{+}_{ij}(\mu_j) \leq 0$ entail that

$$\begin{aligned}
\kappa^{+}_{ij}(\mu_j) &= \rho(y^i_g|\tilde{\boldsymbol{a}}^i) - (\mu_j - \tilde{v}^i_j) \inf_{a_j \in A(i,j)} \left[\frac{\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)}{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j} \right] \\
&= \rho(y^i_g|\tilde{\boldsymbol{a}}^i) - (\mu_j - \tilde{v}^i_j) \left[\frac{1}{\sup_{a_j \in A(i,j)} \left(\frac{u_j(a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \tilde{v}^i_j}{\rho(y^i_b|a_j, \tilde{\boldsymbol{a}}^i_{-j}) - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)} \right)} \right] \\
&= \rho(y^i_g|\tilde{\boldsymbol{a}}^i) - (\mu_j - \tilde{v}^i_j) \left[\frac{1}{\alpha(i,j)} \right] \\
&\leq 0
\end{aligned}$$

This provides a lower bound on μ_j:

$$\mu_j \geq \tilde{v}^i_j + \alpha(i,j)\rho(y^i_g|\tilde{\boldsymbol{a}}^i) = \tilde{v}^i_j + \alpha(i,j)[1 - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)]$$

This bound must hold for every $i \in N$ and every $j \neq i$. Hence, we have

$$\mu_j \geq \max_{i \neq j} \left(\tilde{v}^i_j + \alpha(i,j)[1 - \rho(y^i_b|\tilde{\boldsymbol{a}}^i)] \right)$$

which is Condition 3.

Now we derive Condition 4 (the necessary condition on the discount factor). The minimum discount factor $\underline{\delta}_\mu$ required for V_μ to be a self-generating set solves the optimization problem

$$\underline{\delta}_\mu = \max_{v \in V_\mu} \delta \qquad \text{subject to } v \in \mathscr{B}(V_\mu; \delta)$$

where $\mathscr{B}(V_\mu; \delta)$ is the set of payoff profiles that can be decomposed on V_μ under discount factor δ. Since $\mathscr{B}(V_\mu; \delta) = \cup_{i \in N} \mathscr{B}(V_\mu; \delta, \tilde{\boldsymbol{a}}^i)$, the above optimization problem can be reformulated as

$$\underline{\delta}_\mu = \max_{v \in V_\mu} \min_{i \in N} \delta \qquad \text{subject to } v \in \mathscr{B}(V_\mu; \delta, \tilde{\boldsymbol{a}}^i). \tag{2.28}$$

To solve the optimization problem (2.28), we explicitly express the constraint $v \in \mathscr{B}(V_\mu; \delta, \tilde{\boldsymbol{a}}^i)$ using the results derived above.

Some intuition may be useful. Suppose that i is active and j is an inactive player. Recall that player j's feasible $\gamma_j(y^i_g)$ and $\gamma_j(y^i_b)$ must satisfy (2.27). There

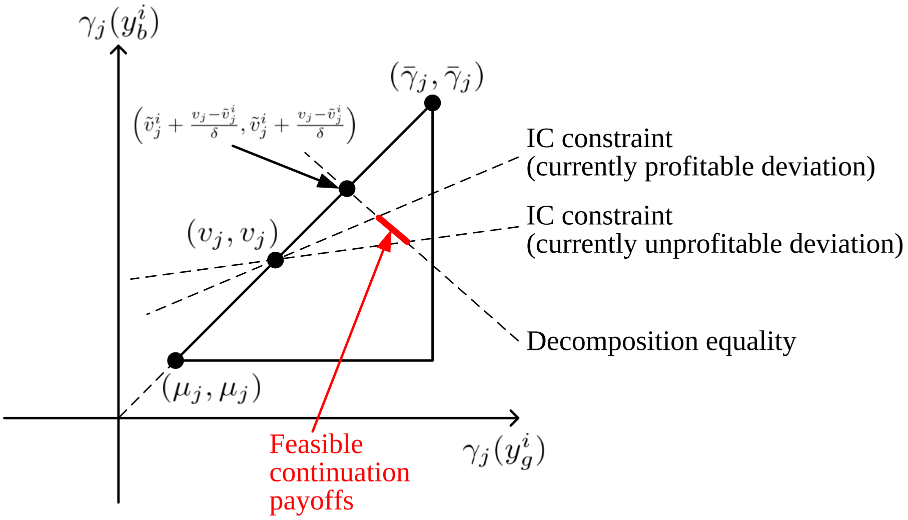

$\gamma_j(y_b^i)$
$(\bar{\gamma}_j, \bar{\gamma}_j)$
$\left(\tilde{v}_j^i + \frac{v_j - \tilde{v}_j^i}{\delta}, \tilde{v}_j^i + \frac{v_j - \tilde{v}_j^i}{\delta}\right)$
IC constraint
(currently profitable deviation)
IC constraint
(currently unprofitable deviation)
(v_j, v_j)
Decomposition equality
(μ_j, μ_j)
Feasible
continuation
payoffs
$\gamma_j(y_g^i)$

Figure 2.14: Illustrations of the feasible continuation payoffs when $\kappa_{ij}^+ \leq 0$. $\bar{\gamma}_j = \frac{1}{\lambda_j}\left(1 - \sum_{k \neq j} \lambda_k \mu_k\right)$.

are many $\gamma_j(y_g^i)$ and $\gamma_j(y_b^i)$ that satisfy (2.27). In Fig. 2.14, we show the feasible continuation payoffs that satisfy (2.27) when $\kappa_{ij}^+(v_j) \leq 0$. We can see that all the continuation payoffs on the heavy line segment are feasible. The line segment is on the line that represents the decomposition equality $\rho(y_g^i|\tilde{\boldsymbol{a}}^i) \cdot \gamma_j(y_g^i) + (1 - \rho(y_g^i|\tilde{\boldsymbol{a}}^i)) \cdot \gamma_j(y_b^i) = \tilde{v}_j^i + \frac{v_j - \tilde{v}_j^i}{\delta}$, and is bounded by the IC constraint on currently profitable deviations $\kappa_{ij}^+ \cdot \gamma_j(y_g^i) + (1 - \kappa_{ij}^+) \cdot \gamma_j(y_b^i) \leq v_j$ and the IC constraint on currently unprofitable deviations $\kappa_{ij}^- \cdot \gamma_j(y_g^i) + (1 - \kappa_{ij}^-) \cdot \gamma_j(y_b^i) \geq v_j$. Among all the feasible continuation payoffs, denoted $\gamma'(y)$, we choose the one, denoted $\gamma^*(y)$, such that for all $j \neq i$, $\gamma_j^*(y_g^i)$ and $\gamma_j^*(y_b^i)$ make the IC constraint on currently profitable deviations in (2.27) binding. This is because under the same discount factor δ, if there is any feasible continuation payoff $\gamma'(y)$ in the self-generating set, the one that makes the IC constraint on currently profitable deviations binding is also in the self-generating set. The reason is that, as can be seen from Fig. 2.14, the continuation payoff $\gamma_j^*(y)$ that makes the IC constraint binding has the smallest $\gamma_j^*(y_g^i) = \min \gamma_j'(y_g^i)$ and the largest $\gamma_j^*(y_b^i) = \max \gamma_j'(y_b^i)$. Formally we establish the following Lemma.

Lemma 1 *Fix a payoff profile v and a discount factor δ. Suppose that v is decomposed by $\tilde{\boldsymbol{a}}^i$. If there are any feasible continuation payoffs $\gamma'(y_g^i) \in V_\mu$ and $\gamma'(y_b^i) \in V_\mu$ that satisfy (2.27) for all $j \neq i$, there there exist feasible continuation payoffs $\gamma^*(y_g^i) \in V_\mu$ and $\gamma^*(y_b^i) \in V_\mu$ such that the IC constraint on currently profitable deviations in (2.27) is binding for all $j \neq i$.*

Proof 1 *Given feasible continuation payoffs $\gamma'(y_g^i) \in V_\mu$ and $\gamma'(y_b^i) \in V_\mu$, we construct $\gamma^*(y_g^i) \in V_\mu$ and $\gamma^*(y_b^i) \in V_\mu$ that are feasible and make the IC constraint on currently profitable deviations in (2.27) binding for all $j \neq i$.*

Specifically, we set $\gamma_j^*(y_g^i)$ and $\gamma_j^*(y_b^i)$ such that the IC constraint on currently profitable deviations in (2.27) is binding. Such $\gamma_j^*(y_g^i)$ and $\gamma_j^*(y_b^i)$ have the following property: $\gamma_j^*(y_g^i) \leq \gamma_j'(y_g^i)$ and $\gamma_j^*(y_b^i) \geq \gamma_j'(y_b^i)$ for all $\gamma_j'(y_g^i)$ and $\gamma_j'(y_b^i)$ that satisfy (2.27). We prove this property by contradiction. Suppose that there exist $\gamma_j'(y_g^i)$ and $\gamma_j'(y_b^i)$ that satisfy (2.27) and $\gamma_j'(y_g^i) = \gamma_j^*(y_g^i) - \Delta$ with $\Delta > 0$. Based on the decomposition equality, we have

$$\gamma_j'(y_b^i) = \gamma_j^*(y_b^i) + \left(\frac{\rho(y_g^i|\tilde{\boldsymbol{a}}^i)}{1 - \rho(y_g^i|\tilde{\boldsymbol{a}}^i)} \right) \Delta$$

We can see that the IC constraint on currently profitable deviations is violated:

$$\begin{aligned} & \kappa_{ij}^+ \, \gamma_j'(y_g^i) + (1 - \kappa_{ij}^+) \, \gamma_j'(y_b^i) \\ = \ & \kappa_{ij}^+ \, \gamma_j^*(y_g^i) + (1 - \kappa_{ij}^+) \, \gamma_j^*(y_b^i) + \left[-\kappa_{ij}^+ \, \Delta + (1 - \kappa_{ij}^+) \left(\frac{\rho(y_g^i|\tilde{\boldsymbol{a}}^i)}{1 - \rho(y_g^i|\tilde{\boldsymbol{a}}^i)} \right) \Delta \right] \\ = \ & v_j + (1 - \kappa_{ij}^+) \left[\frac{\rho(y_g^i|\tilde{\boldsymbol{a}}^i)}{1 - \rho(y_g^i|\tilde{\boldsymbol{a}}^i)} - \frac{\kappa_{ij}^+}{1 - \kappa_{ij}^+} \right] \Delta \\ > \ & v_j \end{aligned}$$

where the last inequality results from $\kappa_{ij}^+ \leq 0$. Hence, we have $\gamma_j^*(y_g^i) \leq \gamma_j'(y_g^i)$ and $\gamma_j^*(y_b^i) \geq \gamma_j'(y_b^i)$ for all $\gamma_j'(y_g^i)$ and $\gamma_j'(y_b^i)$ that satisfy (2.27).

Next, we prove that if $\gamma'(y) \in V_\mu$, then $\gamma^*(y) \in V_\mu$. To prove $\gamma^*(y) \in V_\mu$, we need to show that $\gamma_j^*(y_g^i) \geq \mu_j$ and $\gamma_j^*(y_b^i) \geq \mu_j$ for all $j \in N$. For $j \neq i$, we have

$\gamma_j^*(y_g^i) \geq \gamma_j^*(y_b^i) \geq \gamma_j'(y_b^i) \geq \mu_j$. *For i, we have*

$$\gamma_i^*(y_g^i) = \frac{1}{\lambda_i}\left(1 - \sum_{j\neq i}\lambda_j\gamma_j^*(y_g^i)\right) \geq \frac{1}{\lambda_i}\left(1 - \sum_{j\neq i}\lambda_j\gamma_j'(y_g^i)\right) = \gamma_i'(y_g^i) \geq \mu_i$$

This proves the lemma.

Using this Lemma, we can calculate the continuation payoffs of the inactive player $j \neq i$:

$$\begin{aligned}
\gamma_j(y_g^i) &= \frac{\left(\frac{1}{\delta}(1-\kappa_{ij}^+) - [1-\rho(y_g^i|\tilde{\boldsymbol{a}}^i)]\right)v_j - (\frac{1}{\delta}-1)(1-\kappa_{ij}^+)\tilde{v}_j^i}{\rho(y_g^i|\tilde{\boldsymbol{a}}^i) - \kappa_{ij}^+} \\
&= \frac{v_j}{\delta} - \left(\frac{1-\delta}{\delta}\right)\tilde{v}_j^i + \left(\frac{1-\delta}{\delta}\right)[1-\rho(y_g^i|\tilde{\boldsymbol{a}}^i)]\alpha(i,j), \\
\gamma_j(y_b^i) &= \frac{\left[\rho(y_g^i|\tilde{\boldsymbol{a}}^i) - \frac{1}{\delta}\kappa_{ij}^+\right]v_j + (\frac{1}{\delta}-1)\kappa_{ij}^+\,\tilde{v}_j^i}{\rho(y_g^i|\tilde{\boldsymbol{a}}^i) - \kappa_{ij}^+} \\
&= \frac{v_j}{\delta} - \left(\frac{1-\delta}{\delta}\right)\tilde{v}_j^i - \left(\frac{1-\delta}{\delta}\right)\rho(y_g^i|\tilde{\boldsymbol{a}}^i)\alpha(i,j).
\end{aligned}$$

The active player's continuation payoffs can be determined based on the inactive players' continuation payoffs since $\gamma(y) \in V$. We calculate the active player i's continuation payoffs as

$$\begin{aligned}
\gamma_i(y_g^i) &= \frac{v_i}{\delta} - \left(\frac{1-\delta}{\delta}\right)\tilde{v}_i^i - \left(\frac{1-\delta}{\delta}\right)[1-\rho(y_g^i|\tilde{\boldsymbol{a}}^i)]\frac{1}{\lambda_i}\sum_{j\neq i}\lambda_j\alpha(i,j), \\
\gamma_i(y_b^i) &= \frac{v_i}{\delta} - \left(\frac{1-\delta}{\delta}\right)\tilde{v}_i^i + \left(\frac{1-\delta}{\delta}\right)\rho(y_g^i|\tilde{\boldsymbol{a}}^i)\frac{1}{\lambda_i}\sum_{j\neq i}\lambda_j\alpha(i,j)
\end{aligned}$$

Hence, the constraint $v \in \mathscr{B}(V_\mu; \delta, \tilde{\boldsymbol{a}}^i)$ on discount factor δ is equivalent to

$$\gamma(y) \in V_\mu \text{ for all } y \in Y \Leftrightarrow \gamma_i(y) \geq \mu_i \text{ for all } i \in N, y \in Y$$

Since $\kappa_{ij}^+(\mu_j) \leq 0$, we have $\gamma_j(y) \geq v_j$ for all $y \in Y$, which means that $\gamma_j(y) \geq \mu_j$ for all $y \in Y$. Hence, we only need the discount factor to have the property that $\gamma_i(y) \geq \mu_i$ for all $y \in Y$. Since $\gamma_i(y_g^i) < \gamma_i(y_b^i)$, we need $\gamma_i(y_g^i) \geq \mu_i$, which leads to

$$\delta \geq \frac{1}{1 + \lambda_i(v_i - \mu_i)/\left[\lambda_i(\tilde{v}_i^i - v_i) + \sum_{j\neq i}\lambda_j \cdot (1-\rho(y_g^i|\tilde{\boldsymbol{a}}^i))\alpha(i,j)\right]}.$$

Hence, the optimization problem (2.28) is equivalent to

$$\underline{\delta}_{\mu} = \max_{v \in V_{\mu}} \min_{i \in N} x_i(v) \tag{2.29}$$

where

$$x_i(v) \triangleq \frac{1}{1 + \lambda_i(v_i - \mu_i) / \left(\lambda_i(\tilde{v}_i^i - v_i) + \sum_{j \neq i} \lambda_j [1 - \rho(y_g^i | \tilde{\boldsymbol{a}}^i)] \alpha(i,j) \right)}$$

Since $x_i(v)$ is decreasing in v_i, the payoff v^* that maximizes $\min_{i \in N} x_i(v)$ must satisfy $x_i(v^*) = x_j(v^*)$ for all i and j. Now we find the payoff v^* such that $x_i(v^*) = x_j(v^*)$ for all i and j.

Define

$$z \triangleq \frac{\lambda_i(v_i^* - \mu_i)}{\lambda_i(\tilde{v}_i^i - v_i^*) + \sum\limits_{j \neq i} \lambda_j [1 - \rho(y_g^i | \tilde{\boldsymbol{a}}^i)] \alpha(i,j)}$$

Then we have

$$\lambda_i(1 + z) v_i^* = \lambda_i(\mu_i + z \tilde{v}_i^i) - z \sum_{j \neq i} \lambda_j [1 - \rho(y_g^i | \tilde{\boldsymbol{a}}^i)] \alpha(i,j)$$

from which it follows that

$$z = \frac{1 - \sum\limits_i \lambda_i \mu_i}{\sum\limits_i \left(\lambda_i \tilde{v}_i^i + \sum\limits_{j \neq i} \lambda_j [1 - \rho(y_g^i | \tilde{\boldsymbol{a}}^i)] \alpha(i,j) \right) - 1}$$

Hence, the minimum discount factor is $\underline{\delta}(\mu) = \frac{1}{1+z}$; substituting the definition of z yields Condition 4. This completes the proof that these Conditions 1-4 are necessary for V_μ to be a self-generating set.

It remains to show that these necessary Conditions are also sufficient, which is accomplished in the proof of Theorem 2. This completes the proof of Theorem 1. □

Proof of Theorem 2 In view of the results of APS, it suffices to show that under Conditions 1-4 of Theorem 1, the algorithm yields a decomposition of each target vector $v(t) \in V_\mu$.

For convenience, we summarize how we decompose any $v(t) \in V_\mu$ based on the algorithm as follows. We first find the active player i according to

$$i = \max_j \left\{ \arg\max_{j \in N} d_j(v(t)) \right\},$$

where

$$d_j(v(t)) = \frac{\lambda_j [v_j(t) - \mu_j]}{\lambda_j [\tilde{v}_j^j - v_j(t)] + \sum_{k \neq j} \lambda_k \, \alpha(j,k) \rho(y_b^j | \tilde{\boldsymbol{a}}^j)}.$$

Then in the algorithm we update $v(t+1)$ based on the signal y^t. Essentially, we are assigning the continuation payoff vectors $\gamma(y)$ following the signals. Specifically, the continuation payoff vectors are assigned as follows:

$$\begin{aligned}
\gamma_i(y_g^i) &= \tilde{v}_i^i + (1/\delta)(v_i(t) - \tilde{v}_i^i) - (1/\delta - 1)(1/\lambda_i) \sum_{j \neq i} \lambda_j \alpha(i,j) \rho(y_b^i | \tilde{\boldsymbol{a}}^i), \\
\gamma_j(y_g^i) &= \tilde{v}_j^i + (1/\delta)(v_j(t) - \tilde{v}_j^i) + (1/\delta - 1)\alpha(i,j)\rho(y_b^i | \tilde{\boldsymbol{a}}^i), \forall j \neq i,
\end{aligned}$$

and

$$\begin{aligned}
\gamma_i(y_b^i) &= \tilde{v}_i^i + (1/\delta)(v_i(t) - \tilde{v}_i^i) + (1/\delta - 1)(1/\lambda_i) \sum_{j \neq i} \lambda_j \alpha(i,j) \rho(y_g^i | \tilde{\boldsymbol{a}}^i), \\
\gamma_j(y_b^i) &= \tilde{v}_j^i + (1/\delta)(v_j(t) - \tilde{v}_j^i) - (1/\delta - 1)\alpha(i,j)\rho(y_g^i | \tilde{\boldsymbol{a}}^i), \forall j \neq i.
\end{aligned}$$

We need to verify that under Conditions 1-4 of Theorem 1, the above continuation payoff vectors $\gamma(y_g^i)$ and $\gamma(y_b^i)$ satisfy 1) the decomposition equalities, 2) the incentive compatibility constraints, and 3) that $\gamma(y_g^i) \in V_\mu$ and $\gamma(y_b^i) \in V_\mu$.

It is straightforward to check that the decomposition equalities are satisfied. The incentive compatibility constraints for the inactive players j reduce to Condition 1, and those for the active player i reduce to Condition 2.

We proceed to verify that $\gamma(y_g^i) \in V_\mu$ and $\gamma(y_b^i) \in V_\mu$. It is straightforward to verify that $\gamma(y_g^i) \in V$ and $\gamma(y_b^i) \in V$. We only need to show $\gamma_j(y_g^i) \geq \mu_j$ and $\gamma_j(y_b^i) \geq \mu_j$ for all $j \in N$. Since $\alpha(i,j) > 0$, we can observe that $\gamma_j(y_g^i) > \gamma_j(y_b^i)$ for all $j \neq i$ and $\gamma_i(y_g^i) < \gamma_i(y_b^i)$. Hence, it suffices to show $\gamma_j(y_b^i) \geq \mu_j$ for all $j \neq i$ and $\gamma_i(y_g^i) \geq \mu_i$.

For any inactive player j, we have

$$\begin{aligned}
& \gamma_j(y_b^i) \geq \mu_j \\
\Leftrightarrow \quad & \tilde{v}_j^i + (1/\delta)(v_j(t) - \tilde{v}_j^i) - (1/\delta - 1)\alpha(i,j)\rho(y_g^i|\tilde{\boldsymbol{a}}^i) \geq \mu_j \\
\Leftrightarrow \quad & (1/\delta)v_j(t) - \mu_j \geq (1/\delta - 1)\tilde{v}_j^i + (1/\delta - 1)\alpha(i,j)\rho(y_g^i|\tilde{\boldsymbol{a}}^i) \\
\Leftarrow \quad & (1/\delta)\mu_j - \mu_j \geq (1/\delta - 1)\tilde{v}_j^i + (1/\delta - 1)\alpha(i,j)\rho(y_g^i|\tilde{\boldsymbol{a}}^i) \\
\Leftrightarrow \quad & \mu_j \geq \tilde{v}_j^i + \alpha(i,j)\rho(y_g^i|\tilde{\boldsymbol{a}}^i) \\
\Leftarrow \quad & \text{Condition 3 of Theorem 1.}
\end{aligned}$$

For the active player i, we have

$$\begin{aligned}
& \gamma_i(y_g^i) \geq \mu_i \\
\Leftrightarrow \quad & \tilde{v}_i^i + (1/\delta)(v_i(t) - \tilde{v}_i^i) - (1/\delta - 1)(1/\lambda_i)\sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i) \geq \mu_i \\
\Leftrightarrow \quad & (1/\delta)\left[v_i(t) - \tilde{v}_i^i - (1/\lambda_i)\sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i)\right] \geq \mu_i - \tilde{v}_i^i - (1/\lambda_i)\sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i) \\
\Leftrightarrow \quad & \delta \geq \frac{\tilde{v}_i^i - v_i(t) + (1/\lambda_i)\sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i)}{\tilde{v}_i^i - \mu_i + (1/\lambda_i)\sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i)} \\
\Leftrightarrow \quad & \delta \geq \frac{1}{1 + \lambda_i(v_i(t) - \mu_i) / \left[\lambda_i(\tilde{v}_i^i - v_i(t)) + \sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i)\right]} \\
\Leftarrow \quad & \underline{\delta}_\mu \geq \frac{1}{1 + \lambda_i(v_i(t) - \mu_i) / \left[\lambda_i(\tilde{v}_i^i - v_i(t)) + \sum_{j \neq i} \lambda_j \alpha(i,j)\rho(y_b^i|\tilde{\boldsymbol{a}}^i)\right]} \\
\Leftrightarrow \quad & \underline{\delta}_\mu \geq \frac{1}{1 + d_i(v(t))}.
\end{aligned}$$

According to the proof of Theorem 1, the above $\underline{\delta}_\mu$ in Condition 4 of Theorem 1 is calcualted by solving the optimization problem (2.29), which is equivalent to

$$\underline{\delta}_\mu = \max_{v \in V_\mu} \min_{j \in N} \frac{1}{1 + d_j(v)}.$$

From the above, we have $\underline{\delta}_\mu \geq \min_{j \in N} \frac{1}{1+d_j(v)}$ for any $v \in V_\mu$. Under the given $v(t)$, the active player i is chosen such that $d_i(v(t))$ is the largest (i.e. $\frac{1}{1+d_i(v(t))}$ is the smallest). Hence, we have $\underline{\delta}_\mu \geq \min_{j \in N} \frac{1}{1+d_j(v(t))} = \frac{1}{1+d_i(v(t))}$. This yields $\gamma_i(y_g^i) \geq \mu_i$. □

Proof of Theorem 3 Given a parameter ξ, the algorithm uses the target vector as the continuation value to compute N indicators; let $\Xi(0)$ be the set of parameters for which no two of these indicators are equal. For each parameter in $\Xi(0)$, the algorithm computes continuation values following the good signal and the bad signal and then uses each of these continuation values to compute N indicators; let $\Xi(1) \subset \Xi(0)$ be the set of parameters for which no two of these indicators are equal. Proceeding by induction, we define a decreasing sequence of sets $\Xi(0) \supset \Xi(1) \supset \cdots \supset \Xi(T)$; let Ξ_T be the complement of $\Xi(T)$. Notice that the indicators are continuous functions of the parameters so the ordering of the indicators is locally constant provided no two indicators are equal. Hence for each $\xi \in \Xi(T) = \Xi \setminus \Xi(T)$ then there is a small open neighborhood Z of ξ so that if $\xi' \in Z$ then the strategies $\boldsymbol{\pi}_{\xi'}, \boldsymbol{\pi}_{\xi}$ generate the same ordering of indicators in each of the first T periods. In particular, $\boldsymbol{\pi}_{\xi'}(h) = \boldsymbol{\pi}_{\xi}(h)$ for each history $h \in \mathcal{H}^T$; that is, $\xi \to \boldsymbol{\pi}_{\xi}^T$ is locally constant on the complement of Ξ_T. It remains only to show that Ξ_T is closed and has measure 0. In fact, Ξ_T is a finite union of lower-dimensional submanifolds; this is a consequence of general facts about semi-algebraic sets and the observation that all the indicators are continuous semi-algebraic functions of the parameters, no two of which coincide on any open set. See [23], [24]. □

Proof of Theorem 4 Propositions 2, 3 show that Conditions 1, 2 are necessary conditions for the existence of an efficient PPE for *any* discount factor. Suppose therefore that Conditions 1,2 are satisfied. It is easily checked that the following definitions of μ_1^*, μ_2^* guarantee that Condition 3 of Theorem 1 are satisfied:

$$\mu_1^* = \tilde{v}_1^2 + \alpha(2,1)[1-\rho(y_b^2|\tilde{\boldsymbol{a}}^2)],\ \mu_2^* = \tilde{v}_2^1 + \alpha(1,2)[1-\rho(y_b^1|\tilde{\boldsymbol{a}}^1)].$$

Finally, if

$$\delta \geq \delta^* \triangleq \left(1 + \frac{1-\lambda_1\mu_1^* - \lambda_2\mu_2^*}{\sum\limits_i \left[\lambda_i \tilde{v}_i^i + \lambda_{-i}\,\alpha(i,-i)\,\rho(y_b^i|\tilde{\boldsymbol{a}}^i)\right] - 1}\right)^{-1},$$

then Condition 4 of Theorem 1 is also satisfied. It follows from Theorem 1 that for each $\delta \geq \delta^*$, V_{μ^*} is a self-generating set, so every target vector in V_{μ^*} can be achieved in a PPE. Hence $E(\delta) \supset V_{\mu^*}$ for every $\delta \in [\delta^*, 1)$. To see that $V_{\mu^*} = E(\delta)$ for every $\delta \in [\delta^*, 1)$, simply note that for each δ the set $E(\delta)$ is closed and convex, hence an interval, hence of the form V_μ for some μ. However, Condition 3 of Theorem 1 guarantees that $\mu \geq \mu^*$ which completes the proof. □

CHAPTER 3

Resource Sharing With Decentralized Information

3.1 Introduction

Power systems are currently undergoing drastic changes on both the supply and the demand side. On the supply side, renewable energy (e.g. wind energy, solar energy) is increasingly used to reduce the environmental damage caused by conventional energy generation; however, this often introduces high fluctuations in the amount of energy generated. To cope with these fluctuations (uncertainty) in energy generation, energy storage is increasingly used as an important solution [70]. In this chapter, we determine the *optimal* economic dispatch strategies and the *optimal* demand side management (DSM) strategies in the presence of energy storage.

Specifically, we consider a power system consisting of several energy generators on the supply side, an independent system operator (ISO) that operates the system, and multiple aggregators and their customers on the demand side. On the supply side, the ISO receives energy purchase requests from the aggregators as well as reports of (parameterized) energy generation cost functions from the generators and, based on these, dispatches the energy generators and determines the unit energy prices. On the demand side, the aggregators are located in different geographical areas (i.e. nodes/buses of the power network) and provide energy for their customers (e.g. households, office buildings) in the neighborhood. In the literature, the term "DSM" has been broadly used for different decision problems on the demand side. For example, some papers (see [65]–[68] for representative papers) focus on the interaction between one aggregator and its customers, and refer to DSM as determining the power consumption schedules of the users. Some papers [69]–[77] focus on how multiple aggregators [69]–[72] or a single aggregator [73]–[77] purchase(s) energy from the ISO based on the energy consumption requests from their customers. Our work pertains to the second category.

The key feature that sets apart our work from most existing works [69]–[72] is that all the decision makers in the system are *foresighted*. Each aggregator seeks to minimize its *long-term* cost, consisting of its operational cost of energy storage and its payment for energy purchase. In contrast, in most existing works [69]–[72], the aggregators are *myopic* and seek to minimize their *short-term* (e.g. one-day or even hourly) cost. In the presence of energy storage, foresighted DSM strategies can achieve much lower costs than myopic DSM strategies because the current decisions of the aggregators will affect their future costs. For example, an aggregator can purchase more energy from the ISO than that requested from its customers, and store the unused energy in the energy storage for future use, if it anticipates that the future energy price will be high. Hence, the current purchase from the aggregators will affect how much they will purchase in the future. In this case, it is optimal for the entities to make *foresighted* decisions, taking into account the impact of their current decisions on the future. Since the aggregators deploy foresighted DSM strategies, it is also optimal for the ISO to make foresighted economic dispatch, in order to minimize the *long-term* total cost of the system, consisting of the long-term cost of energy generation and the aggregators' long-term operational cost. Note that although some works [73]–[77] assume that the aggregator has energy storage and is foresighted, they study the decision problem of a *single* aggregator and do not consider the economic dispatch problem of the ISO. When there are multiple aggregators in the system (which is the case in practice), this approach neglects the impact of aggregators' decisions on each other, which leads to suboptimal solutions in terms of minimizing the total cost of the system.

When the ISO and *multiple* aggregators make *foresighted* decisions, it is difficult to obtain the optimal foresighted strategies for two reasons. First, the information is decentralized. The total cost depends on the generation cost functions (e.g. the speed of wind for wind energy generation, the amount of sunshine for solar energy generation, and so on), the status of the transmission lines (e.g. the flow capacity of the transmission lines), the amount of electricity in the energy storage, and the demand from the customers, all of which change over time due to supply and demand uncertainty. However, no entity knows all the above information: the ISO knows only the generation cost functions and the status of the transmission lines, and each aggregator knows only the status of its own energy storage and the demand of its own customers. Hence, the DSM strategy needs to be decentralized, such that each entity can make decisions solely based on its *locally-available* information[1]. Second, the aggregators are coupled in a manner that is unknown to them and changing over time. Specifically, each aggregator's purchase affects the prices[2], and thus the payments of the other aggregators. However, the price is determined by the ISO based on the generation cost functions and the status of the transmission lines, neither of which is known to any aggregator. Hence, each aggregator does not know how its purchase will influence the price, which makes it difficult for the aggregator to make the optimal decision.

To overcome the challenges due to information decentralization and compli-

[1] Even if the aggregators are willing to share all their private information with the ISO such that the ISO can make centralized decisions, the resulting centralized decision problem becomes intractable quickly as the size of the power network grows (e.g. for the IEEE 118-bus system). For large power networks, it is not only desirable, but also necessary to have an decentralized solution in which each entity is able to solve one subproblem after decomposing the intractable centralized problem.

[2] In our model, an aggregator is responsible for all the customers on a node/bus of the power network. Hence, its purchase is significant enough to influence the locational marginal prices (LMPs).

cated coupling, we propose a decentralized DSM strategy based on conjectured prices. Specifically, each aggregator makes decisions based on its conjectured price, and its local information about the status of its energy storage and the demand from its customers. In other words, each aggregator summarizes all the information that is not available to it into its conjectured price. Note, however, that the price is determined based on the generation cost functions and the status of the transmission lines, which is only known to the ISO. Hence, the aggregators' conjectured prices are determined by the ISO. We propose a simple online algorithm for the ISO to update the conjectured prices based on its local information, and prove that by using the algorithm, the ISO obtains the optimal conjectured prices under which the aggregators' (foresighted) best responses minimize the total cost of the system.

In addition, we consider the scenario in which the entities do not know their state transition probabilities a priori. For example, the aggregators may not know how their customers' demands change, and the ISO may not know how the status of the transmission lines evolves. We propose an online learning algorithm for the entities to learn and converge to the optimal DSM strategy. The proposed online learning algorithm utilizes the concept of post-decision states [84][85] and exploits the independence of the state dynamics in the system, which results in a faster learning speed and a better run-time performance compared to conventional learning algorithms such as Q-learning [86]. Simulations demonstrate that the learning algorithm can learn the optimal policy, and can adapt to the nonstationary dynamics of the system.

In summary, the major contributions of our work are as follows:

- We rigorously formalize the *long-term* interaction among the ISO and *multiple* aggregators with energy storage as a multi-user Markov decision process (MU-MDP).

- We propose the *optimal* decentralized *foresighted* demand-side management with energy storage, which minimizes the total system cost. Simulations demonstrate that our proposed solution can reduce the total system cost by 60% and 30%, compared to the state-of-the-art myopic solutions and foresighted solutions based on Lyapunov optimization, respectively.

- To the best of our knowledge, our proposed theoretical framework is the *first* one that *optimally* solve the MU-MDP (in terms of minimizing the total cost) we formulate.

- We propose an learning algorithm that allows the entities to reach the optimal solution even without statistical knowledge of the system dynamics, and track the optimal solution even when the underlying transition probabilities of system dynamics are time-varying (i.e. when the system dynamics are not Markovian).

The rest of this chapter is organized as follows. We provide a more detailed literature review in Section 4.2. We introduce the system model in Section 4.3, and then formulate the design problem in Section 3.4. We describe the proposed optimal decentralized DSM strategy in Section 3.5. Through simulations, we validate our theoretical results and demonstrate the performance gains of the proposed strategy in Section 4.6. Finally, we conclude the chapter in Section 4.7.

3.2 Related Works

3.2.1 Related Works on Demand-Side Management

In Table 4.1, we categorize existing works on DSM in power systems in terms of the assumption on demand/supply uncertain and whether energy storage is adopted when designing the DSM strategy. In short, there exists no work that designs *foresighted* DSM strategies for *multiple* aggregators who seek to minimize

Table 3.1: Comparisons With Related Works on Demand-Side Management.

	Energy storage	Time horizon	Foresighted	# of Aggregators	Supply uncertainty	Demand Uncertainty
[65][66][69]	No	1 day	No	Multiple	No	No
[67]	No	1 day	No	Multiple	Yes	No
[68]	No	1 day	No	Multiple	No	Yes
[70]	Yes	1 day	No	Multiple	No	No
[71][72]	Yes	1 day	No	Multiple	Yes	Yes
[73][74]	Yes	Infinite	Yes	Single	No	Yes
[75]–[77]	Yes	Infinite	Yes	Single	Yes	Yes
Proposed	Yes	Infinite	Yes	Multiple	Yes	Yes

their *long-term* costs (i.e. the average cost in a time horizon much longer than one day).

Some works [65]–[72] proposed *myopic* DSM strategies for *multiple* aggregators who seek to minimize their costs *within one day*. With energy storage, the foresighted strategies which minimize the long-term cost can greatly outperform the myopic strategies. For example, in a myopic strategy, the aggregator may tend to purchase as little power as possible as long as the demand is fulfilled, in order to minimize the current operational cost of its energy storage. However, the optimal strategy should take into consideration the future price, and balance the trade-off between the current operational cost and the future saving in energy purchase.

Other works [73]–[77] proposed *foresighted* DSM strategies for a *single* aggregator who seeks to minimize their costs *in an infinite time horizon*. In practice, the power system has many aggregators. With multiple aggregators, it is inefficient for each aggregator to simply adopt the optimal foresighted strategy designed under the assumption of a single aggregator. As we will show in the simulation, with multiple aggregators, the total cost achieved by such a simple adaptation of the optimal single-aggregator strategy is much higher (up to 30%) than the proposed optimal solution. This is because the single-aggregator strategy aims at

Table 3.2: Comparisons With Related Mathematical Frameworks.

	MDP	MU-MDP [78][79]	Lyapunov Optimization [73]–[77]	Stochastic Control [80]	Stochastic Games [81]	This work
# of decision makers	Single	Multiple	Single	Multiple	Multiple	Multiple
Decentralized information	N/A	Yes	N/A	Yes	No	Yes
Coupling among users	N/A	Weak	N/A	Strong	Strong	Strong
Optimal	Yes	Yes	Yes	No	Yes	Yes
Constructive	Yes	Yes	Yes	Yes	No	Yes

achieving individual minimum cost, instead of the total cost. Due to the coupling among the aggregators, the outcome in which individual costs are minimized may be very different from the outcome in which the total cost is minimized.

We will provide a more technical comparison with the existing works in Table 3.6, after we described the proposed framework.

3.2.2 Related Theoretical Frameworks

Decision making in a dynamically changing environment has been studied and formulated as Markov decision processes (MDPs). Most MDP-based works have been dedicated to solving single-user decision problems. There have been few works [78][79] on multi-user MDPs (MU-MDPs). The works on MU-MDPs [78][79] focus on *weakly coupled* MU-MDPs, where the term "weakly coupled" is coined by [78] to denote the assumption that one user's action does not directly affect the others' current payoffs. The users are coupled only through some linking constraints on their actions (for example, the sum of their actions, e.g. the sum data rate, should not exceed some threshold, e.g. the available bandwidth). However, once a user chooses its own action, its current payoff and its state transition are determined and do not depend on the other users' actions. In contrast, in this work, the users are *strongly coupled*, namely one user's action directly affect the others' current payoffs. For example, one aggregator's energy purchase affects the unit price of energy, which has impact on the other aggregators' payments. There are few works in stochastic control that model the users' interaction as strongly coupled [80]. However, the main focus of [80] is to prove the existence of a Nash equilibrium (NE) strategy. There is no performance analysis/guarantee of the proposed NE strategy.

The interaction among users with strong coupling is modeled as a stochastic game [81] in the game theory literature. However, in standard stochastic games, the state of the system is known to all the players. Hence, we cannot model the interaction of entities in our work as a stochastic game, because different entities have different private states unknown to the others. In addition, the results in [81] are not constructive. They focus on *what* payoff profiles are achievable, but not *how* to achieve those payoff profiles (i.e. their methods are not constructive). In contrast, we propose an algorithm to compute the optimal strategy profile.

In Table 4.2, we compare our work with existing theoretical frameworks. Note that we will provide a more technical comparison with the Lyapunov optimization and MU-MDP frameworks in Table 3.6, after we described the proposed framework.

3.3 System Model

3.3.1 The System Setup

We consider a smart grid with one ISO indexed by 0, G generators indexed by $g = 1, 2, \ldots, G$, I aggregators indexed by $i = 1, 2, \ldots, I$, and L transmission lines (see Fig. 3.1 for an illustration). The ISO schedules the energy generation of generators and determines the unit prices of energy for the aggregators. The generators provide the ISO with the information of their energy generation cost

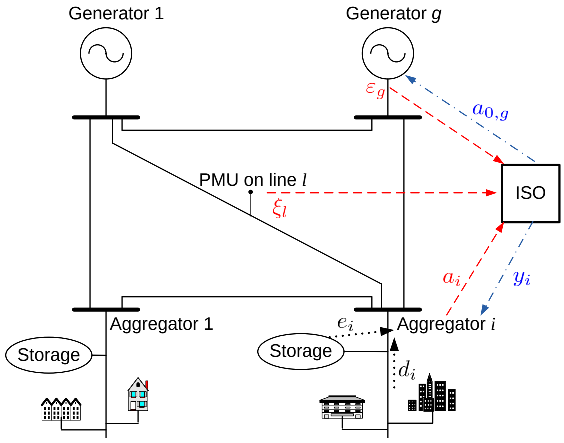


Figure 3.1: The system model of the smart grid. The information flow to the ISO is denoted by red dashed lines, the information flow to the aggregators is denoted by black dotted lines, and the information flow sent from the ISO is denoted by blue dash-dot lines.

functions, based on which the ISO can minimize the total cost of the system. Since the ISO determines how much energy each generator should produce, we do not model generators as decision makers in the system; instead, we abstract them by their energy generation cost functions. Each aggregator, equipped with energy storage, provides energy for its customers (e.g. residential households, commercial buildings), and determines how much energy to buy from the ISO. In summary, the decision makers (or the entities) in the system are the ISO and the I aggregators. We denote the set of aggregators by $\mathcal{I} = \{1, \ldots, I\}$. In the following, we refer to the ISO or an aggregator generally as entity $i \in \{0\} \cup \mathcal{I}$, with entity 0 being the ISO and entity $i \in \mathcal{I}$ being aggregator i.

As discussed before, different entities possess different local information. Specifically, the ISO receives reports of the energy generation cost functions, denoted by $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_G)$, from the generators, and measures the status of the transmission lines such as the phases or capacities, denoted by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, by using the phasor measurement units (PMUs). We summarize the energy gener-

Table 3.3: Each entity's knowledge, and the corresponding results.

Knowledge	Solutions
ISO: realizations and transition probabilities of s_0 Aggregator i: realizations and transition probabilities of s_i	Sec. 3.5.2
ISO: only realizations of s_0 Aggregator i: only realizations of s_i	Sec. 3.5.3

ation cost functions and the status of the transmission lines into the ISO's state $s_0 = (\boldsymbol{\varepsilon}, \boldsymbol{\xi}) \in S_0$, which is unknown to the aggregators[3]. Each aggregator receives energy consumption requests from its customers, and manages its energy storage. We summarize the aggregate demand d_i from aggregator i's customers and the amount e_i of energy in aggregator i's storage into aggregator i's state $s_i = (d_i, e_i) \in S_i$, which is only known to aggregator i. We assume that all the sets $S_0, \ldots, S_I$ of states are finite. We highlight which information is available to which entity in Table 3.3.

The ISO's action is how much energy each generator should produce, denoted by $a_0 \in A_0(s_0) \subset \mathbb{R}_+^G$, where $A_0(s_0)$ is the action set under state s_0. Each aggregator i's action is how much energy to purchase from the ISO, denoted by $a_i \in A_i(s_i) \subset \mathbb{R}_+$, where $A_i(s_i)$ is the action set under state s_i. We denote the joint action profile of the aggregators as $\boldsymbol{a} = (a_1, \ldots, a_N)$, and the joint action profile of all the aggregators other than i as $\boldsymbol{a}_{-i}$. We allow the action set to be dependent on the current state, in order to impose constraints on each entity's behavior. For example, we require that the aggregator must fulfill its customers' demand. Hence, given aggregator i's state $s_i = (d_i, e_i)$, we have

$$A_i(s_i) = \{a_i : d_i \leq a_i + e_i \leq E_i\},$$

where E_i is the maximum capacity of aggregator i's storage. We could also

[3] Note that the status of the transmission lines are the phase and capacity. Such information, and sometimes even the topology of the network of transmission lines, is critical energy infrastructure information (CEII) and is unknown to the aggregators. Note also that an aggregator may know the congestion on the bus in which it is located. However, the congestion is not the state of the ISO.

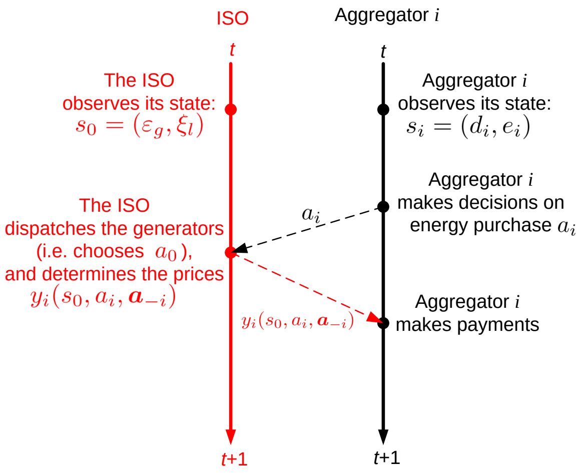

ISO
Aggregator *i*
t
t
The ISO observes its state: $s_0 = (\varepsilon_g, \xi_l)$
Aggregator *i* observes its state: $s_i = (d_i, e_i)$
Aggregator *i* makes decisions on energy purchase a_i
a_i
The ISO dispatches the generators (i.e. chooses a_0), and determines the prices $y_i(s_0, a_i, \boldsymbol{a}_{-i})$
$y_i(s_0, a_i, \boldsymbol{a}_{-i})$
Aggregator *i* makes payments
t+1
t+1

Figure 3.2: Illustration of the interaction between the ISO and aggregator i (i.e. their decision making and information exchange) in one period.

impose constraints on the charging/discharging rates of the storage. The aggregator charges the storage when $a_i > d_i$ and discharges the storage when $a_i < d_i$. Hence, the maximum charging/discharging rate constraint can be written as $-r_i^{discharge} \leq a_i - d_i \leq r_i^{charge}$, where $r_i^{discharge}$ and r_i^{charge} are the maximum discharging and charging rates, respectively.

We divide time into periods $t = 0, 1, 2, \ldots$, where the duration of a period is determined by how fast the demand or supply changes or how frequently the energy trading decisions are made. In each period t, the entities act as follows (see Fig. 3.2 for illustration):

- The ISO observes its state s_0.
- Each aggregator i observes its state s_i.
- Each aggregator i chooses its action a_i, namely how much energy to purchase from the ISO, and tells its amount a_i of energy purchase to the ISO.
- Based on its state s_0 and the aggregators' action profile $\boldsymbol{a}$, the ISO de-

termines the price[4] $y_i(s_0, \boldsymbol{a}) \in Y_i$ of electricity at each aggregator i, and announces it to each aggregator i. The ISO also determines its action a_0, namely how much energy each generator should produce.

- Each aggregator i pays $y_i(s_0, \boldsymbol{a}) \cdot a_i$ to the ISO.[5]

The instantaneous cost of each entity depends on its current state and its current action. Each aggregator i's total cost consists of two parts: the operational cost and the payment. Each aggregator i's operational cost $c_i : S_i \times A_i \to \mathbb{R}$ is a convex increasing function of its action a_i. An example operational cost function of an aggregator can be

$$c_i(s_i, a_i) = m_i(e_i, a_i - d_i),$$

where $m_i(e_i, a_i)$ is the maintenance cost of the energy storage that is convex [70]. It may depend on both the amount of energy in the storage and the charging/discharging rate $a_i - d_i$. Then we write each aggregator i's total cost, which is the cost aggregator i aims to minimize, as the sum of the operational cost and the payment, namely $\bar{c}_i = c_i + y_i(s_0, a_i, \boldsymbol{a}_{-i}) \cdot a_i$. Note that each aggregator's payments depends on the others' actions through the price. Although each aggregator i observes its realized price y_i, it does not know how its action a_i influences the price y_i, because the price depends on the others' actions $\boldsymbol{a}_{-i}$ and the ISO's state s_0, neither of which is known to aggregagtor i.

The energy generation cost of generator g is denoted $c_g(\varepsilon_g, a_{0,g})$, which is assumed to be convex increasing in the energy production level $a_{0,g}$. An example cost function can be

$$c_g(\varepsilon_g, a_{0,g}) = (q_{0,g} + q_{1,g} \cdot a_{0,g} + q_{2,g} \cdot a_{0,g}^2) + q_{r,g} \cdot (a_{0,g} - a_{0,g}^-)^2,$$

[4] We do not model the pricing as the ISO's action, because it does not affect the ISO's payoff, i.e. the social welfare (this is because the payment from the aggregators to the ISO is a monetary transfer within the system and does not count in the social welfare).

[5] Since we consider the interaction among the ISO and the aggregators only, we neglect the payments from the ISO to the generators, which are not included in the total cost anyway, because the payments are transferred among the entities in the system.

where $a_{0,g}^{-}$ is the production level in the previous time slot. In this case, the energy generation cost function of generator g is a vector $\varepsilon_g = (q_{0,g}, q_{1,g}, q_{2,g}, q_{r,g}, a_{0,g}^{-})$. In the cost function, $q_{0,g} + q_{1,g} \cdot a_{0,g} + q_{2,g} \cdot a_{0,g}^2$ is the quadratic cost of producing a_0 amount of energy [65][66], and $q_{r,g} \cdot (a_{0,g} - s_{0,g})^2$ is the ramping cost of changing the energy production level. We denote the total generation cost by $c_0 = \sum_{g=1}^{G} c_g$. The ISO's cost, denoted $\bar{c}_0$, is then the sum of generation costs and the aggregators' costs, i.e. $\bar{c}_0 = \sum_{i=0}^{N} c_i$.

Note, importantly, that the example cost functions above are for illustrative purpose; we can define a variety of cost functions as long as they satisfy the convexity assumption.

We assume that each entity's state transition is Markovian[6], namely its current state depends only on its previous state and its previous action. Under the Markovian assumption, we denote the transition probability of entity i's state s_i by $\rho_i(s_i'|s_i, a_i)$. This assumption holds for the following reasons. The ISO's state consists of the energy generation cost functions and the status of the transmission lines. For renewable energy generation, the energy generation cost function is modeled by the amount of available renewable energy sources (e.g. the wind speed in wind energy, and the amount of sunshine in solar energy), which is usually assumed to be i.i.d. [67][71][72]. In our model, we relax the i.i.d. assumption and allow the amount of available renewable energy sources to be correlated across adjacent periods. For conventional energy generation, the energy generation cost function is usually constant when we do not consider ramping costs. If we consider ramping costs, we can include the energy production level at the previous period in the energy generation cost function. For the aggregators, the amount of energy left in the storage depends only on the amount of energy in the previous period and the amount of energy purchases in the current period. The demand

[6] We need this assumption for our theoretical results. As we will show in the simulations, even when the state transition is not Markovian, our proposed solution can track the nonstationary dynamics (i.e. time-varying state transitions).

of the aggregator is the total demand of all its customers. Since the number of customers is large, the temporal correlation of each customer's energy demand can be neglected in the total demand. For this reason, the demand of the aggregator is often assumed to be i.i.d. [75]–[77]. In our model, we relax the i.i.d. assumption and allow the demand of the aggregator to be temporally correlated across adjacent periods.

We also assume that conditioned on the ISO's action a_0 and the aggregators' action profile $\boldsymbol{a}$, each entity's state transition is independent of each other. This assumption holds for the ISO, because the energy generation cost functions and the status of the transmission lines depend on the environments such as weather conditions, and possibly on the previous energy production levels when we consider ramping costs, but not on the aggregators' demand or its energy storage. For each aggregator, its energy storage level depends only on its own state and action, but not on the ISO's or the other aggregators' states. The demand of each aggregator could potentially depend on the ISO's state, because the ISO's state influences the unit price of energy. However, in practice, consumers are not price-anticipating (namely they do not determine how much to consume based on their anticipation of the real-time prices). As a result, it is reasonable to assume that the demand of each aggregator is independent of the ISO's and the other aggregators' states.

3.3.2 The DSM Strategy

At the beginning of each period t, each aggregator i chooses an action based on all the information it has, namely the history of its private states and the history of its prices. We write each aggregator i's history in period t as $h_i^t = (s_i^0, y_i^0; s_i^1, y_i^1; \ldots; s_i^{t-1}, y_i^{t-1}; s_i^t)$, and the set of all possible histories of aggregator i in period t as $\mathcal{H}_i^t = S_i^{t+1} \times Y_i^t$. Hence, each aggregator i's strategy can be written as $\pi_i : \cup_{t=0}^{\infty} \mathcal{H}_i^t \to A_i$. Similarly, we write the ISO's history in period t as $h_0^t = (s_0^0, \boldsymbol{y}^0; s_0^1, \boldsymbol{y}^1; \ldots; s_0^{t-1}, \boldsymbol{y}^{t-1}; s_0^t)$, where $\boldsymbol{y}^t$ is the collection of prices at period t, and the set of all possible histories of the ISO in period t as $\mathcal{H}_0^t = S_0^{t+1} \times \prod_{i \in \mathcal{N}} Y_i^t$. Then the ISO's strategy can be written as $\pi_0 : \cup_{t=0}^{\infty} \mathcal{H}_i^t \to A_i$. The joint strategy profile of all the entities is written as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. Since each entity's strategy depends only on its local information, the strategy $\boldsymbol{\pi}$ is *decentralized*[7]. Among all the decentralized strategies, we are interested in stationary decentralized strategies, in which the action to take depends only on the current information, and this dependence does not change with time. Specifically, entity i's *stationary* strategy is a mapping from its set of states to its set of actions, namely $\pi_i^s : S_i \to A_i$. Since we focus on stationary strategies, we drop the superscript s, and write π_i as entity i's stationary strategy.

The joint strategy profile $\boldsymbol{\pi}$ and the initial state $(s_0^0, s_1^0, \ldots, s_N^0)$ induce a probability distribution over the sequences of states and prices, and hence a probability distribution over the sequences of total costs $\bar{c}_i^0, \bar{c}_i^1, \ldots$. Taking expectation with respect to the sequences of stage-game payoffs, we have entity i's expected long-term cost given the initial state as

$$\bar{C}_i(\boldsymbol{\pi}|(s_0^0, s_1^0, \ldots, s_I^0)) = \mathbb{E}\left\{(1-\delta)\sum_{t=0}^{\infty}\left(\delta^t \cdot \bar{c}_i^t\right)\right\}, \tag{3.1}$$

where $\delta \in [0, 1)$ is the discount factor.

3.4 The Design Problem

The designer, namely the ISO, aims to maximize the social welfare, namely minimize the long-term total cost in the system. In addition, we need to satisfy the constraints due to the capacity of the transmission lines, the supply-demand requirements, and so on. We denote the constraints by $\boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \leq \boldsymbol{0}$, where $\boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \in \mathbb{R}^N$ with N being the number of constraints. We assume that the

[7] As we will later, the proposed strategy requires the ISO and the aggregators to exchange some information (i.e. the conjectured prices). As in the works based on network utility maximization [66], such strategies are called decentralized because the entities make decisions based on local information.

electricity flow can be approximated by the direct current (DC) flow model, in which case the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \leq \mathbf{0}$ are linear in each a_i. Hence, the design problem can be formulated as

$$\begin{aligned} \min\nolimits_{\boldsymbol{\pi}} \quad & \sum_{s_0^0, s_1^0, \ldots, s_I^0} \left\{ C_0(\boldsymbol{\pi}|(s_0^0, s_1^0, \ldots, s_I^0)) + \sum_{i \in \mathcal{I}} C_i(\boldsymbol{\pi}|(s_0^0, s_1^0, \ldots, s_I^0)) \right\} \quad (3.2) \\ s.t. \quad & \boldsymbol{f}(s_0, \pi_0(s_0), \pi_1(s_1), \ldots, \pi_I(s_I)) \leq \mathbf{0}, \ \forall (s_0, s_1, \ldots, s_N). \end{aligned}$$

Note that in the above optimization problem, we use aggregator i's cost C_i instead of its total cost $\bar{C}_i$, because its payment is transferred to the ISO and is thus canceled in the total cost. Note also that we sum up the social welfare under all the initial states. This can be considered as the expected social welfare when the initial state is uniformly distributed. The optimal stationary strategy profile that maximizes this expected social welfare will also maximize the social welfare given any initial state. We write the solution to the design problem as $\boldsymbol{\pi}^\star$ and the optimal value of the design problem as $C^\star$.

3.5 Optimal Foresighted Demand Side Management

In this section, we derive the optimal foresighted DSM strategy assuming that each entity knows its own state transition probabilities.

3.5.1 The aggregator's Decision Problem and Its Conjectured Price

Contrary to the designer, each aggregator aims to minimize its own long-term total cost $\bar{C}_i(\boldsymbol{\pi}|(s_0^0, s_1^0, \ldots, s_N^0))$. In other words, each aggregator i solves the following problem:

$$\pi_i = \arg\max_{\pi_i'} \bar{C}_i(\pi_i', \boldsymbol{\pi}_{-i}|(s_0^0, s_1^0, \ldots, s_N^0)).$$

Assuming that the aggregator knows all the information, the optimal solution to the above problem should satisfy the following:

$$V(s_0, s_i, \boldsymbol{s}_{-i}) =$$

$$\max_{a_i \in A_i} (1-\delta)\bar{c}_i(s_0, s_i, a_i, \boldsymbol{a}_{-i}) + \delta \cdot \sum_{s_0', s_i', \boldsymbol{s}_{-i}'} \left\{ \rho_0(s_0'|s_0) \prod_{j \in \mathcal{N}} \rho_j(s_j'|s_j, a_j) V(s_0', s_i', \boldsymbol{s}_{-i}') \right\}.$$

Note that the above equations would be the Bellman equations, if the aggregator knew all the information such as the other aggregators' strategies $\boldsymbol{\pi}_{-i}$ and states $\boldsymbol{s}_{-i}$, and the ISO's state s_0. However, such information is never known to the aggregator. Hence, we need to separate the influence of the other entities from each aggregator's decision problem.

One way to decouple the interaction among the aggregators is to endow each aggregator with a conjectured price. In general, the conjecture informs the aggregator of what price it should anticipate given its state and its action. However, in the presence of decentralized information, such a complicated conjecture is hard, if not impossible, to form. Specifically, aggregator i's conjectured price should depend not only on aggregator i's action and state, but also on the ISO's state. Hence, no entity possess *all the necessary information* to form the conjecture. For this reason, in this work, we adopt a simple conjecture, namely the price does not depend on the aggregator's state and action[8]. In this case, the conjectures can be formed by the ISO based on its local information and then communicated to the aggregators. Denote the conjectured price as $\tilde{y}_i$, we can rewrite aggregator i's decision problem as

$$\tilde{V}^{\tilde{y}_i}(s_i) = \max_{a_i \in A_i} (1-\delta)\left[c_i(s_i, a_i) + \tilde{y}_i \cdot a_i\right] + \delta \cdot \sum_{s_i'} \left[\rho_i(s_i'|s_i, a_i) \tilde{V}^{\tilde{y}_i}(s_i') \right].$$

Clearly, we can see from the above equations that given the conjectured price $\tilde{y}_i$, each aggregator can make decisions based only on its local information.

[8] We could propose more complicated conjectures which may depend on the aggregators' states and actions. However, as we will prove later, this simple conjecture is sufficient to achieve the social optimum.

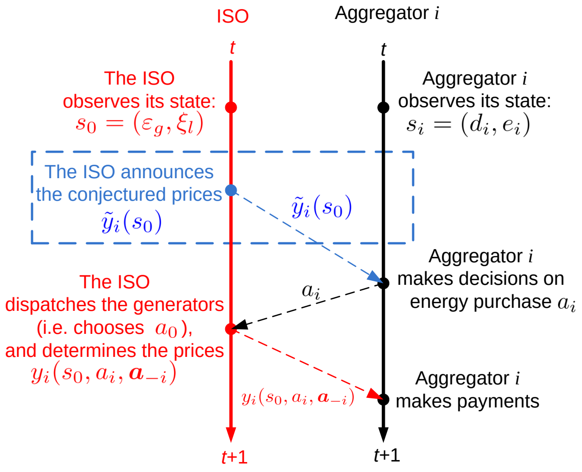

ISO
Aggregator i
t
t
The ISO observes its state: $s_0 = (\varepsilon_g, \xi_l)$
Aggregator i observes its state: $s_i = (d_i, e_i)$
The ISO announces the conjectured prices $\tilde{y}_i(s_0)$
$\tilde{y}_i(s_0)$
Aggregator i makes decisions on energy purchase a_i
a_i
The ISO dispatches the generators (i.e. chooses a_0), and determines the prices $y_i(s_0, a_i, \boldsymbol{a}_{-i})$
$y_i(s_0, a_i, \boldsymbol{a}_{-i})$
Aggregator i makes payments
t+1
t+1

Figure 3.3: Illustration of the entities' decision making and information exchange in the design framework based on conjectured prices.

In Fig. 3.3, we illustrate the entities' decision making and information exchange in the design framework based on conjectured prices. Comparing Fig. 3.3 with Fig. 3.2 of the system without conjectured prices, we can see that in the proposed design framework, the ISO sends the conjectured prices to the aggregators before the aggregators make decisions. This additional procedure of exchanging conjectured prices allows the ISO to lead the aggregators to the optimal DSM strategies. Note that the conjectured price is generally not equal to the real price charged at the end of the period, and is not equal to the expectation of the real price in the future. In this sense, the conjectured prices can be considered as control signals sent from the ISO to the aggregators, which can help the aggregators to compute the optimal strategies. In Section 4.6, we will compare the conjectured price with the expected real price by simulation.

The remaining question is how to determine the optimal conjectured prices, such that when each aggregator reacts based on its conjectured price, the resulting strategy profile maximizes the social welfare.

3.5.2 The Optimal Decentralized DSM Strategy

The optimal conjectured prices depend on the ISO's state, which is known to the ISO only. Hence, we propose a distributed algorithm used by the ISO to iteratively update the conjectured prices and by the aggregators to update their optimal strategies. The algorithm will converge to the optimal conjectured prices and the optimal strategy profile that achieves the minimum total system cost $C^\star$.

At each iteration k, given the conjectured price $\tilde{y}_i^{(k)}$, each aggregator i solves

$$\tilde{V}_i^{\tilde{y}_i^{(k)}}(s_i) = \max_{a_i \in A_i}(1-\delta)\left[c_i(s_i,a_i) + \tilde{y}_i^{(k)} \cdot a_i\right] + \delta \cdot \sum_{s_i'}\left[\rho_i(s_i'|s_i,a_i)\tilde{V}_i^{\tilde{y}_i^{(k)}}(s_i')\right],$$

and obtains the optimal value function $\tilde{V}_i^{\tilde{y}_i^{(k)}}$ as well as the corresponding optimal strategy $\pi_i^{\tilde{y}_i^{(k)}}$ under the current conjectured price $\tilde{y}_i^{(k)}$.

Similarly, given the conjectured prices $\tilde{\boldsymbol{y}}_0^{(k)} \in \mathbb{R}^G$, the ISO solves

$$\tilde{V}_0^{\tilde{\boldsymbol{y}}_0^{(k)}}(s_0) = \min_{a_i \in A_i}(1-\delta)\left[\sum_g c_g(s_0,a_0) + \tilde{\boldsymbol{y}}_0^{(k)T} \cdot a_0\right] + \delta \cdot \sum_{s_0'}\left[\rho_0(s_0'|s_0,a_0)\tilde{V}_0^{\tilde{\boldsymbol{y}}_0^{(k)}}(s_0')\right],$$

and obtains the optimal value function $\tilde{V}_0^{\tilde{y}_i^{(k)}}$ as well as the corresponding optimal strategy $\pi_0^{\tilde{y}_i^{(k)}}$ under the current conjectured price $\tilde{y}_i^{(k)}$.

Then the ISO updates the conjectured prices as follows:

$$\tilde{y}_i^{(k+1)} = \left(\boldsymbol{\lambda}^{(k+1)}(s_0)\right)^T \cdot \frac{\partial f(s_0, \boldsymbol{a})}{\partial a_i},$$

where $\boldsymbol{\lambda}^{(k+1)}(s_0) \in \mathbb{R}^N$ is calculated as

$$\boldsymbol{\lambda}^{(k+1)}(s_0) = \left\{\boldsymbol{\lambda}^{(k)}(s_0) + \Delta(k) \cdot \boldsymbol{f}\left(s_0, \pi_0^{\tilde{\boldsymbol{y}}_0(k)}(s_0), \frac{\sum_{s_1}\pi_1^{\tilde{y}_1(k)}(s_1)}{|S_1|}, \ldots, \frac{\sum_{s_I}\pi_I^{\tilde{y}_I(k)}(s_I)}{|S_I|}\right)\right\}^+,$$

where $\Delta(k) \in \mathbb{R}_{++}$ is the step size, and $\{x\}^+ = \max\{x, 0\}$.

Note that in the above update of conjectures, to calculate (the subgradient) $\boldsymbol{\lambda}^{(k)}$, the ISO needs to know the average amount of purchase $\frac{\sum_{s_i}\pi_i^{\tilde{y}_1(k)}(s_i)}{|S_i|}$ from

Table 3.4: Distributed algorithm to compute the optimal decentralized DSM strategy.

```
Input: Each entity's performance loss tolerance $\epsilon_i$

Initialization: Set $k = 0$, $\bar{a}_i(0) = 0, \forall i \in \mathcal{I}$, $\tilde{y}_i(0) = 0, \forall i \in \mathcal{I} \cup \{0\}$.
repeat
    Each aggregator $i$ solves
        $\tilde{V}_i^{\tilde{y}_i^{(k)}(s_0)}(s_i) = \max_{a_i \in A_i} (1-\delta) \left[c_i(s_i, a_i) + \tilde{y}_i^{(k)}(s_0) \cdot a_i\right] + \delta \cdot \sum_{s_i'} \left[\rho_i(s_i'|s_i, a_i) \tilde{V}_i^{\tilde{y}_i^{(k)}(s_0)}(s_i')\right]$
    The ISO solves
        $\tilde{V}_0^{\tilde{\boldsymbol{y}}_0(k)}(s_0) = \min_{a_i \in A_i} (1-\delta) \left[\sum_g c_g(s_0, a_0) + \tilde{\boldsymbol{y}}_0(k)^T \cdot a_0\right] + \delta \cdot \sum_{s_0'} \left[\rho_0(s_0'|s_0, a_0) \tilde{V}_0^{\tilde{\boldsymbol{y}}_0(k)}(s_0')\right]$
    Each aggregator $i$ reports its purchase request $\pi_i^{\tilde{y}_i^{(k)}(s_0)}(s_i)$
    The ISO updates $\bar{a}_i(k+1) = \bar{a}_i(k) + \pi_i^{\tilde{y}_i^{(k)}(s_0)}(s_i)$ for all $i \in \mathcal{I}$
    The ISO updates the conjectured prices:
        $\tilde{y}_i^{(k+1)}(s_0) = \left(\boldsymbol{\lambda}(k+1)^T \cdot \frac{\partial f(s_0, \boldsymbol{a})}{\partial a_i}\right)^T$, where $\Delta(k) = \frac{1}{k+1}$ and
            $\boldsymbol{\lambda}(k+1) = \left\{\boldsymbol{\lambda}(k) + \Delta(k) \cdot \boldsymbol{f}\left(s_0, \pi_0^{\tilde{\boldsymbol{y}}_0(k)}(s_0), \frac{\bar{a}_1(k+1)}{k+1}, \ldots, \frac{\bar{a}_I(k+1)}{k+1}\right)\right\}^+$
until $\|\tilde{V}_i^{\tilde{y}_i^{(k+1)}(s_0)} - \tilde{V}_i^{\tilde{y}_i^{(k)}(s_0)}\| \leq \epsilon_i$
```

each aggregator i. This requires additional information exchange from the aggregator to the ISO. Moreover, the aggregator may not be willing to report such information to the ISO. To reduce the amount of information exchange and preserve privacy, we propose that the ISO calculates the empirical mean values of the aggregators' purchases in the run-time (which results in stochastic subgradients). We summarize the algorithm in Table 3.4, and prove that the algorithm can achieve the optimal social welfare in the following theorem.

Theorem 5 *The algorithm in Table 3.4 converges to the optimal strategy profile, namely*

$$\lim_{k \to \infty} \left| \sum_{s_0^0, s_1^0, \ldots, s_I^0} \left\{ C_0(\boldsymbol{\pi}^{\tilde{y}^{(k)}} | (s_0^0, s_1^0, \ldots, s_I^0)) + \sum_{i \in \mathcal{I}} C_i(\boldsymbol{\pi}^{\tilde{y}^{(k)}} | (s_0^0, s_1^0, \ldots, s_I^0)) \right\} - C^\star \right| = 0.$$

Proof 2 *See the appendix.*

We summarize the information needed by each entity in Table 3.5. We can see

Table 3.5: Information needed by each entity to implement the algorithm.

Entity	Information at each step k
The ISO	The purchase request $\pi_i^{\tilde{y}_i^{(k)}(s_0)}(s_i)$ of each aggregator
Each aggregator i	Conjecture on its price $\tilde{y}_i^{(k)}(s_0)$

that the amount of information exchange at each iteration is small $(O(I))$, compared to the amount of information unavailable to each entity ($\prod_{j\neq i}|S_i|$ states plus the strategies $\boldsymbol{\pi}_{-i}$). In other words, the algorithm enables the entities to exchange a small amount $(O(I))$ of information and reach the optimal DSM strategy that achieves the same performance as when each entity knows the complete information about the system.

We briefly discuss the complexity of implementing the algorithm in terms of the dimensionality of the Bellman equations solved by each entity. For each aggregator, it solves the Bellman equation that has the the same dimensionality as the cardinality of its state space, namely $|S_i|$. For each ISO, the dimensionality of its state space is large, because the generation cost functions $\boldsymbol{\varepsilon}$ are a vector of length G and the status of the transmission lines is a vector of length L. However, the ISO's decision problem can be decomposed due to the following observation. Note that the generators' energy generation cost functions are independent of each other. Then we have the following theorem.

Theorem 6 *Given the conjectured price $\tilde{\boldsymbol{y}}_0(k)$, the ISO's value function $\tilde{V}_0^{\tilde{\boldsymbol{y}}_0(k)}$ can be calculated by $\tilde{V}_0^{\tilde{\boldsymbol{y}}_0(k)}(s_0)=\sum_{g=1}^{G}\tilde{V}_{0,g}^{\tilde{y}_{0,g}(k)}(\varepsilon_g)$, where $\tilde{V}_{0,g}^{\tilde{y}_{0,g}(k)}$ solves*

$$\tilde{V}_{0,g}^{\tilde{y}_{0,g}(k)}(\varepsilon_g)=$$

$$\min_{a_{0,g}}(1-\delta)\left[c_g(\varepsilon_g,a_{0,g})+\tilde{y}_{0,g}(k)^T\cdot a_{0,g}\right]+\delta\cdot\sum_{\varepsilon_g'}\left[\rho_0(\varepsilon_g'|\varepsilon_g,a_{0,g})\tilde{V}_{0,g}^{\tilde{y}_{0,g}(k)}(\varepsilon_g')\right].$$

Proof 3 *The proof follows directly from Lemma 2 in the appendix.*

From the above proposition, we know that the dimensionality of the ISO's decision problem is $\sum_{g=1}^{G}|\mathcal{E}_g|$, where $|\mathcal{E}_g|$ is the cardinality of the set of generator g's generation cost functions. The dimensionality increases linearly with the number of generators, instead of exponentially with the number of generators and transmission lines without decomposition.

3.5.3 Learning Unknown Dynamics

In practice, each entity may not know the dynamics of its own states (i.e., its own state transition probabilities) or even the set of its own states. When the state dynamics are not known a priori, each entity cannot solve their decision problems using the distributed algorithm in Table 3.4. In this case, we can adapt the online learning algorithm based on post-decision state (PDS) in [84], which was originally proposed for wireless video transmissions, to our case.

The main idea of the PDS-based online learning is to learn the post-decision value function, instead of the normal value function. Each aggregator i's post-decision value function is defined as $U_i(\tilde{d}_i, \tilde{e}_i)$, where $(\tilde{d}_i, \tilde{e}_i)$ is the post-decision state. The difference from the normal state is that the PDS $(\tilde{d}_i, \tilde{e}_i)$ describes the status of the system after the purchase action is made but before the demand in the next period arrives. Hence, the relationship between the PDS and the normal state is

$$\tilde{d}_i = d_i, \;\; \tilde{e}_i = e_i + (a_i - d_i).$$

Then the post-decision value function can be expressed in terms of the normal value function as follows:

$$U_i(\tilde{d}_i, \tilde{e}_i) = \sum_{d_i'} \rho_i(d_i', \tilde{e}_i - (a_i - \tilde{d}_i) | \tilde{d}_i, \tilde{e}_i) \cdot V_i(d_i', \tilde{e}_i - (a_i - \tilde{d}_i)).$$

In PDS-based online learning, the normal value function and the post-decision value function are updated in the following way:

$$\begin{aligned} V_i^{(k+1)}(d_i^{(k)}, e_i^{(k)}) &= \max_{a_i} (1-\delta) \cdot c_i(d_i^{(k)}, e_i^{(k)}, a_i) + \delta \cdot U_i^{(k)}(d_i^{(k)}, e_i^{(k)} + (a_i - d_i^{(k)})), \\ U_i^{(k+1)}(d_i^{(k)}, e_i^{(k)}) &= (1-\alpha^{(k)}) U_i^{(k)}(d_i^{(k)}, e_i^{(k)}) + \alpha^{(k)} \cdot V_i^{(k)}(d_i^{(k)}, e_i^{(k)} - (a_i - d_i^{(k)})). \end{aligned}$$

We can see that the above updates do not require any knowledge about the state dynamics. It is proved in [84] that the PDS-based online learning will converge to the optimal value function.

3.5.4 Detailed Comparisons with Existing Frameworks

Since we have introduced our proposed framework, we can provide a detailed comparison with the existing theoretical framework. The comparison is summarized in Table 3.6.

First, the proposed framework reduces to the myopic optimization framework when we set the discount factor $\delta = 0$. In this case, the problem reduces to the classic economic dispatch problem.

Second, the Lyapunov optimization framework is closely related to the PDS-based online learning. In fact, it could be considered as a special case of the PDS-based online learning when we set the post-decision value function as $U_i(s_i) = c_i(s_i, a_i) + (e_i + a_i)^2 - e_i^2$, and choose the action that minimizes the post-decision value function in the run-time. However, the Lyapunov drift in the above post-decision value function depends only on the status of the energy storage, but not on the demand. In contrast, in our PDS-based online learning, we explicitly considers the impact of the demand when updating the normal and post-decision value functions.

Finally, the key difference between our proposed framework and the framework for MU-MDP [78][79] is how we penalize the constraints $f(s_0, a_0, \boldsymbol{a})$. In particular, the framework in [78][79], if directly applied in our model, would define only one Lagrangian multiplier for all the constraints under different states s_0. This leads to performance loss in general [79]. In contrast, we define different Lagrangian multipliers to penalize the constraints under different states s_0, and potentially enable the proposed framework to achieve the optimality (which is indeed the case

Table 3.6: Relationship between the proposed and existing theoretical frameworks.

Framework	Relationship	Representative works
Myopic	$\delta = 0$	[67]
Lyapunov optimization	PDS value function $U_i(s_i) = c_i(s_i, a_i) + (e_i + a_i - d_i)^2 - e_i^2$	[75][76]
MU-MDP	Lagrangian multiplier $\boldsymbol{\lambda}(s_0) = \boldsymbol{\lambda}$ for all s_0	[78][79]

as have been proved in Theorem 5).

3.6 Simulation Results

In this section, we validate our theoretical results and compare against existing DSM strategies through extensive simulations. We use the widely-used IEEE test power systems with the data (e.g. the topology, the admittances and capacity limits of transmission lines) provided by University of Washington Power System Test Case Archive [87]. We describe the other system parameters as follows (these system parameters are used by default; any changes in certain scenarios will be specified):

- One period is one hour. The discount factor is $\delta = 0.99$.
- The demand of aggregator i at period t is uniformly distributed among the interval $[d_i(t \mod 24) - \Delta d_i(t \mod 24), d_i(t \mod 24) + \Delta d_i(t \mod 24)]$. In other words, the distribution of demand is time-varying across a day. We let the peak hours for all the aggregators to be from 17:00 to 22:00. The mean value $d_i(t \mod 24)$ and the range $\Delta d_i(t \mod 24)$ of aggregator i's demand are described as follows (values are adapted from [88]):

$$d_i(t \mod 24) = \begin{cases} 50 + (i-1) \cdot 0.5 \text{ MW} & \text{if } t \mod 24 \in [17, 22] \\ 25 + (i-1) \cdot 0.5 \text{ MW} & \text{otherwise} \end{cases} \tag{3.3}$$

and

$$\Delta d_i(t \mod 24) = \begin{cases} 5 \text{ MW} & \text{if } t \mod 24 \in [17, 22] \\ 2 \text{ MW} & \text{otherwise} \end{cases} \tag{3.4}$$

- All the aggregators have energy storage of the same capacity 25 MW.
- All the aggregators have the same linear energy storage cost function [70]:

  $$c_i(s_i, a_i) = 2 \cdot (a_i - d_i)^+,$$

  namely the maintenance cost grows linearly with the charging/discharging rates.
- We index the energy generators starting from the renewable energy generators. All the renewable energy generators have linear energy generation cost functions: [88]

  $$c_g(a_{0,g}) = g \cdot a_{0,g},$$

  where the unit energy generation cost has the same value as the index of the generator (these values are adapted from [88], which cited that the unit energy generation cost ranges from \$0.19/MWh to \$10/MWh). Although the energy generation cost function is deterministic, the maximum amount of energy production is stochastic (due to wind speed, the amount of sunshine, and so on). The maximum amounts of energy production of all the renewable energy generators follow the same uniform distribution in the range of $[90, 110]$ MW.
- The rest of energy generators are conventional energy generators that use coal, all of which have the same energy generation cost function: [70]

  $$c_g(a_{0,g}) = \underbrace{\frac{1}{2}(a_{0,g})^2}_{\text{generation cost}} + \underbrace{\frac{1}{10}(a_{0,g} - a_{0,g}^-)^2}_{\text{ramping cost}}.$$

  In other words, the conventional energy generators have fixed (i.e. not stochastic) generation cost functions.

- The status of the transmission lines is their capacity limits. The nominal values of the capacity limits are the same as specified in the data provided by [87]. In each period, we randomly select a line with equal probability, and decrease its capacity limit by 10%.

We compare the proposed DSM strategies with the following schemes.

- Centralized optimal strategies ("Centralized"): We assume that there is a central controller who knows everything about the system and solves the long-term cost minimization problem as a single-user MDP. This scheme serves as the benchmark optimum.

- Myopic strategies ("Myopic") [65]–[72]: In each period t, the aggregators myopically minimizes their current costs, and based on their actions, the ISO minimizes the current total generation cost.

- Single-user Lyapunov optimization ("Lyapunov") [73]–[76]: We let each aggregator adopt the stochastic optimization technique proposed in [73]–[76]. Based on the aggregators' purchases, the ISO minimizes the current total generation cost.

3.6.1 Learning and Tracking The Optimal Policies Without Knowledge of State Transition Probabilities

Before comparing against the other solutions, we show that the proposed PDS-based learning algorithm converges to the optimal solution (namely the optimal value function is learned). The optimal solution is obtained by the proposed algorithm in Table 3.4 assuming the statistical knowledge of the system dynamics. We consider the IEEE 14-bus system, in which each aggregator has a energy storage of 45 MW. For illustrative purpose, we show the convergence of the learning algorithm in terms of the average long-term costs only for two aggregators in

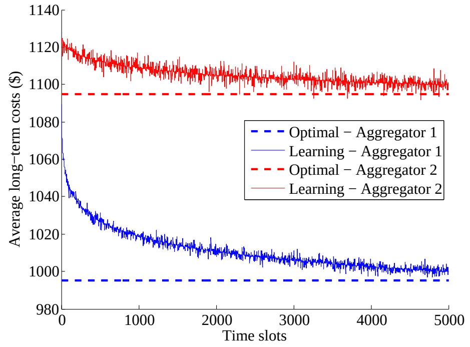


Figure 3.4: Convergence of the PDS-based learning algorithm. The two aggregators' average long-term costs converge to the optimal one under the PDS-based learning algorithm.

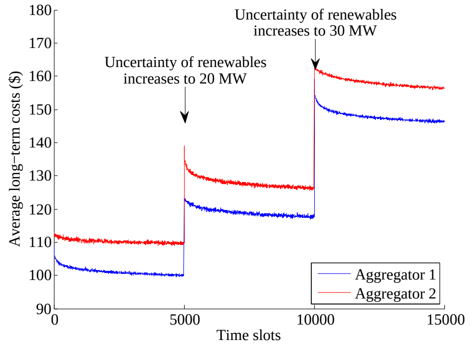


Figure 3.5: The PDS-based learning algorithm tracks the optimal solution when the underlying distribution of the energy generation is time-varying.

Fig. 3.4.

We also demonstrate that the proposed PDS-based learning lagorithm can track the optimal solution when the state transition probabilities are time-varying. Note that we assume that the maximum amounts of energy production by the renewable energy generators are uniformly distributed in the range of [90, 110] mW. In Fig. 3.5, we change the range of the uniform distribution to 20 mW and 30 mW (i.e. increase the uncertainty of renewable energy) every 5000 time slots. We can see that the learning algorithm can track the optimal solution even when the underlying distribution of the energy generation is time-varying.

3.6.2 Performance Evaluation

Now we evaluate the performance of the proposed DSM strategy in various scenarios.

3.6.2.1 Impact of the energy storage

First, we study the impact of the energy storage on the performance of different schemes. We assume that all the generators are conventional energy generators using fossil fuel, in order to rule out the impact of the uncertainty in renewable energy generation (which will be examined next). The performance criterion is the total cost per hour normalized by the number of buses in the system. We compare the normalized total cost achieved by different schemes when the capacity of the energy storage increases from 5 MW to 45 MW.

Fig. 3.6–3.8 show the normalized total cost achieved by different schemes under IEEE 14-bus system, IEEE 30-bus system, and IEEE 118-bus system, respectively. Note that we do not show the performance of the centralized optimal strategy under IEEE 118-bus system, because the number of states in the centralized MDP is so large that it is intractable to compute the optimal solution. This also shows the computational tractability and the scalability of the proposed distributed algorithm. Under IEEE 14-bus and 30-bus systems, we can see that the proposed DSM strategy achieves almost the same performance as the centralized optimal strategy. The slight optimality gap comes from the performance loss experienced during the convergence process of the conjectured prices. Compared to the DSM strategy based on single-user Lyapunov optimization, our proposed strategy can reduce the total cost by around 30% in most cases. Compared to the myopic DSM strategy, our reduction in the total cost is even larger and increases with the capacity of the energy storage (up to 60%).

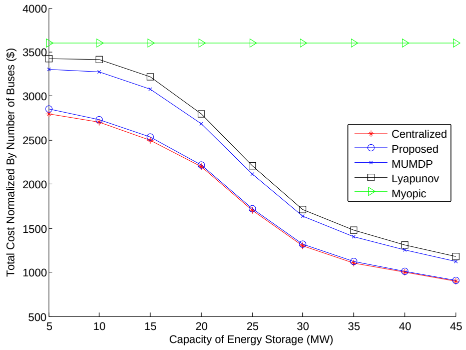


Figure 3.6: The normalized total cost per hour versus the capacity of the energy storage in the IEEE 14-bus system.

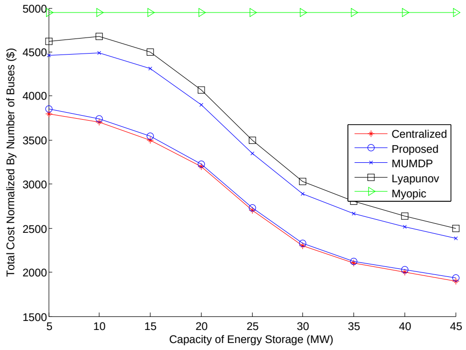


Figure 3.7: The normalized total cost per hour versus the capacity of the energy storage in the IEEE 30-bus system.

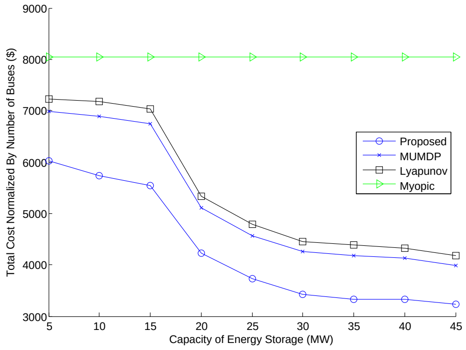


Figure 3.8: The normalized total cost per hour versus the capacity of the energy storage in the IEEE 118-bus system.

3.6.2.2 Impact of the uncertainty in renewable energy generation

Now we examine the impact of the uncertainty in renewable energy generation. For a given test system, we let half of the generators to be renewable energy generators. Recall that the maximum amounts of energy production of the renewable energy generators are stochastic and follow the same uniform distribution. We set the mean value of the maximum amount of energy production to be 100 MW, and vary the range of the uniform distribution. A wider range indicates a higher uncertainty in renewable energy production. Hence, we define the uncertainty in renewable energy generation as the maximum deviation from the mean value in the uniform distribution.

Fig. 3.9 shows the normalized total cost under different degrees of uncertainty in renewable energy generation. Again, the proposed strategy achieves the performance of the centralized optimal strategy in the IEEE 14-bus system. We can see that the costs achieved by all the schemes increase with the uncertainty in renewable energy generation. This happens for the following reasons. Since the renewable energy is cheaper, the ISO will dispatch renewable energy whenever possible, and dispatch conventional energy for the residual demand. However, when the renewable energy generation has larger uncertainty, the variation in the residual demand is higher, which results in a higher variation in the conventional energy dispatched and thus a larger ramping cost. To reduce the ramping cost, the ISO needs to be more conservative in dispatching the renewable energy, which results in a higher total cost. However, we can also see from the simulation that when the aggregators have larger capacity to store energy, the increase of the total cost with the uncertainty is smaller. This is because the energy storage can smooth the demand, in order to mitigate the impact of uncertainty in the renewable energy generation. This shows the value of energy storage to reduce the cost.

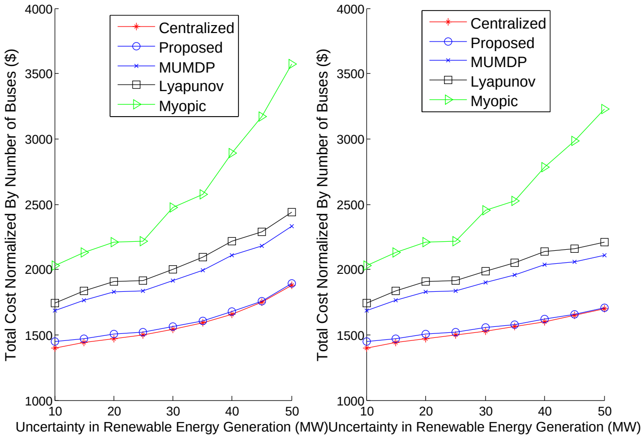


Figure 3.9: The normalized total cost per hour versus the uncertainty in renewable energy generation in the IEEE 14-bus system. The aggregators have energy storage of capacity 25 MW and 50 MW, respectively.

3.6.2.3 Fairness

Now we investigate how the individual costs of the aggregators are influenced by the capacity of their energy storage. In particular, we are interested in whether some aggregators are affected by having smaller energy storage. We assume that half of the aggregators have energy storage of capacity 50 MW, while the other half have energy storage of much smaller capacity 10 MW. In Fig. 3.10, we compare the average individual cost of the aggregators with smaller energy storage and that of the aggregators with larger energy storage. We can see that the average cost of the aggregators with smaller energy storage does increase with the uncertainty in renewable energy generation. Hence, the aggregators with higher energy storage have an advantage over those with smaller energy storage, because they have high flexibility in coping with the price fluctuation.

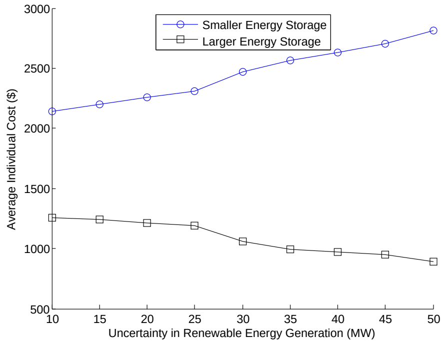


Figure 3.10: The average individual costs of the aggregators with different energy storage in the IEEE 118-bus system.

3.6.3 The Conjectured Prices

We compare the conjectured prices with the expected real prices. In our simulation, each aggregator i's conjectured price is the conjectured price that the proposed algorithm in Table 3.4 converges to, namely $\lim_{k\rightarrow\infty} \tilde{y}_i^{(k)}$. We can also calculate the expected real price as follows. The optimal DSM strategy that the algorithm in Table 3.4 converges to will induce a probability distribution over the states. In each state, we calculate the locational marginal price for each aggregator based on the actions taken at this state. Then we calculate the expected LMP of each aggregator, which is the expected real price at which each aggregator pays for the energy.

In Fig. 3.11, we show the highest conjectured price among all the aggregators and the lowest expected real price among all the aggregators, because it is infeasible to plot the prices for all the aggregators in the figure. Hence, the difference of the conjectured price and the real expected price for each individual aggregator is no larger than the difference shown in the figure. First, as we can see from the figure, the prices go down when the capacity of the energy storage increases. This is because with energy storage, the congestion due to high energy purchase demand decreases, which in turn decreases the congestion cost (technically, the

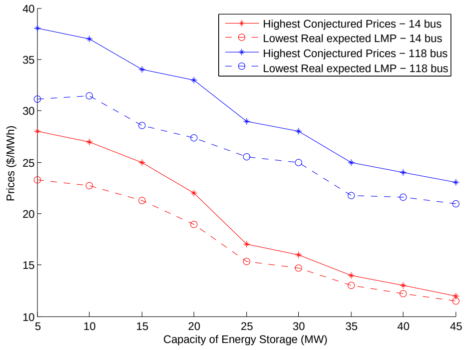


Figure 3.11: The conjectured prices and the real expected locational marginal prices (LMPs) in IEEE 14-bus and 118-bus systems.

Lagrangian multiplier associated with the capacity constraint is smaller). Second, in our simulations (not shown in the figure), we observe that the conjectured prices are always higher than the expected real prices. Hence, the conjectured price gives each aggregator an overestimate of the real expected price. An overestimate is better than an underestimate, in the sense that each aggregator has a guarantee of how much it will pay in the worst case. Finally, we can see from the figure that the conjectured price is close to the real expected price. The maximum difference between these two prices is less than 20% in all the considered scenarios.

3.6.4 Comparisons of the DSM strategies

In this section, to better understand why the proposed DSM strategies outperform the other strategies, we present a simple example and illustrate why the proposed strategy achieves a lower cost. To keep the illustration simple, we reduce the number of states. Specifically, we assume that the demand has three states: “high”, “medium”, “low”, which corresponds to the highest, medium, and lowest values of the uniform distribution described in (3.3)(3.4). Similarly, the energy storage has three states: “empty”, “half”, and “full”. The maximum capacity of the renewable energy generator has two values, corresponding to the highest and lowest values in the uniform distribution described in the basic simulation setup at the beginning of Section 4.6. To distinguish the description of the demand state, we suppose that the renewable energy generator harnesses solar energy, and refer to its states as “sunny” and “cloudy” instead of “high” and “low”. We do not assume any randomness in the transmission lines. The purchase from each aggregator is also quantized into three levels: large, moderate, and small.

In Table 3.7, we compare the actions chosen by different strategies under different states in an IEEE 14-bus system. Due to space limitation, we cannot show the strategies of all the aggregators, but only one of them. The state is a three tuple that consists of the demand, the energy storage, and the renewable energy generation capacity. Although we have reduced the number of states, there are still $3 \times 3 \times 2 = 18$ states, which is hard to show in one table. Instead, we only show the actions in some representative states, in which different strategies take very different actions.

First, we can observe that the myopic strategy takes actions based on the demand and the energy storage exclusively. The myopic strategy aims to minimize the current operational cost of the energy storage as long as the demand can be satisfied. Hence, it chooses to purchase small amount of energy as long as the demand can be fulfilled by the energy left in the storage. Second, as we discussed at the end of Section 3.5, the strategy based on Lyapunov optimization does not take into account the demand dynamics. As we can see from the table, the strategy based on Lyapunov optimization takes actions based on the energy storage and the renewable energy generation capacity exclusively. It will purchase large amount of energy as long as it is sunny (which means that the capacity of the renewable energy generator is high and hence the price is low). In contrast, the proposed strategy considers all the three states when making decisions. For example, when the states are (low,empty,sunny) and (high,empty,sunny), the strategy based on Lyapunov optimization always chooses to purchase large amount of energy, while the proposed strategy will purchase moderate amount of energy when the demand

Table 3.7: Comparisons of different strategies.

State	high, full, sunny	high, full, cloudy	low, full, sunny	low, empty, sunny	high, empty, cloudy	high, empty, sunny
Myopic	small	small	small	small	large	large
Lyapunov	large	small	large	large	small	large
MU-MDP	large	small	small	small	small	large
Proposed	large	low	moderate	moderate	low	large

is low. Finally, the strategy based on MU-MDP also considers all the three states, and takes similar actions as the proposed strategy. However, the strategy based on MU-MDP takes more conservative actions (e.g. purchases small amount of energy when the proposed strategy purchases moderate amount of energy). This is because there is only one Lagrangian multiplier under all the states, and to ensure the feasibility of the constraints, the Lagrangian multiplier has to be set larger. This results in a harsher penalty in the objective function. Hence, the actions taken are more conservative to ensure that the line capacity constraints are satisfied.

3.7 Conclusion

In this chapter, we proposed a methodology for performing optimal foresighted DSM strategies that minimize the long-term total cost of the power system. We overcame the hurdles of information decentralization and complicated coupling among the various entities present in the system, by decoupling their decision problems using conjectured prices. We proposed an online algorithm for the ISO to update the conjectured prices, such that the conjectured prices can converge to the optimal ones, based on which the entities make optimal decisions that minimize the long-term total cost. We prove that the proposed method can achieve the social optimum, and demonstrate through simulations that the proposed foresighted DSM significantly reduces the total cost compared to the optimal myopic DSM (up to 60% reduction), and the foresighted DSM based on the Lyapunov optimization framework (up to 30% reduction).

3.8 Appendix

Due to limited space, we only give a detailed proof sketch. The proof consists of three key steps. First, we prove that by penalizing the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$ into the objective function, the decision problems of different entities can be decentralized. Hence, we can derive optimal decentralized strategies for different entities under given Lagrangian multipliers. Then we prove that the update of Lagrangian multipliers converges to the optimal ones under which there is no duality gap between the primal problem and the dual problem, due to the convexity assumptions made on the cost functions. Finally, we validate the calculation of the conjectured prices.

First, suppose that there is a central controller that knows everything about the system. Then the optimal strategy to the design problem (4.6) should result in a value function V^* that satisfies the following Bellman equation: for all $s_0, s_1, \ldots, s_I$, we have

$$\begin{aligned} V^*(s_0, \boldsymbol{s}) = \quad & \max_{a_0, \boldsymbol{a}} \quad \left\{ (1-\delta) \cdot \sum_{i=0}^{I} c_i(s_i, a_i) + \delta \cdot \sum_{s_0', \boldsymbol{s}'} \rho(s_0', \boldsymbol{s}' | s_0, \boldsymbol{s}, a_0, \boldsymbol{a}) V^*(s_0', \boldsymbol{s}') \right\} \\ & s.t. \quad \boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \leq 0. \end{aligned} \tag{3.5}$$

Defining a Lagrangian multiplier $\boldsymbol{\lambda}(s_0) \in \mathbb{R}_+^N$ associated with the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \leq 0$, and penalizing the constraints on the objective function, we get the following Bellman equation:

$$V^{\boldsymbol{\lambda}}(s_0, \boldsymbol{s}) = \max_{a_0, \boldsymbol{a}} \left\{ (1-\delta) \cdot \left[\sum_{i=0}^{I} c_i(s_i, a_i) + \boldsymbol{\lambda}^T(s_0) \cdot \boldsymbol{f}(s_0, a_0, \boldsymbol{a}) \right] \right. \tag{3.6}$$

$$\left. + \delta \cdot \sum_{s_0', \boldsymbol{s}'} \rho(s_0', \boldsymbol{s}' | s_0, \boldsymbol{s}, a_0, \boldsymbol{a}) V^{\boldsymbol{\lambda}}(s_0', \boldsymbol{s}') \right\}.$$

In the following lemma, we can prove that (3.6) can be decomposed.

Lemma 2 *The optimal value function $V^{\boldsymbol{\lambda}}$ that solves* (3.6) *can be decomposed as $V^{\boldsymbol{\lambda}}(s_0, \boldsymbol{s}) = \sum_{i=0}^{I} V_i^{\boldsymbol{\lambda}}(s_i)$ for all $(s_0, \boldsymbol{s})$, where $V_i^{\boldsymbol{\lambda}}$ can be computed by entity i locally by solving*

$$V_i^{\boldsymbol{\lambda}(s_0)}(s_i) = \max_{a_i} \left\{ (1-\delta) \cdot \left[c_i(s_i, a_i) + \boldsymbol{\lambda}^T(s_0) \cdot f_i(s_0, a_i) \right] + \delta \cdot \sum_{s_i'} \rho_i(s_i'|s_i, a_i) V_i^{\boldsymbol{\lambda}}(s_i') \right\} \tag{3.7}$$

Proof 4 *This can be proved by the independence of different entities' states and by the decomposition of the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$. Specifically, in a DC power flow model, the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$ are linear with respect to the actions $a_0, a_1, \ldots, a_I$. As a result, we can decompose the constraints as $\boldsymbol{f}(s_0, a_0, \boldsymbol{a}) = \sum_{i=0}^{I} f_i(s_0, a_i)$.*

We have proved that by penalizing the constraints $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$ using Lagrangian multiplier $\boldsymbol{\lambda}(s_0)$, different entities can compute the optimal value function $V_i^{\boldsymbol{\lambda}(s_0)}$ distributively. Due to the convexity assumptions on the cost functions, we can show that the primal problem (4.6) is convex. In addition, there always exists a strictly feasible solution. Hence, there is no duality gap. In other words, at the optimal Lagrangian multipliers $\boldsymbol{\lambda}^*(s_0)$, the corresponding value function $V^{\boldsymbol{\lambda}^*(s_0)}(s_0, \boldsymbol{s}) = \sum_{i=0}^{I} V_i^{\boldsymbol{\lambda}^*(s_0)}(s_i)$ is equal to the optimal value function $V^*(s_0, \boldsymbol{s})$ of the primal problem (3.5). It is left to show that the update of Lagrangian multipliers converge to the optimal ones. It is a well-known result in dynamic programming that $V^{\boldsymbol{\lambda}(s_0)}$ is convex and piecewise linear in $\boldsymbol{\lambda}(s_0)$, and that the subgradient of $V^{\boldsymbol{\lambda}(s_0)}$ with respect to $\boldsymbol{\lambda}(s_0)$ is $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$ (it is a subgradient since the function $V^{\boldsymbol{\lambda}}(s_0)$ may not be differentiable with respect to $\boldsymbol{\lambda}(s_0)$) [79]. Note that we use the sample mean of a_0 and $\boldsymbol{a}$, whose expectation is the true mean value of a_0 and $\boldsymbol{a}$. Since $\boldsymbol{f}(s_0, a_0, \boldsymbol{a})$ is linear in a_0 and $\boldsymbol{a}$, the subgradient calculated based on the sample mean has the same mean value as the subgradient calculated based on the true mean values. In other words, the update is a stochastic subgradient descent method. It is well-known that when the stepsize $\Delta(k) = \frac{1}{k+1}$, the stochastic subgradient descent on the dual variable (i.e. the conjectured prices) $\boldsymbol{\lambda}$ will converge to the optimal $\boldsymbol{\lambda}^*$ [91].

Finally, we can write the conjectured prices by taking the derivatives of the penalty terms. For aggregator i, its penalty is $\boldsymbol{\lambda}^T(s_0) \cdot f_i(s_0, a_i)$. Hence, its conjectured price is

$$\frac{\partial \boldsymbol{\lambda}^T(s_0) \cdot f_i(s_0, a_i)}{\partial a_i} = \boldsymbol{\lambda}^T(s_0) \cdot \frac{\partial f_i(s_0, a_i)}{\partial a_i} \ . \tag{3.8}$$

CHAPTER 4

Resource Exchange With Imperfect Monitoring

4.1 Introduction

Service exchange platforms have proliferated as the medium that allows the users to exchange services valuable to each other. For instance, emerging new service exchange platforms include crowdsourcing systems (e.g. in Amazon Mechanical Turk and CrowdSource) in which the users exchange labor [92][93], online question-and-answer websites (e.g. in Yahoo! Answers and Quora) in which the users exchange knowledge [93], peer-to-peer (P2P) networks in which the users exchange files/packets [94][95][96], and online trading platforms (e.g. eBay) where the users exchange goods [97]. In a typical service exchange platform, a user plays a dual role: as a *client*, who requests services, and as a *server*, who chooses to provide high-quality or low-quality services. Common features of many service exchange platforms are: the user population is large and users are anonymous. In other words, each user interacts with a randomly-matched partner without knowing its partner's identity (However, the platform does know the identify of the interacting users.). The absence of a fixed partner and the anonymity of the users create incentive problems – namely the users tend to "free-ride" (i.e., receive high-quality services from others as a client, while providing low-quality services as a server). In addition, a user generally may not be able to perfectly monitor[1] its partner's action, which makes it even harder to incentivize the users to provide

[1] The monitoring discussed throughout this paper is a user's observation on its current partner's actions. Each user knows nothing about the ongoing interactions among the other pairs of users.

high-quality services.

An important class of incentive mechanisms for service exchange platforms are the rating mechanisms[2] [93]–[104], in which each user is labeled with a rating based on its past behaviors in the system. A rating mechanism consists of a rating update rule and a recommended strategy[3]. The recommended strategy specifies what is the desirable behavior under the current system state (e.g. the current rating profile of the users or the current rating distribution). For example, the rating mechanism may recommend providing high-quality services for all the users when the majority of users have high ratings, while recommending to provide high-quality services only to high-rating users when the majority have low ratings. Then, based on each client's report on the quality of service, the rating mechanism revises each server's rating according to the rating update rule. Generally speaking, the ratings of the users who comply with (resp. deviate from) the recommended behaviors go up (resp. down). Hence, each user's rating summarizes its past behavior in the system. By keeping track of all the users' ratings and recommending them to reward (resp. punish) the users with high (resp. low) ratings, the rating mechanism gives incentives to the users to obtain high ratings by rewarding them indirectly, through recommending other users to provide them with high-quality services.

Existing rating mechanisms have been shown to work well when monitoring and reporting are perfect. However, when monitoring and reporting are subject to errors, existing rating mechanisms cannot achieve the social optimum [93]–[104]. The errors, which are often encountered in practice, may arise either from the client's own incapability of accurate assessment (for instance, the client, who wants

[2] Note that the rating mechanisms studied in this paper focus on dealing with moral hazard problems, namely the server's quality of service is not perfectly observable. They are different from the rating mechanisms dealing with adverse selection problems, namely the problems of identifying the users' types. See [97, Sec. I] for detailed discussions on the above two classes of rating mechanisms.

[3] Different terminologies have been used in the existing literature. For example, [97][98] used "reputation" for "rating", and [98] used "social norm" for "recommended strategy".

to translate some sentences into a foreign language, cannot accurately evaluate the server's translation), or from some system errors (for example, the client's report on the server's service quality is missing due to network errors)[4]. In the presence of errors, the server's rating may be wrongly updated. Hence, even if the users follow the recommended desirable behavior, the platform may still fall into some "bad" states in which many users have low ratings due to erroneous rating updates. In these bad states, the users with low ratings receive low-quality services, resulting in large performance loss compared to the social optimum. This performance loss in the bad states is the major reason for the inefficiency of the existing rating mechanisms.

In this paper, we propose the first rating mechanisms that can achieve the social optimum even under imperfect monitoring. A key feature of the proposed rating mechanism is the *nonstationary* recommended strategy, which recommends different behaviors under the same system state, depending on when this state occurs (for example, the rating mechanism may not always recommend punishing users with low ratings in the bad states). Note, importantly, that the rating mechanism does not just randomize over different behaviors with a fixed probability in a state. Instead, it recommends different behaviors in the current state based on the history of past states. We design the recommended strategy carefully, such that the punishments happen frequently enough to provide sufficient incentives for the users, but not too frequently to reduce the performance loss incurred in the bad states. The more patient the users are (i.e. the larger discount factor they have), the less frequent are the punishments. As a result, the designed rating mechanism can asymptotically achieve the social optimum as the users become increasingly patient (i.e. as the discount factor approaches 1). This

[4] Note that the errors in this paper are not caused by the strategic behaviors of the users. In other words, the clients report the service quality truthfully, and do not misreport intentionally to manipulate the rating mechanism for their own interests. If the clients may report strategically, the mechanism can let the platform to assess the service quality (still, with errors) to avoid strategic reporting.

is in contrast with the existing rating mechanisms with *stationary* recommended strategies, whose performance loss does not vanish even as the users' patience increases. Another key feature of the proposed rating mechanism is the use of differential punishments that punish users with different ratings differently. In Section 4.4, we show that the absence of any one of these two features in our mechanism will result in performance loss that does not vanish even when the users are arbitrarily patient.

We prove that the social optimum can be achieved by simple rating mechanisms, which assign *binary* ratings to the users and recommend a small set of *three* recommended behaviors. We provide design guidelines of the rating update rules in socially-optimal rating mechanisms, and a low-complexity online algorithm to construct the nonstationary recommended strategies. The algorithm essentially solves a system of two linear equations with two variables in each period, and can be implemented with a memory of a fixed size (although by the definition of nonstationary strategies, it appears that we may need a memory growing with time to store the history of past states), because we can appropriately summarize the history of past states (by the solution to the above linear equations).

The rest of the paper is organized as follows. In Section 4.2, we discuss the differences between our work and related works. In Section 4.3, we describe the model of service exchange systems with rating mechanisms. Then we design the optimal rating mechanisms in Section 4.5. Simulation results in Section 4.6 demonstrate the performance improvement of the proposed rating mechanism. Finally, Section 4.7 concludes the paper.

Table 4.1: Related Works on Rating Protocols.

	Rating update error	Recommended strategy	Discount factor	Performance loss
[93][94]	$\to 0$	Stationary	< 1	Yes
[95][96]	> 0	Stationary	< 1	Yes
[97]	> 0	Stationary/Nonstationary	< 1	Yes
[98]–[103]	$= 0$	Stationary	$\to 1$	Yes
[104]	$\to 0$	Stationary	$\to 1$	Yes
This work	> 0	Nonstationary	< 1	No

4.2 Related Works

4.2.1 Related Works on Rating Protocols

Rating mechanisms were originally proposed by [98] for a large anonymous society, in which users are repeatedly randomly matched to play the Prisoners' dilemma game. Assuming perfect monitoring, [98] proposed a simple rating mechanism that can achieve the social optimum: any user who has defected will be assigned with the lowest rating forever and will be punished by its future partners. Subsequent research has been focusing on extending the results to more general games (see [99][100][101][103]), or on discovering alternative mechanisms (for example, [102] showed that cooperation can be sustained if each user can observe its partner's past actions). However, all these works assumed perfect monitoring and were aimed at dealing with the incentive problems caused by the anonymity of users and the lack of fixed partnership; they did not study the impact of imperfect monitoring. Under imperfect observation/reporting, the system will collapse under their rating mechanisms because all the users will eventually end up with having low ratings forever due to errors.

Some works [93][94][104] assumed imperfect monitoring, but focused on the limit case when the monitoring tends to be perfect. The conclusion of these works is that the social optimum can be achieved in the limit case when the monitoring becomes "almost perfect" (i.e., when the rating update error goes to zero).

Only a few works [95]–[97] analyzed rating mechanisms under imperfect monitoring with *fixed nonzero* monitoring errors. For a variety of rating mechanisms studied in [95]–[97], the performance loss with respect to the social optimum is quantified in terms of the rating update error. These results confirm that existing rating mechanisms suffer from (severe) performance loss under rating update errors. Note that the model in [97] is fundamentally different than ours. In [97], there is only a single long-lived seller (server), while all the buyers (clients) are short-lived. Under this model, it is shown in [97] that the rating mechanism is bounded away from social optimum even when nonstationary strategies are used. In contrast, we show that under our model with long-lived servers and clients, we can achieve the social optimum by nonstationary strategies with differential punishments. In the following, we discuss the intuitions of how to achieve the social optimum under our model.

There are two sources of inefficiency. One source of inefficiency comes from the stationary recommended strategies, which recommends the same behavior under the same state [93]–[96][98]–[104]. As we have discussed earlier, the inefficiency of the existing rating mechanisms comes from the punishments triggered in the "bad" states. Specifically, to give incentives for the users to provide high-quality services, the rating mechanism *must* punish the low-rating users under certain rating distributions (i.e. under certain "bad" states). When the users are punished (i.e. they are provided with low-quality services), the average payoffs in these states are far below the social optimum. In the presence of rating update errors, the bad states happen with a probability bounded above zero (the lower bound depends only on the rating update error). As a result, the low payoffs occur with a frequency bounded above zero, which incurs an efficiency loss that cannot vanish unless the rating update error goes to zero.

Another source of inefficiency is the lack of differential punishments. As will be proved in Section 4.4, the rating mechanisms with no differential punishment have performance loss even when nonstationary recommended strategies are used.

This paper is the first to propose a class of rating mechanisms that achieve the social optimum even when update errors do not tend to zero. Our mechanisms rely on (explicitly-constructed) *nonstationary* strategies with differential punishments. The key intuitions of why the proposed mechanism achieves social optimum are as follows. First, nonstationary strategies punish the users in the bad states only when necessary, depending on the history of past states. In this way, nonstationary strategies can lower the frequency of punishment in the bad states to a level just enough to provide sufficient incentives for the users to provide high-quality services. In addition, differential punishment further reduces the loss in social welfare by transferring payoffs from low-rating users to high-rating users, instead of lowering everyone's payoff with non-differential punishment.

In Table 4.1, we compare the proposed work with existing rating mechanisms.

4.2.2 Related Works in Game Theory Literature

Our results are related to folk theorem results for repeated games [9] and stochastic games [81]. However, these existing folk theorem results [9][81] cannot be directly applied to our model. First, the results in [9] are derived for repeated games, in which every stage game is the same. Our system is modeled as a stochastic game, in which the stage games may be different because of the rating distributions.

Second, there do exist folk theorems for stochastic games [81], but they also do not apply to our model. The folk theorems [81] apply to standard stochastic games, in which the state must satisfy the following properties: 1) the state, together with the plan profile, uniquely determines the stage-game payoff, and 2) the state is known to all the users. In our model, since each user's stage game payoff depends on its own rating, each user's rating must be included in the state and be known to all the users. In other words, if we model the system as a standard stochastic game in order to apply the folk theorems, we need to define the state as the rating profile of all the users (not just the rating distribution). Then, the folk theorem states that the social optimum can be asymptotically achieved by strategies that depend on the history of rating profiles. However, in our model, the players do not know the full rating profile, but only know the rating distribution. Hence, the strategy can use only the information of rating distributions.[5] Whether such strategies can achieve the social optimum is not known according to the folk theorems; we need to prove the existence of socially optimal strategies that use only the information of rating distributions.

In addition, our results are fundamentally different from the folk theorem results [9][81] in nature. First, [9][81] focus on the limit case when the discount factor goes to one, which is not realistic because the users are not sufficiently patient. More importantly, the results in [9][81] are not constructive. They focus on *what* payoff profiles are achievable, but cannot show *how* to achieve those payoff profiles. They do not determine a lower bound on discount factors that admit equilibrium strategy profiles yielding the target payoff profile, and hence cannot construct equilibrium strategy profiles. By contrast, we do determine a lower bound on discount factors that admit equilibrium strategy profiles yielding the target payoff profile, and do construct equilibrium strategy profiles.

4.2.3 Related Mathematical Frameworks

Rating mechanisms with stationary recommended strategies can be designed by extending Markov decision processes (MDPs) in two important and non-trivial ways [93][94][105][106]: 1) since there are multiple users, the value of each state is a vector of all the users' values, instead of a scalar in standard MDPs, and 2)

[5] We insist on restricting to strategies that depend only on the history of rating distributions because in practice, 1) the platform may not publish the full rating profile due to informational and privacy constraints, and 2) even if the platform does publish such information, it is impractical to assume that the users can keep track of it.

Table 4.2: Related Mathematical Frameworks.

	Standard MDP	Extended MDP [93][94][105][106]	Self-generating sets [9][14][81]	This work
# of users	Single	Multiple	Multiple	Multiple
Value function	Single-valued	Single-valued	Set-valued	Set-valued
Incentive constraints	No	Yes	Yes	Yes
Strategies	Stationary	Stationary	Nonstationary	Nonstationary
Discount factor	< 1	< 1	$\to 1$	< 1
Constructive	Yes	Yes	No	Yes

the incentive compatibility constraints of self-interested users need to be fulfilled (e.g., the values of "good" states, in which most users have high ratings, should be sufficiently larger than those of "bad" states, such that users are incentivized to obtain high ratings), while standard MDPs do not impose such constraints.

In this paper, we make a significant step forward with respect to the state-of-the-art rating mechanisms with stationary strategies: we design rating mechanisms where the recommended strategies can be nonstationary. The proposed design leads to significant performance improvements, but is also significantly more challenging from a theoretical perspective. The key challenge is that nonstationary strategies may choose different actions under the same state, resulting in possibly different current payoffs in the same state. Hence, the value function under nonstationary strategies are *set-valued*, which significantly complicates the analysis, compared to *single-valued* value functions under stationary strategies[6].

The mathematical framework of analyzing nonstationary strategies with set-valued value functions was proposed as a theory of self-generating sets in [14]. It was widely used in game theory to prove folk theorems in repeated games [9] and stochastic games [81]. We have discussed our differences from the folk theorem results [9][81] in the previous subsection.

[6] In randomized stationary strategies, although different actions may be taken in the same state *after randomization*, the probability of actions chosen is fixed. In the Bellman equation, we need to use the *expected payoff before randomization* which is fixed in the same state, instead of the realized payoffs after randomization. Hence, the value function is still single-valued.

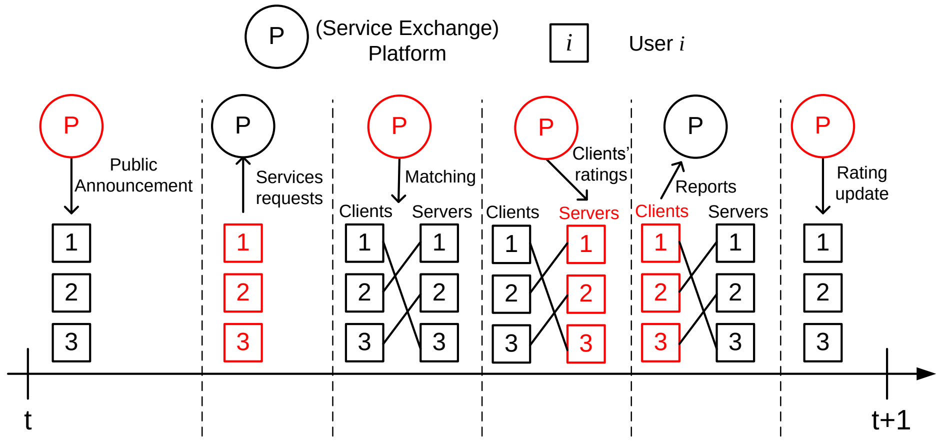


Figure 4.1: Illustration of the rating mechanism in one period.

In Table 4.2, we compare our work with existing mathematical frameworks.

4.3 System Model and Problem Formulation

4.3.1 System Model

4.3.1.1 The Rating Mechanism

We consider a service exchange platform with a set of N users, denoted by $\mathcal{N} = \{1, \ldots, N\}$. Each user can provide some services (e.g. data in P2P networks, labor in Amazon Mechanic Turk) valuable to the other users. The rating mechanism assigns each user i a binary label $\theta_i \in \Theta \triangleq \{0, 1\}$, and keep record of the *rating profile* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$. Since the users usually stay in the platform for a long period of time, we divide time into periods indexed by $t = 0, 1, 2, \ldots$. In each period, the rating mechanism operates as illustrated in Fig. 4.1, which can be roughly described as follows:

- Each user requests services as a *client*.
- Each user, as a *server*, is matched to another user (its client) based on a matching rule.

- Each server chooses to provide high-quality or low-quality services.
- Each client reports its assessment of the service quality to the rating mechanism, who will update the server's rating based on the report.

Next, we describe the key components in the rating mechanism in details.

Public announcement: At the beginning of each period, the platform makes public announcement to the users. The public announcement includes the rating distribution and the recommended plan in this period. The *rating distribution* indicates how many users have rating 1 and rating 0, respectively. Denote the rating distribution by $\boldsymbol{s}(\boldsymbol{\theta}) = (s_0(\boldsymbol{\theta}), s_1(\boldsymbol{\theta}))$, where $s_1(\boldsymbol{\theta}) = \sum_{i \in \mathcal{N}} \theta_i$ is the number of users with rating 1, and $s_0(\boldsymbol{\theta}) = N - s_1(\boldsymbol{\theta})$ is the number of users with rating 0. Denote the set of all possible rating distributions by S. Note that the platform does not disclose the rating profile $\boldsymbol{\theta}$ for privacy concerns. The platform also recommends a desired behavior in this period, called *recommended plan*. The recommended plan is a contingent plan of which service quality the server should choose based on its own rating and its client's rating. Formally, the recommended plan, denoted by α_0, is a mapping $\alpha_0 : \Theta \times \Theta \to \{0, 1\}$, where 0 and 1 represent "low-quality service" and "high-quality service", respectively. Then $\alpha_0(\theta_c, \theta_s)$ denotes the recommended service quality for a server with rating θ_s when it is matched to a client with rating θ_c. We write the set of recommended plans as $\mathcal{A} = \{\alpha | \alpha : \Theta \times \Theta \to \{0, 1\}\}$. We are particularly interested in the following three plans. The plan α^{a} is the *altruistic* plan:

$$\alpha^{\mathrm{a}}(\theta_c, \theta_s) = 1, \forall \theta_c, \theta_s \in \{0, 1\}, \tag{4.1}$$

where the server provides high-quality service regardless of its own and its client's ratings. The plan α^{f} is the *fair* plan:

$$\alpha^{\mathrm{f}}(\theta_c, \theta_s) = \begin{cases} 0 & \theta_s > \theta_c \\ 1 & \theta_s \leq \theta_c \end{cases}, \tag{4.2}$$

where the server provides high-quality service when its client has higher or equal ratings. The plan α^{s} is the *selfish* plan:

$$\alpha^{\mathrm{s}}(\theta_c, \theta_s) = 0, \forall \theta_c, \theta_s \in \{0, 1\}, \tag{4.3}$$

where the server provides low-quality service regardless of its own and its client's ratings. Note that we can consider the selfish plan as a non-differential punishment in which everyone receives low-quality services, and consider the fair plan as a differential punishment in which users with different ratings receive different services.

Service requests: The platform receives service requests from the users. We assume that there is no cost in requesting services, and that each user always have demands for services. Hence, all the users will request services.

Matching: The platform matches each user i, as a client, to another user $m(i)$ who will serve i, where m is a matching $m : \mathcal{N} \rightarrow \mathcal{N}$. Since the platform cannot match a user to itself, we write the set of all possible matchings as $M = \{m : m \text{ b}ijective,\ m(i) \neq i, \forall i \in \mathcal{N}\}$. The mechanism defines a random matching rule, which is a probability distribution μ on the set of all possible matchings M. In this paper, we focus on the uniformly random matching rule, which chooses every possible matching $m \in M$ with the same probability. The analysis can be easily generalized to the cases with non-uniformly random matching rules, as long as the matching rules do not distinguish users with the same rating.

Clients' ratings: The platform will inform each server of its client's rating, such that each server can choose its service quality based on its own and its client's ratings.

Reports: After the servers serve their clients, the platform elicits reports from the clients about their service quality. However, the report is *inaccurate*, either by the client's incapability of accurate assessment (for instance, the client, who wants to translate some sentences into a foreign language, cannot accurately evaluate the server's translation) or by some system error (for example, the data/file sent by the server is missing due to network errors). We characterize the erroneous report by a mapping $R : \{0,1\} \rightarrow \Delta(\{0,1\})$, where $\Delta(\{0,1\})$ is the probability distribution over $\{0,1\}$. For example, $R(1|q)$ is the probability that the client reports "high quality" given the server's actual service quality q. In this paper, we focus on reports of the following form

$$R(r|q) = \begin{cases} 1-\varepsilon, & r = q \\ \varepsilon, & r \neq q \end{cases}, \ \forall r, q \in \{0,1\}, \tag{4.4}$$

where $\varepsilon \in [0, 0.5)$ is the report error probability.[7] Note, however, that reporting is not strategic: the client reports truthfully, but with errors. If the clients report strategically, the mechanism can let the platform to assess the service quality (still, with errors) to avoid strategic reporting. For simplicity, we assume that the report error is symmetric, in the sense that reporting high and low qualities have the same error probability. Extension to asymmetric report errors is straightforward.

Rating update: Given the clients' reports, the platform updates the servers' ratings according to the *rating update rule*, which is defined as a mapping $\tau : \Theta \times \Theta \times \{0,1\} \times \mathcal{A} \rightarrow \Delta(\Theta)$. For example, $\tau(\theta_s'|\theta_c, \theta_s, r, \alpha_0)$ is the probability of the server's updated rating being θ_s', given the client's rating θ_c, the server's own rating θ_s, the client's report r, and the recommended plan α_0. We focus on the following class of rating update rules (see Fig. 4.2 for illustration):

$$\tau(\theta_s'|\theta_c, \theta_s, r, \alpha_0) = \begin{cases} \beta_{\theta_s}^{+}, & \theta_s' = 1, r \geq \alpha_0(\theta_c, \theta_s) \\ 1-\beta_{\theta_s}^{+}, & \theta_s' = 0, r \geq \alpha_0(\theta_c, \theta_s) \\ 1-\beta_{\theta_s}^{-}, & \theta_s' = 1, r < \alpha_0(\theta_c, \theta_s) \\ \beta_{\theta_s}^{-}, & \theta_s' = 0, r < \alpha_0(\theta_c, \theta_s) \end{cases}.$$

[7] We confine the report error probability ε to be smaller than 0.5. If the error probability ε is 0.5, the report contains no useful information about the service quality. If the error probability is larger than 0.5, the rating mechanism can use the opposite of the report as an indication of the service quality, which is equivalent to the case with the error probability smaller than 0.5.

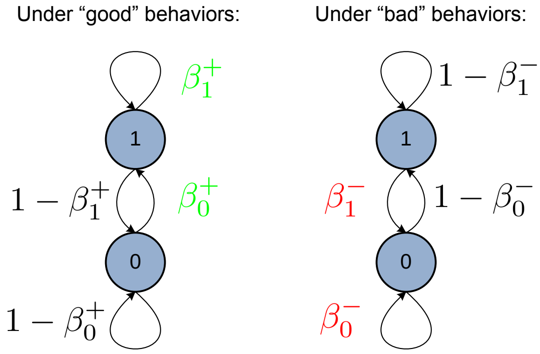


Figure 4.2: Illustration of the rating update rule. The circle denotes the rating, and the arrow denotes the rating update with corresponding probabilities.

In the above rating update rule, if the reported service quality is not lower than the recommended service quality, a server with rating θ_s will have rating 1 with probability $\beta_{\theta_s}^+$; otherwise, it will have rating 0 with probability $\beta_{\theta_s}^-$. Other more elaborate rating update rules may be considered. But we show that this simple one is good enough to achieve the social optimum.

Recommended strategy: The final key component of the rating mechanism is the recommended *strategy*, which determines what recommended plan should be announced in each period. In each period t, the mechanism keeps track of the history of rating distributions, denoted by $\boldsymbol{h}^t = (\boldsymbol{s}^0, \ldots, \boldsymbol{s}^t) \in S^{t+1}$, and chooses the recommended plan based on $\boldsymbol{h}^t$. In other words, the recommended strategy is a mapping from the set of histories to its plan set, denoted by $\pi_0 : \cup_{t=0}^{\infty} S^{t+1} \rightarrow \mathcal{A}$. Denote the set of all recommended strategies by Π. Note that although the rating mechanism knows the rating profile, it determines the recommended plan based on the history of rating distributions, because 1) this reduces the computational and memory complexity of the protocol, and 2) it is easy for the users to follow since they do not know the rating profile. Moreover, since the plan set $\mathcal{A}$ has 16 elements, the complexity of choosing the plan is large. Hence, we consider the strategies that choose plans from a subset $\mathcal{B} \subseteq \mathcal{A}$, and define $\Pi(\mathcal{B})$ as the set of

Table 4.3: Gift-Giving Game Between A Client and A Server.

	high-quality	low-quality
request	$(b, -c)$	$(0, 0)$

strategies restricted on the subset $\mathcal{B}$ of plans.

In summary, the rating mechanism can be represented by the design parameters: the rating update rule and the recommended strategy, and can be denoted by the tuple (τ, π_0).

4.3.1.2 Payoffs

Once a server and a client are matched, they play the gift-giving game in Table 4.3 [93]–[98][102][104], where the row player is the client and the column player is the server. We normalize the payoffs received by the client and the server when a server provides low-quality services to 0. When a server provides high-quality services, the client gets a benefit of $b > 0$ and the worker incurs a cost of $c \in (0, b)$. In the unique Nash equilibrium of the gift-giving game, the server will provide low-quality services, which results in a zero payoff for both the client and the server. Note that as in [93]–[98][102][104], we assume that the same gift-giving game is played for all the client-server pairs. This assumption is reasonable when the number of users is large. Since b can be considered as a user's expected benefit across different servers, and c as its expected cost of high-quality service across different clients, the users' expected benefits/costs should be approximately the same when the number of users is large. This assumption is also valid when the users have different realized benefit and cost in each period but the same expected benefit b and expected cost c across different periods.

Expected payoff in one period: Based on the gift-giving game, we can calculate each user's expected payoff obtained in one period. A user's expected payoff in one period depends on its own rating, the rating distribution, and the users' plans. We write user i's plan as $\alpha_i \in \mathcal{A}$, and the plan profile of all the users as $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$. Then user i's expected payoff in one period is $u_i(\theta_i, \boldsymbol{s}, \boldsymbol{\alpha})$. For illustration, we calculate the users' expected payoffs under several important scenarios, assuming that all the users follow the recommended plan (i.e. $\alpha_i = \alpha_0, \ \forall i \in \mathcal{N}$). When the altruistic plan α^{a} is recommended, all the users receive the same expected payoff in one period as

$$u_i(\theta_i, \boldsymbol{s}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) = b - c, \ \forall i, \theta_i, \boldsymbol{s},$$

where $\alpha \cdot \mathbf{1}_N$ is the plan profile in which every user chooses plan α. Similarly, when the selfish plan α^{s} is recommended, all the users receive zero expected payoff in one period, namely

$$u_i(\theta_i, \boldsymbol{s}, \alpha^{\mathrm{s}} \cdot \mathbf{1}_N) = 0, \ \forall i, \theta_i, \boldsymbol{s}.$$

When the fair plan α^{f} is recommended, the users receive expected payoffs in one period as follows

$$u_i(\theta_i, \boldsymbol{s}, \alpha^{\mathrm{f}} \cdot \mathbf{1}_N) = \begin{cases} \frac{s_0 - 1}{N-1} \cdot b - c, & \theta_i = 0 \\ b - \frac{s_1 - 1}{N-1} \cdot c, & \theta_i = 1 \end{cases} . \tag{4.5}$$

Under the fair plan, the users with rating 1 receive a payoff higher than $b - c$, because they get high-quality services from everyone and provide high-quality services only when matched to clients with rating 1. In contrast, the users with rating 0 receive a payoff lower than $b-c$. Hence, the fair plan α^{f} can be considered as a differential punishment.

Discounted average payoff: Each user i has its own strategy $\pi_i \in \Pi$. Write the joint strategy profile of all the users as $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. Then given the initial rating profile $\boldsymbol{\theta}^0$, the recommended strategy π_0 and the joint strategy profile $\boldsymbol{\pi}$ induce a probability distribution over the sequence of rating profiles $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots$. Taking the expectation with respect to this probability distribution, each user i

receives a discounted average payoff $U_i(\boldsymbol{\theta}^0, \pi_0, \boldsymbol{\pi})$ calculated as

$$U_i(\boldsymbol{\theta}^0, \pi_0, \boldsymbol{\pi}) = \mathbb{E}_{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots} \left\{ (1-\delta) \sum_{t=0}^{\infty} \delta^t u_i \left(\theta_i^t, \boldsymbol{s}(\boldsymbol{\theta}^t), \boldsymbol{\pi}(\boldsymbol{s}(\boldsymbol{\theta}^0), \ldots, \boldsymbol{s}(\boldsymbol{\theta}^t)) \right) \right\}$$

where $\delta \in [0, 1)$ is the common discount factor of all the users. The discount factor δ is the rate at which the users discount future payoffs, and reflects the patience of the users. A more patient user has a larger discount factor. Note that the recommended strategy π_0 does affect the users' discounted average payoffs by affecting the evolution of the rating profile (i.e. by affecting the expectation operator $\mathbb{E}_{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots}$).

4.3.1.3 Definition of The Equilibrium

The platform adopts *sustainable rating mechanisms*, which specifies a tuple of rating update rule and recommended strategy (τ, π_0), such that the users find it in their self-interests to follow the recommended strategy. In other words, the recommended strategy should be an equilibrium.

Note that the interaction among the users is neither a repeated game [9] nor a standard stochastic game [81]. In a repeated game, every stage game is the same, which is clearly not true in the platform because users' stage-game payoff $u_i(\theta_i, \boldsymbol{s}, \boldsymbol{\alpha})$ depends on the rating distribution $\boldsymbol{s}$. In a standard stochastic game, the state must satisfy: 1) the state and the plan profile uniquely determines the stage-game payoff, and 2) the state is known to all the users. In the platform, the user's stage-game payoff $u_i(\theta_i, \boldsymbol{s}, \boldsymbol{\alpha})$ depends on its own rating θ_i, which should be included in the state and be known to all the users. Hence, if we were to model the interaction as a standard stochastic game, we need to define the state as the rating profile $\boldsymbol{\theta}$. However, the rating profile is not known to the users in our formulation.

To reflect our restriction on recommended strategies that depend only on rating distributions, we define the equilibrium as public announcement equilibrium

(PAE), since the strategy depends on the publicly announced rating distributions. Before we define PAE, we need to define the continuation strategy $\pi|_{\boldsymbol{h}^t}$, which is a mapping $\pi|_{\boldsymbol{h}^t} : \cup_{k=0}^{\infty} \mathcal{H}^k \rightarrow A$ with $\pi|_{\boldsymbol{h}^t}(\boldsymbol{h}^k) = \pi(\boldsymbol{h}^t \boldsymbol{h}^k)$, where $\boldsymbol{h}^t \boldsymbol{h}^k$ is the concatenation of $\boldsymbol{h}^t$ and $\boldsymbol{h}^k$.

Definition 3 *A pair of a recommended strategy and a symmetric strategy profile* $(\pi_0, \pi_0 \cdot \mathbf{1}_N)$ *is a PAE, if for all* $t \geq 0$*, for all* $\tilde{\boldsymbol{h}}^t \in \mathcal{H}^t$*, and for all* $i \in \mathcal{N}$*, we have*

$$U_i(\tilde{\boldsymbol{\theta}}^t, \pi_0|_{\tilde{\boldsymbol{h}}^t}, \pi_0|_{\tilde{\boldsymbol{h}}^t} \cdot \mathbf{1}_N) \geq U_i(\tilde{\boldsymbol{\theta}}^t, \pi_0|_{\tilde{\boldsymbol{h}}^t}, (\pi_i|_{\tilde{\boldsymbol{h}}^t}, \pi_0|_{\tilde{\boldsymbol{h}}^t} \cdot \mathbf{1}_{N-1})), \ \forall \pi_i|_{\tilde{\boldsymbol{h}}^t} \in \Pi,$$

where $(\pi_i|_{\tilde{\boldsymbol{h}}^t}, \pi_0|_{\tilde{\boldsymbol{h}}^t} \cdot \mathbf{1}_{N-1})$ *is the continuation strategy profile in which user* i *deviates to* $\pi_i|_{\tilde{\boldsymbol{h}}^t}$ *and the other users follow the strategy* $\pi_0|_{\tilde{\boldsymbol{h}}^t}$*.*

Note that in the definition, we allow a user to deviate to any strategy $\pi_i \in \Pi$, even if the recommended strategy π_0 is restricted to a subset $\mathcal{B}$ of plans. Hence, the rating mechanism is robust, in the sense that a user cannot gain even when it uses more complicated strategies. Note also that although a rating mechanism can choose the initial rating profile $\boldsymbol{\theta}^0$, we require a recommended strategy to fulfill the incentive constraints under all the initial rating profiles. This adds to the flexibility in choosing the initial rating profile.

PAE is stronger than the Nash equilibrium (NE), because PAE requires the users to not deviate following *any* history, while NE requires the users to not deviate following the histories that happen in the equilibrium. In this sense, PAE can be considered as a special case of public perfect equilibrium (PPE) in standard repeated and stochastic games, where the strategies depend only on the rating distribution.

4.3.2 The Rating Protocol Design Problem

The goal of the rating mechanism designer is to choose a rating mechanism (τ, π_0), such that the social welfare at the equilibrium is maximized in the worst case (with respect to different initial rating profiles). Maximizing the worst-case performance gives us a much stronger performance guarantee than maximizing the performance under a given initial rating profile. Given the rating update error ε, the discount factor δ, and the subset $\mathcal{B}$ of plans, the rating mechanism design problem is formulated as:

$$\begin{aligned} W(\varepsilon, \delta, \mathcal{B}) = \quad & \max_{\tau, \pi_0 \in \Pi(\mathcal{B})} \quad \min_{\boldsymbol{\theta}^0 \in \Theta^N} \frac{1}{N} \sum_{i \in \mathcal{N}} U_i(\boldsymbol{\theta}^0, \pi_0, \pi_0 \cdot \mathbf{1}_N) \\ & \qquad s.t. \qquad (\pi_0, \pi_0 \cdot \mathbf{1}_N) \text{ is a PAE.} \end{aligned} \tag{4.6}$$

Note that $W(\varepsilon, \delta, \mathcal{B})$ is strictly smaller than the social optimum $b-c$ for any ε, δ, and $\mathcal{B}$. This is because to exactly achieve $b-c$, the protocol must recommend the altruistic plan α^{a} all the time (even when someone shirks), which cannot be an equilibrium. However, we can design rating mechanisms such that for any fixed $\varepsilon \in [0, 0.5)$, $W(\varepsilon, \delta, \mathcal{B})$ can be arbitrarily close to the social optimum. In particular, such rating mechanisms can be simple, in that $\mathcal{B}$ can be a small subset of three plans (i.e. $\mathcal{B} = A^{\mathrm{a}fs} \triangleq \{\alpha^{\mathrm{a}}, \alpha^{\mathrm{f}}, \alpha^{\mathrm{s}}\}$).

4.4 Sources of Inefficiency

To illustrate the importance of designing optimal, yet simple rating schemes, as well as the challenges associated with determining such a design, in this section, we discuss several simple rating mechanisms that appear to work well intuitively, and show that they are actually bounded away from the social optimum even when the users are arbitrarily patient. We will illustrate why they are inefficient, which gives us some insights on how to design socially-optimal rating mechanisms.

4.4.1 Stationary Recommended Strategies

4.4.1.1 Analysis

We first consider rating mechanisms with stationary recommended strategies, which determine the recommended plan solely based on the current rating distribution. Since the game is infinitely-repeatedly played, given the same rating distribution, the continuation game is the same regardless of when the rating distribution occurs. Hence, similar to MDP, we can assign value functions $V_\theta^{\pi_0} : S \to \mathbb{R}$, $\forall \theta$ for a stationary strategy π_0, with $V_\theta^{\pi_0}(\boldsymbol{s})$ being the continuation payoff of a user with rating θ at the rating distribution $\boldsymbol{s}$. Then, we have the following set of equalities that the value function needs to satisfy:

$$\begin{aligned} V_{\theta_i}^{\pi_0}(\boldsymbol{s}) &= (1-\delta)\cdot u_i(\pi_0(\boldsymbol{s}), \pi_0(\boldsymbol{s})\cdot \mathbf{1}_N) \\ &+ \delta \cdot \sum_{\theta_i', \boldsymbol{s}'} \Pr(\theta_i', \boldsymbol{s}' | \theta_i, \boldsymbol{s}, \pi_0(\boldsymbol{s}), \pi_0(\boldsymbol{s})\cdot \mathbf{1}_N) \cdot V_{\theta_i'}^{\pi_0}(\boldsymbol{s}'), \forall i \in \mathcal{N}, \end{aligned} \tag{4.7}$$

where $\Pr(\theta_i', \boldsymbol{s}' | \theta_i, \boldsymbol{s}, \pi_0(\boldsymbol{s}), \pi_0(\boldsymbol{s})\cdot \mathbf{1}_N)$ is the transition probability. We can solve for the value function from the above set of equalities, which are similar to the Bellman equations in MDP. However, note that obtaining the value function is not the final step. We also need to check the incentive compatibility (IC) constraints. For example, to prevent user i from deviating to plan α', the following inequality has to be satisfied:

$$\begin{aligned} V_{\theta_i}^{\pi_0}(\boldsymbol{s}) &\geq (1-\delta)\cdot u_i(\pi_0(\boldsymbol{s}), (\alpha', \pi_0(\boldsymbol{s})\cdot \mathbf{1}_{N-1})) \\ &+ \delta \cdot \sum_{\theta_i', \boldsymbol{s}'} \Pr(\theta_i', \boldsymbol{s}' | \theta_i, \boldsymbol{s}, \pi_0(\boldsymbol{s}), (\alpha', \pi_0(\boldsymbol{s})\cdot \mathbf{1}_{N-1})) \cdot V_{\theta_i'}^{\pi_0}(\boldsymbol{s}'), \forall i \in \mathcal{N}. \end{aligned} \tag{4.8}$$

Given a rating mechanism with a stationary recommended strategy π_0, if its value function satisfies all the IC constraints, we can determine the social welfare of the rating mechanism. For example, suppose that all the users have an initial rating of 1. Then, all of them achieve the expected payoff $V_1^{\pi_0}(0, N)$, which is the social welfare achieved under this rating mechanism.

Note that given a recommended strategy π_0, it is not difficult to compute the value function by solving the set of linear equations in (4.7) and check the IC constraints according to the set of linear inequalities in (4.8). However, it is difficult to derive structural results on the value function (e.g. whether the state with more rating-1 users has a higher value), and thus difficult to know the structures of the optimal recommended strategy (e.g. whether the optimal recommended strategy is a threshold strategy). The difficulty mainly comes from the complexity of the transition probabilities $\Pr(\theta_i', \boldsymbol{s}' | \theta_i, \boldsymbol{s}, \pi_0(\boldsymbol{s}), \pi_0(\boldsymbol{s}) \cdot \mathbf{1}_N)$. For example, assuming $\pi_0(\boldsymbol{s}) = \alpha^{\mathrm{a}}$, we have

$$\Pr(1, \boldsymbol{s}' | 1, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) = x_1^+ \cdot$$
$$\textstyle\sum_{k=\max\{0, s_1'-1-(N-s_1)\}}^{\min\{s_1-1, s_1'-1\}} \binom{s_1-1}{k} (x_1^+)^k (1-x_1^+)^{s_1-1-k} \binom{N-s_1}{s_1'-1-k} (x_0^+)^{s_1'-1-k} (1-x_0^+)^{N-s_1-(s_1'-1-k)},$$

where $x_1^+ \triangleq (1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)$ is the probability that a rating-1 user's rating remains to be 1, and $x_0^+ \triangleq (1-\varepsilon)\beta_0^+ + \varepsilon(1-\beta_0^-)$ is the probability that a rating-0 user's rating goes up to 1. We can see that the transition probability has combinatorial numbers in it and is complicated. Hence, although the stationary strategies themselves are simpler than the nonstationary strategies, they are *harder to compute*, in the sense that it is difficult to derive structural results for rating mechanisms with stationary recommended strategy. In contrast, we are able to develop a unified design framework for socially-optimal rating mechanisms with nonstationary recommended strategies.

4.4.1.2 Inefficiency

We measure the efficiency of the rating mechanisms with stationary recommended strategies using the "price of stationarity" (PoStat), defined as

$$\mathrm{PoStat}(\varepsilon, \mathcal{B}) = \frac{\lim_{\delta \to 1} W^s(\varepsilon, \delta, \mathcal{B})}{b-c}, \tag{4.9}$$

where $W^s(\varepsilon, \delta, \mathcal{B})$ is the optimal value of a modified optimization problem (4.6) with an additional constraint that π_0 is stationary.

Table 4.4: Normalized social welfare of stationary strategies restricted on $A^{\mathrm{a}fs}$.

δ	0.7	0.8	0.9	0.99	0.999	0.9999
Normalized welfare	0.690	0.700	0.715	0.720	0.720	0.720

Note that PoStat$(\varepsilon, \mathcal{B})$ measures the efficiency of a class of rating mechanisms (not a specific rating mechanism), because we optimize over all the rating update rules and stationary recommended strategies restricted on $\mathcal{B}$. PoStat is a number between 0 and 1. A small PoStat indicates a low efficiency.

Through simulation, we can compute PoStat$(0.1, A^{\mathrm{a}fs}) = 0.720$. In other words, even with differential punishment α^{f}, the performance of stationary strategies is bounded away from social optimum. We compute the PoStat in a platform with $N = 5$ users, the benefit $b = 3$, the cost $c = 1$, and $\varepsilon = 0.1$. Under each discount factor δ, we assign values between 0 and 1 with a 0.1 grid to $\beta_\theta^+, \beta_\theta^-$ in the rating update rule, namely we try 11^4 rating update rules to select the optimal one. For each rating update rule, we try all the $3^{N+1} = 729$ stationary recommended strategies restricted on $A^{\mathrm{a}fs}$. In Table 4.4, we list normalized social welfare under different discount factors.

As mentioned before, the inefficiency of stationary strategies is due to the punishment exerted under certain rating distributions. For example, the optimal recommended strategies discussed above recommend the selfish or fair plan when at least one user has rating 0, resulting in performance loss. One may think that when the users are more patient (i.e. when the discount factor is larger), we can use milder punishments by lowering the punishment probabilities β_1^- and β_0^-, such that the rating distributions with many low-rating users happen with less frequency. However, simulations on the above strategies show that, to fulfill the IC constraints, the punishment probabilities cannot be made arbitrarily small even when the discount factor is large. For example, Table 4.5 shows the minimum punishment probability β_1^- (which is smaller than β_0^-) of rating mechanisms restricted

Table 4.5: Minimum punishment probabilities of rating mechanisms restricted on $A^{\mathrm{a}fs}$ when $\varepsilon = 0.1$.

δ	0.7	0.8	0.9	0.99	0.999	0.9999	0.99999
Minimum β_1^-	0.8	0.8	0.6	0.6	0.3	0.3	0.3

on $A^{\mathrm{a}fs}$ under different discount factors. In other words, the rating distributions with many low-rating users will happen with some probabilities bounded above zero, with a bound independent of the discount factor. Hence, the performance loss is bounded above zero regardless of the discount factor. Note that in a nonstationary strategy, we could choose whether to punish in rating distributions with many low-rating users, depending on the history of past rating distributions. This adaptive adjustment of punishments allows nonstationary strategies to potentially achieve the social optimum.

4.4.2 Lack of Differential Punishments

We have discussed in the previous subsection the inefficiency of stationary strategies. Now we consider a class of nonstationary strategies restricted on the subset of plans $A^{\mathrm{a}s}$. Under this class of strategies, all the users are rewarded (by choosing α^{a}) or punished (by choosing α^{s}) simultaneously. In other words, there is no differential punishment that can "transfer" some payoff from low-rating users to high-rating users. We quantify the performance loss of this class of nonstationary strategies restricted on $A^{\mathrm{a}s}$ as follows.

Proposition 4 *For any $\varepsilon > 0$, we have*

$$\lim_{\delta \to 1} W(\varepsilon, \delta, A^{as}) \leq b - c - \zeta(\varepsilon), \tag{4.10}$$

where $\zeta(\varepsilon) > 0$ for any $\varepsilon > 0$.

Proof: The proof is similar to the proof of [97, Proposition 6]; see Appendix 4.8.1. □

The above proposition shows that the maximum social welfare achievable by $(\pi_0, \pi \cdot \mathbf{1}_N) \in \Pi(A^{\mathrm{a}s}) \times \Pi^N(A^{\mathrm{a}s})$ at the equilibrium is bounded away from the social optimum $b - c$, unless there is no rating update error. Note that the performance loss is independent of the discount factor. In contrast, we will show later that, if we can use the fair plan α^{f}, the social optimum can be asymptotically achieve when the discount factor goes to 1. Hence, the differential punishment introduced by the fair plan is crucial for achieving the social optimum.

4.5 Socially Optimal Design

In this section, we design rating mechanisms that asymptotically achieve the social optimum at the equilibrium, even when the rating update rule $\varepsilon > 0$. In our design, we use the APS technique, named after the authors of the seminal paper [14], which is also used to prove the folk theorem for repeated games in [9] and for stochastic games in [81]. We will briefly introduce the APS technique first. Meanwhile, more importantly, we will illustrate why we *cannot* use APS in our setting in the same way as [9] and [81] did. Then, we will show how we use APS in a different way in our setting, in order to design the optimal rating mechanism and to construct the equilibrium strategy. Finally, we analyze the performance of a class of simple but suboptimal strategies, which sheds light on why the proposed strategy can achieve the social optimum.

4.5.1 The APS Technique

APS [14] provides a characterization of the set of PPE payoffs. It builds on the idea of *self-generating sets* described as follows. Note that APS is used for standard stochastic games, and recall from our discussion in Section 4.2 that the state of the standard stochastic game is the rating profile $\boldsymbol{\theta}$. Then define a set $\mathcal{W}^{\boldsymbol{\theta}} \subset \mathbb{R}^N$ for every state $\boldsymbol{\theta} \in \Theta^N$, and write $(\mathcal{W}^{\boldsymbol{\theta}'})_{\boldsymbol{\theta}' \in \Theta^N}$ as the collection of these sets. Then we have the following definitions [14][81][8]. First, we say a payoff profile $\boldsymbol{v}(\boldsymbol{\theta}) \in \mathbb{R}^N$ is *decomposable* on $(\mathcal{W}^{\boldsymbol{\theta}'})_{\boldsymbol{\theta}' \in \Theta^N}$ given $\boldsymbol{\theta}$, if there exists a recommended plan α_0, an plan profile $\boldsymbol{\alpha}^*$, and a *continuation payoff function* $\boldsymbol{\gamma} : \Theta^N \rightarrow \cup_{\boldsymbol{\theta}' \in \Theta^N} \mathcal{W}^{\boldsymbol{\theta}'}$ with $\boldsymbol{\gamma}(\boldsymbol{\theta}') \in \mathcal{W}^{\boldsymbol{\theta}'}$, such that for all $i \in \mathcal{N}$ and for all $\alpha_i \in A$,

$$\begin{aligned} v_i &= (1-\delta) u_i(\boldsymbol{\theta}, \alpha_0, \boldsymbol{\alpha}^*) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha_0, \boldsymbol{\alpha}^*) \qquad (4.11) \\ &\geq (1-\delta) u_i(\boldsymbol{\theta}, \alpha_0, \alpha_i, \boldsymbol{\alpha}_{-i}^*) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha_0, \alpha_i, \boldsymbol{\alpha}_{-i}^*). \end{aligned}$$

Then, we say a set $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta} \in \Theta^N}$ is a self-generating set, if for any $\boldsymbol{\theta}$, every payoff profile $\boldsymbol{v}(\boldsymbol{\theta}) \in \mathcal{W}^{\boldsymbol{\theta}}$ is decomposable on $(\mathcal{W}^{\boldsymbol{\theta}'})_{\boldsymbol{\theta}' \in \Theta^N}$ given $\boldsymbol{\theta}$. The important property of self-generating sets is that any self-generating set is a set of PPE payoffs [14][81][8].

Based on the idea of self-generating sets, [9] and [81] proved the folk theorem for repeated games and stochastic games, respectively. However, we cannot use APS in the same way as [9] and [81] did for the following reason. We assume that the users do not know the rating profile of every user, and restrict our attention on symmetric PA strategy profiles. This requires that each user i cannot use the continuation payoff function $\gamma_i(\boldsymbol{\theta})$ directly. Instead, each user i should assign the same continuation payoff for the rating profiles that have the same rating distribution, namely $\gamma_i(\boldsymbol{\theta}) = \gamma_i(\boldsymbol{\theta}')$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ such that $\boldsymbol{s}(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta}')$.

4.5.2 Socially Optimal Design

As mentioned before, the social optimum $b - c$ can be exactly achieved only by servers providing high-quality service all the time, which is not an equilibrium. Hence, we aim at achieving the social optimum $b - c$ asymptotically. We define the asymptotically socially optimal rating mechanisms as follows.

Definition 4 (Asymptotically Socially Optimal Rating Mechanisms) *Given a rating update error* $\varepsilon \in [0, 0.5)$, *we say a rating mechanism* $(\tau(\varepsilon), \pi_0(\varepsilon, \xi, \delta) \in \Pi)$

is asymptotically socially optimal under ε*, if for any small performance loss* $\xi > 0$*, we can find a* $\underline{\delta}(\xi)$*, such that for any discount factor* $\delta > \underline{\delta}(\xi)$*, we have*

- $(\pi_0(\xi, \delta), \pi_0(\xi, \delta) \cdot \mathbf{1}_N)$ *is a PAE;*
- $U_i(\boldsymbol{\theta}^0, \pi_0, \pi_0 \cdot \mathbf{1}_N) \geq b - c - \xi, \ \forall i \in \mathcal{N}, \ \forall \boldsymbol{\theta}^0.$

Note that in the asymptotically socially optimal rating mechanism, the rating update rule depends only on the rating update error, and works for any tolerated performance loss ξ and for any the discount factor $\delta > \underline{\delta}$. The recommended strategy π_0 is a class of strategies parameterized by $(\varepsilon, \xi, \delta)$, and works for any $\varepsilon \in [0, 0.5)$, any $\xi > 0$ and any discount factor $\delta > \underline{\delta}$ under the rating update rule $\tau(\varepsilon)$.

First, we define a few auxiliary variables first for better exposition of the theorem. Define $\kappa_1 \triangleq \frac{b}{\frac{N-2}{N-1}b-c} - 1$ and $\kappa_2 \triangleq 1 + \frac{c}{(N-1)b}$. In addition, we write the probability that a user with rating 1 has its rating remain at 1 if it follows the recommended altruistic plan α^{a} as:

$$x_1^+ \triangleq (1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-).$$

Write the probability that a user with rating 1 has its rating remain at 1 if it follows the recommended fair plan α^{f} as:

$$x_{s_1(\boldsymbol{\theta})} \triangleq \left[(1-\varepsilon)\frac{s_1(\boldsymbol{\theta})-1}{N-1} + \frac{N-s_1(\boldsymbol{\theta})}{N-1}\right]\beta_1^+ + \left(\varepsilon\frac{s_1(\boldsymbol{\theta})-1}{N-1}\right)(1-\beta_1^-).$$

Write the probability that a user with rating 0 has its rating increase to 1 if it follows the recommended plan α^{a} or α^{f}:

$$x_0^+ \triangleq (1-\varepsilon)\beta_0^+ + \varepsilon(1-\beta_0^-).$$

Theorem 7 *Given any rating update error* $\varepsilon \in [0, 0.5)$*,*

- (Design rating update rules): *A rating update rule* $\tau(\varepsilon)$ *that satisfies*

- *Condition 1 (following the recommended plan leads to a higher rating):*

$$\beta_1^+ > 1 - \beta_1^- \text{ and } \beta_0^+ > 1 - \beta_0^-,$$

- *Condition 2 (Enough "reward" for users with rating 1):*

$$x_1^+ = (1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-) > \frac{1}{1+\frac{c}{(N-1)b}},$$

- *Condition 3 (Enough "punishment" for users with rating 0):*

$$x_0^+ = (1-\varepsilon)\beta_0^+ + \varepsilon(1-\beta_0^-) < \frac{1-\beta_1^+}{\frac{c}{(N-1)b}},$$

can be the rating update rule in a asymptotically socially-optimal rating mechanism.

- (Optimal recommended strategies): *Given the rating update rule $\tau(\varepsilon)$ that satisfies the above conditions, any small performance loss $\xi > 0$, and any discount factor $\delta \geq \underline{\delta}(\varepsilon, \xi)$ with $\underline{\delta}(\varepsilon, \xi)$ defined in Appendix 4.8.2, the recommended strategy $\pi_0(\varepsilon, \xi, \delta) \in \Pi(A^{afs})$ constructed by Table 4.6 is the recommended strategy in a asymptotically socially-optimal rating mechanism.*

Proof: See Appendix 4.8.3 for the entire proof. We provide a proof sketch here.

The proof builds on the theory of self-generating sets [14], which can be considered as the extension of Bellman equations in dynamic programming to the cases with multiple self-interested users using nonstationary strategies. We can decompose each user i's discounted average payoff into the current payoff and the continuation payoff as follows:

$$\begin{aligned}
& U_i(\boldsymbol{\theta}^0, \pi_0, \boldsymbol{\pi}) \\
= \;& \mathbb{E}_{\boldsymbol{\theta}^1,\ldots}\left\{(1-\delta)\sum_{t=0}^{\infty}\delta^t u_i\left(\theta_i^t, \boldsymbol{s}(\boldsymbol{\theta}^t), \boldsymbol{\pi}(\boldsymbol{s}(\boldsymbol{\theta}^0), \ldots, \boldsymbol{s}(\boldsymbol{\theta}^t)\right)\right\} \\
= \;& (1-\delta)\cdot \underbrace{u_i\left(\theta_i^0, \boldsymbol{s}(\boldsymbol{\theta}^0), \boldsymbol{\pi}(\boldsymbol{s}(\boldsymbol{\theta}^0))\right)}_{\text{current payoff at } t=0} \\
+ \;& \underbrace{\delta \cdot \mathbb{E}_{\boldsymbol{\theta}^2,\ldots}\left\{(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1} u_i\left(\theta_i^t, \boldsymbol{s}(\boldsymbol{\theta}^t), \boldsymbol{\pi}(\boldsymbol{s}(\boldsymbol{\theta}^0), \ldots, \boldsymbol{s}(\boldsymbol{\theta}^t))\right)\right\}}_{\text{continuation payoff starting from } t=1}.
\end{aligned}$$

We can see that the continuation payoff starting from $t = 1$ is the discounted average payoff as if the system starts from $t = 1$. Suppose that the users follow the recommended strategy. Since the recommended strategy and the rating update rule do not differentiate users with the same rating, we can prove that the users with the same rating have the same continuation payoff starting from any point. Hence, given π_0 and $\boldsymbol{\pi} = \pi_0 \cdot \mathbf{1}_N$, the decomposition above can be simplified into

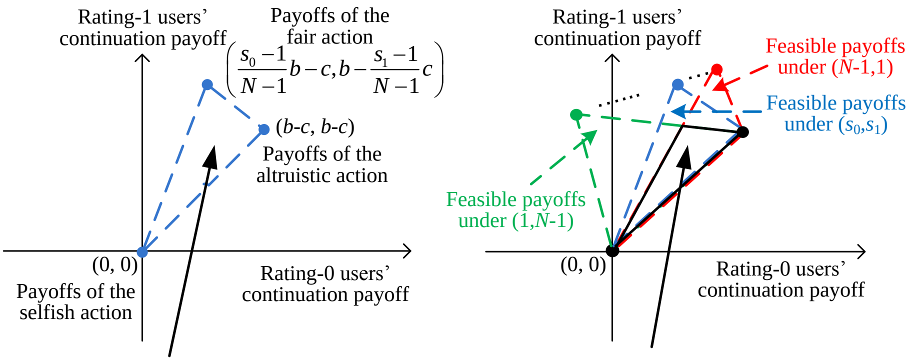
$$v_\theta^{\pi_0}(\boldsymbol{s}) = (1-\delta) \cdot u\left(\theta, \boldsymbol{s}, \alpha_0 \cdot \mathbf{1}_N\right) + \delta \cdot \sum_{\theta'=0}^{1} \sum_{\boldsymbol{s}'} q(\theta', \boldsymbol{s}' | \theta, \boldsymbol{s}, \alpha_0 \cdot \mathbf{1}_N) \cdot v_{\theta'}^{\pi_0}(\boldsymbol{s}'), \tag{4.12}$$

where $q(\theta', \boldsymbol{s}' | \theta, \boldsymbol{s}, \alpha \cdot \mathbf{1}_N)$ is the probability that the user has rating θ' and the rating distribution is $\boldsymbol{s}'$ in the next period given the user's current rating θ, the current rating distribution $\boldsymbol{s}$, and the action profile $\alpha \cdot \mathbf{1}_N$, and $v_\theta^{\pi_0}(\boldsymbol{s})$ is the continuation payoff of the users with rating θ starting from the initial rating distribution $\boldsymbol{s}$.

The differences between (4.12) and the Bellman equations are 1) the "value" of state $\boldsymbol{s}$ in (4.12) is a vector comprised of rating-1 and rating-0 users' values, compared to scalar values in Bellman equations, and 2) the value of state $\boldsymbol{s}$ is not fixed in (4.12), because the action α_0 taken under state $\boldsymbol{s}$ is not fixed in a nonstationary strategy (this is also the key difference from the analysis of stationary strategies; see (4.7) where the action taken at state $\boldsymbol{s}$ is fixed to be $\pi_0(\boldsymbol{s})$). In addition, for an equilibrium recommended strategy, the decomposition needs to satisfy the following incentive constraints: for all α,

$$\begin{gathered} v_\theta^{\pi_0}(\boldsymbol{s}) \geq \\ (1-\delta) \cdot u\left(\theta, \boldsymbol{s}, (\alpha, \alpha_0 \cdot \mathbf{1}_{N-1})\right) + \delta \cdot \textstyle\sum_{\theta'=0}^{1} \sum_{\boldsymbol{s}'} \rho(\theta', \boldsymbol{s}' | \theta, \boldsymbol{s}, (\alpha, \alpha_0 \cdot \mathbf{1}_{N-1})) \cdot v_{\theta'}^{\pi_0}(\boldsymbol{s}'). \end{gathered} \tag{4.13}$$

To analyze nonstationary strategies, we use the theory of self-generating sets. Note, however, that [14] does not tell us how to construct a self-generating set, which is exactly the major difficulty to overcome in our proof. In our proof, we construct the following self-generating set. First, since the strategies depend on rating distributions only, we let $\mathcal{W}(\boldsymbol{\theta}) = \mathcal{W}(\boldsymbol{\theta}')$ for any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ that have the


Rating-1 users' continuation payoff
Payoffs of the fair action
$\left(\frac{s_0-1}{N-1}b-c, b-\frac{s_1-1}{N-1}c\right)$
(b-c, b-c)
Payoffs of the altruistic action
(0, 0)
Payoffs of the selfish action
Rating-0 users' continuation payoff
Rating-1 users' continuation payoff
Feasible payoffs under (N-1,1)
Feasible payoffs under (s_0,s_1)
Feasible payoffs under (1,N-1)
(0, 0)
Rating-0 users' continuation payoff

(a) Feasible payoffs in one state (rating distribution (s_0,s_1))

(b) Common feasible payoffs in all the states (rating distributions)

Figure 4.3: Illustration of how to build the self-generating set. The left figure shows the set of feasible payoffs in one state (i.e. under the rating distribution (s_0, s_1)). The right figure shows the sets of feasible payoffs in different states (i.e. rating distributions) and their intersection, namely the set of common feasible payoffs in all the states (i.e. under all the rating distributions).

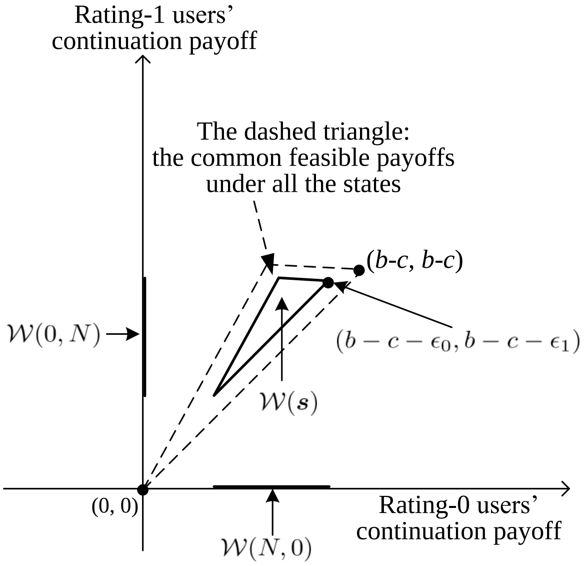

Rating-1 users' continuation payoff
The dashed triangle: the common feasible payoffs under all the states
(b-c, b-c)
$\mathcal{W}(0, N)$
$(b-c-\epsilon_0, b-c-\epsilon_1)$
$\mathcal{W}(\boldsymbol{s})$
(0, 0)
$\mathcal{W}(N, 0)$
Rating-0 users' continuation payoff

Figure 4.4: Illustration of the self-generating set, which is a triangle within the set of common feasible payoffs in Fig. 4.3.

same rating distribution. Hence, in the rest of the proof sketch, we write the self-generating set as $\{\mathcal{W}(\boldsymbol{s})\}_{\boldsymbol{s}}$, which is illustrated in Fig. 4.3 and Fig. 4.4. Fig. 4.3 shows how to construct the self-generating set. The left of Fig. 4.3 shows the feasible payoffs in one state, and the right shows the common feasible payoffs in all the states (we consider the common feasible payoffs such that we can use the same $\mathcal{W}(\boldsymbol{s})$ under all the states $\boldsymbol{s}$). The self-generating set is a subset of the common feasible payoffs, as illustrated in Fig. 4.4. When the users have different ratings (i.e. $1 \leq s_0 \leq N-1$), the set $\mathcal{W}(\boldsymbol{s})$ is the triangle shown in Fig. 4.4, which is congruent to the triangle of common feasible payoffs (shown in dashed lines), and has the upper right vertex at $(b-c-\epsilon_0, b-c-\epsilon_1)$ with $\epsilon_0, \epsilon_1 \leq \xi$. We have the analytical expression for the triangle in Appendix 4.8.3. When all the users have the same rating (i.e. $s_0 = 0$ or $s_0 = N$), only one component in $\boldsymbol{v}(\boldsymbol{s})$ is relevant. Hence, the sets $\mathcal{W}((N,0))$ and $\mathcal{W}((0,N))$ are line segments determined by the ranges of v_0 and v_1 in the triangle, respectively.

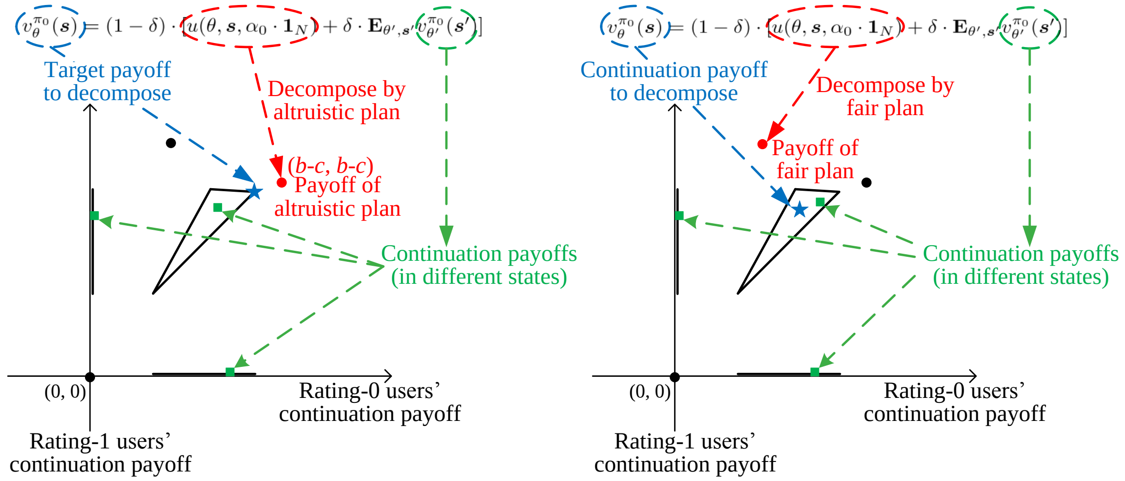
In addition, we simplify the decomposition (4.12) and (4.13) by letting the continuation payoffs $v_{\theta'}^{\pi_0}(\boldsymbol{s}') = v_{\theta'}^{\pi_0}$ for all $\boldsymbol{s}'$. Hence, for a given $\boldsymbol{s}$ and a payoff vector $\boldsymbol{v}(\boldsymbol{s})$, the continuation payoffs $\boldsymbol{v}' = (v_0', v_1')$ can be determined by solving the following two linear equations:

$$\begin{cases} v_0(\boldsymbol{s}) = (1-\delta)u\left(0, \boldsymbol{s}, \alpha_0 \mathbf{1}_N\right) + \delta \sum_{\theta'=0}^{1} q(\theta'|0, \alpha_0 \mathbf{1}_N) v_{\theta'}' \\ v_1(\boldsymbol{s}) = (1-\delta)u\left(1, \boldsymbol{s}, \alpha_0 \mathbf{1}_N\right) + \delta \sum_{\theta'=0}^{1} q(\theta'|1, \alpha_0 \mathbf{1}_N) v_{\theta'}' \end{cases} \tag{4.14}$$

where $q(\theta'|\theta, \alpha_0)$ is the probability that the next rating is θ' for a user with rating θ under the plan profile $\alpha_0 \cdot \mathbf{1}_N$.

Based on the above simplification, the collection of sets $\{\mathcal{W}(\boldsymbol{s})\}_{\boldsymbol{s}}$ in Fig. 4.4 is a self-generating set, if for any $\boldsymbol{s}$ and any payoff vector $\boldsymbol{v}(\boldsymbol{s}) \in \mathcal{W}(\boldsymbol{s})$, we can find a plan α_0 such that the continuation payoffs $\boldsymbol{v}'$ calculated from (4.14) lie in the triangle and satisfy the incentive constraints in (4.13).

In summary, we can prove that the collection of sets $\{\mathcal{W}(\boldsymbol{s})\}_{\boldsymbol{s}}$ illustrated in Fig. 4.4 is a self-generating set under certain conditions. Specifically, given a



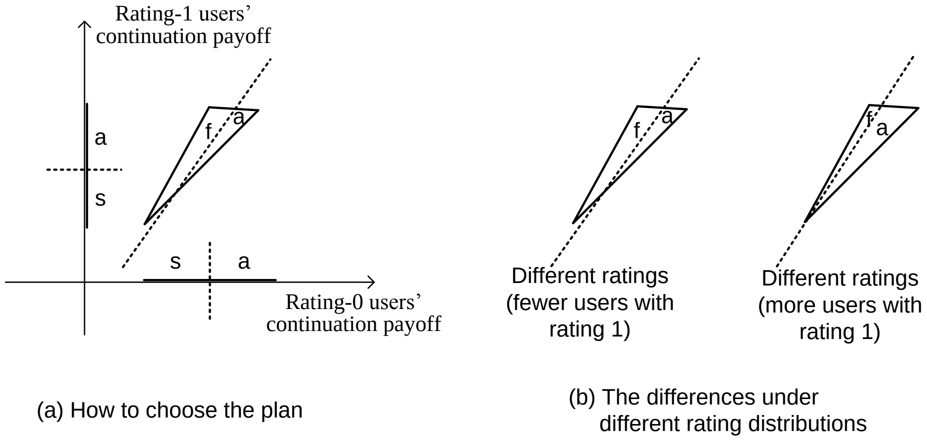
(a) Decomposition in period 0

(b) Decomposition in period 1
(when users have different
ratings in period 1)

Figure 4.5: The decomposition of payoffs. The left figure shows the decomposition in period 0, when the payoff to decompose is the target payoff $(b-c-\epsilon_0, b-c-\epsilon_1)$; the right figure shows the decomposition in period 1, when the payoff to decompose is the continuation payoff starting from period 1 and when the users have different ratings.

performance loss ξ, we construct the corresponding $\{\mathcal{W}(\boldsymbol{s})\}_{\boldsymbol{s}}$, and prove that it is a self-generating set under the following conditions: 1) the discount factor $\delta \geq \underline{\delta}(\varepsilon, \xi)$ with $\underline{\delta}(\varepsilon, \xi)$ defined in Appendix 4.8.2, and 2) the three conditions on the rating update rule in Theorem 7. This proves the first part of Theorem 7.

The corresponding recommended strategy can be determined based on the decomposition of payoffs. Specifically, given the current rating distribution $\boldsymbol{s}$ and the current expected payoffs $\boldsymbol{v}(\boldsymbol{s}) \in \mathcal{W}(\boldsymbol{s})$, we find a plan α_0 such that the continuation payoffs $\boldsymbol{v}'$ calculated from (4.14) lie in the triangle and satisfy the incentive constraints. The decomposition is illustrated in Fig. 4.5. One important issue in the decomposition is which plan should be used to decompose the payoff. We prove that we can determine the plan in the following way (illustrated in Fig. 4.6). When the users have different ratings, choose the altruistic plan α^{a}


Rating-1 users' continuation payoff
a
s
f
a
s
a
Rating-0 users' continuation payoff
f
a
Different ratings (fewer users with rating 1)
f
a
Different ratings (more users with rating 1)

(a) How to choose the plan

(b) The differences under different rating distributions

Figure 4.6: Illustration of how to choose the plan in order to decompose a given payoff. Each self-generating set is partitioned into two parts. In each period, a recommended plan (the altruistic plan "a", the fair plan "f", or the selfish plan "s") is chosen, depending on which part of the self-generating set the expected payoffs fall into.

when $\boldsymbol{v}(\boldsymbol{s})$ lies in the region marked by "a" in the triangle in Fig. 4.6-(b), and choose the fair plan α^{f} otherwise. When the users have the same rating 0 or 1, choose the altruistic plan α^{a} when $v_0(\boldsymbol{s})$ or $v_1(\boldsymbol{s})$ lies in the region marked by "a" in the line segment in Fig. 4.6-(a) or Fig. 4.6-(c), and choose the selfish plan α^{s} otherwise. Note that we can analytically determine the line that separates the two regions in the triangle and the threshold that separates the two regions in the line segment (analytical expressions are omitted due to space limitation; see Appendix 4.8.4 for details). The above decomposition is repeated, and is used to determine the recommended plan in each period based on the current rating distribution $\boldsymbol{s}$ and the current expected payoffs to achieve $\boldsymbol{v}(\boldsymbol{s})$. The procedure described above is exactly the algorithm to construct the recommended strategy, which is described in Table 4.6. Due to space limitation, Table 4.6 is illustrative but not specific. The detailed table that describes the algorithm can be found in Appendix 4.8.4. □

Theorem 7 proves that for any rating update error $\varepsilon \in [0, 0.5)$, we can design an asymptotically optimal rating mechanism. The design of the asymptotically optimal rating mechanism consists of two parts. The first part is to design the rating update rule. First, we should give incentives for the users to provide high-quality service, by setting β_θ^+, the probability that the rating goes up when the service quality is not lower than the recommended quality, to be larger than $1 - \beta_\theta^-$, the probability that the rating goes up when the service quality is lower than the recommended quality. Second, for a user with rating 1, the expected probability that its rating goes up when it complies should be larger than the threshold specified in Condition 2 ($x_{s_1}^+ > x_1^+$ implies that $x_{s_1}^+$ is larger than the threshold, too). This gives users incentives to obtain rating 1. Meanwhile, for a user with rating 0, the expected probability that its rating goes up when it complies, x_0^+, should be smaller than the threshold specified in Condition 3. This provides necessary punishment for a user with rating 0. Note that Conditions 2 and 3 imply that $x_1^+ > x_0^+$. In this way, a user will prefer to have rating 1.

The second part is to construct the equilibrium recommended strategy. Theorem 7 proves that for any feasible discount factor δ no smaller than the lower-bound discount factor $\underline{\delta}(\varepsilon, \xi)$ defined in Appendix 4.8.2, we can construct the corresponding recommended strategy such that each user can achieve an discounted average payoff of at least $b - c - \xi$. Now we show how to construct the recommended strategy. Note that determining the lower-bound discount factor $\underline{\delta}(\varepsilon, \xi)$ analytically is important for constructing the equilibrium $(\pi_0, \pi_0 \cdot \mathbf{1}_N)$, because a feasible discount factor is needed to determine the strategy. In [9] and [81], the lower bound for the discount factor cannot be obtained analytically. Hence, their results are not constructive.

The algorithm in Table 4.6 that constructs the optimal recommended strategy works as follows. In each period, the algorithm updates the continuation payoffs (v_0, v_1), and determines the recommended plan based on the current rating dis-

Table 4.6: Algorithm to construct recommended strategies.

```
Require: $b$, $c$, $\varepsilon$, $\xi$; $\tau(\varepsilon)$, $\delta \geq \underline{\delta}(\varepsilon, \xi)$; $\boldsymbol{\theta}^0$
Initialization: $t = 0$, $\epsilon_0 = \xi$, $\epsilon_1 = \epsilon_0/(1 + \frac{\kappa_2}{\kappa_1})$, $v^\theta = b - c - \epsilon_\theta$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.
repeat
    if $s_1(\boldsymbol{\theta}) = 0$ then
        if $(v^0, v^1)$ lies in region "a" of the horizontal line segment in Fig. 4.6-(a)
            choose recommended plan $\alpha^{\mathrm{a}}$
        else
            choose recommended plan $\alpha^{\mathrm{s}}$
        end
    elseif $s_1(\boldsymbol{\theta}) = N$ then
        if $(v^0, v^1)$ lies in region "a" of the vertical line segment in Fig. 4.6-(c)
            choose recommended plan $\alpha^{\mathrm{a}}$
        else
            choose recommended plan $\alpha^{\mathrm{s}}$
        end
    else
        if $(v^0, v^1)$ lies in region "a" of the triangle in Fig. 4.6-(b)
            choose recommended plan $\alpha^{\mathrm{a}}$
        else
            choose recommended plan $\alpha^{\mathrm{f}}$
        end
    end
    determine the continuation payoffs $(v_0', v_1')$ according to (4.14)
    $t \leftarrow t + 1$, determine the rating profile $\boldsymbol{\theta}^t$, set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^t$, $(v_0, v_1) \leftarrow (v_0', v_1')$
until $\varnothing$
```

tribution and the continuation payoffs. In Fig. 4.6, we illustrate which plan to recommend based on where the continuation payoffs locate in the self-generating sets. Specifically, each set $\mathcal{W}(\boldsymbol{s})$ is partitioned into two parts (the partition lines can determined analytically; see Appendix 4.8.4 for the analytical expressions). When all the users have rating 0 (or 1), we recommend the altruistic plan α^{a} if the continuation payoff v_0 (or v_1) is large, and the selfish plan α^{s} otherwise. When the users have different ratings, we recommend the altruistic plan α^{a} when (v_0, v_1) lies in the region marked by "a" in the triangle in Fig. 4.6-(b), and the fair plan α^{f} otherwise. Note that the partition of $\mathcal{W}(\boldsymbol{s})$ is different under different rating distributions (e.g., the region in which the altruistic plan is chosen is larger when more users have rating 1). Fig. 4.6 also illustrates why the strategy is nonstationary: the recommended plan depends on not only the current rating distribution $\boldsymbol{s}$, but also which region of $\mathcal{V}(\boldsymbol{s})$ the continuation payoffs (v_0, v_1) lie in.

Complexity: Although the design of the optimal recommended strategy is complicated, the implementation is simple. The computational complexity in each period comes from 1) identifying which region the continuation payoffs lie in, which is simple because the regions are divided by a straight line that is analytically determined, and 2) updating the continuation payoffs (v_0, v_1) by (4.14), which can be easily done by solving a set of two linear equations with two variables. The memory complexity is also low: because we summarize the history of past states by the continuation payoffs (v_0, v_1), the protocol does not need to store all the past states.

4.5.3 Whitewashing-Proofness

An important issue in rating mechanisms is whitewashing, namely users with low ratings can register as a new user to clear its history of bad behaviors. We say a rating mechanism is whitewashing-proof, if the cost of whitewashing (e.g. creating a new account) is higher than the benefit from whitewashing. The benefit from

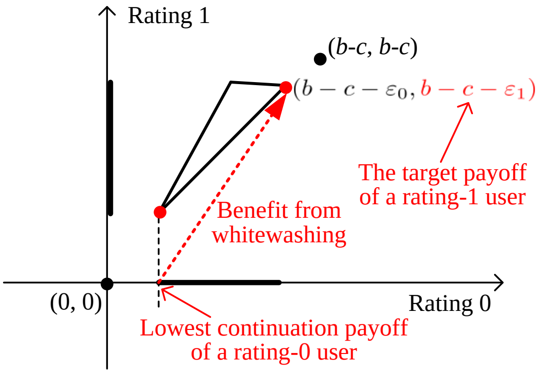


Figure 4.7: Illustration of the target payoff of a rating-1 user and the lowest continuation payoff of a rating-0 user.

whitewashing is determined by the difference between the current continuation payoff of a low-rating user and the target payoff of a high-rating user. Since this difference is relatively small under the proposed rating mechanism, the proposed mechanism is robust to whitewashing.

Proposition 5 *Given the performance loss tolerance $\xi > 0$, the proposed rating mechanism is whitewashing-proof if the cost of whitewashing is larger than $\left(1 - \frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right) \cdot \xi$.*

Proof: We illustrate the proof using Fig. 4.7. In Fig. 4.7, we show the self-generating set again, and point out the target payoff of a rating-1 user and the lowest continuation payoff of a rating-0 user. The difference between these two payoffs is the highest benefit that a rating-0 user can get by whitewashing. Simple calculation tells us that the difference is $\left(1 - \frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right) \cdot \xi$, which completes the proof of Proposition 5. □

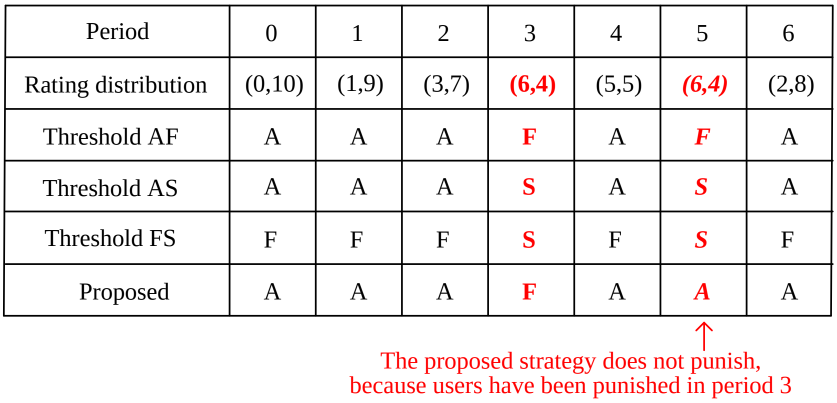
4.6 Simulation Results

We compare against the rating mechanism with threshold-based stationary recommended strategies. In particular, we focus on threshold-based stationary recommended strategies that use two plans. In other words, one plan is recommended when the number of rating-1 users is no smaller than the threshold, and the other plan is recommended otherwise. In particular, we consider threshold-based stationary recommended strategies restricted on $A^{\mathrm{a}f}$, $A^{\mathrm{a}s}$, and $A^{\mathrm{f}s}$, and call them “Threshold AF”, “Threshold AS”, and “Threshold FS”, respectively. We focus on threshold-based strategies because it is difficult to find the optimal stationary strategy in general when the number of users is large (the number of stationary strategies grows exponentially with the number of users). In our experiments, we fix the following system parameters: $N = 10, b = 3, c = 1$.

In Fig. 4.8, we first illustrate the evolution of the states and the recommended plans taken under the proposed rating mechanism and the rating mechanism with the Threshold AF strategy. The threshold is set to be 5. Hence, it recommends the altruistic plan when at least half of the users have rating 1, and recommends the fair plan otherwise. We can see that in the proposed strategy, the plans taken can be different at the same state. In particular, in “bad” states (6,4) and (7,3) at time slot 3 and 5, respectively, the proposed rating mechanism may recommend the fair plan (as a punishment) and the altruistic plan (i.e. do not punish because the punishment happens in time slot 3), while the stationary mechanism always recommends the fair plan to punish the low-rating users.

Then in Fig. 4.9, we show the price of stationarity of three representative stationary rating mechanisms: the one with the optimal Threshold AF strategy, the one with the optimal Threshold AS strategy, and the one with the optimal Threshold FS strategy. We can see from Fig. 4.9 that as the rating update error increases, the efficiency of stationary rating mechanisms decreases, and drops to 0 when the error probability is large (e.g. when $\varepsilon > 0.4$). In contrast, the proposed rating mechanism can achieve arbitrarily close to the social optimum.

In Fig. 4.10, we illustrate the lower-bound discount factors under different performance loss tolerances and rating update errors. As expected, when the per-



Period	0	1	2	3	4	5	6
Rating distribution	(0,10)	(1,9)	(3,7)	**(6,4)**	(5,5)	***(6,4)***	(2,8)
Threshold AF	A	A	A	**F**	A	***F***	A
Threshold AS	A	A	A	**S**	A	***S***	A
Threshold FS	F	F	F	**S**	F	***S***	F
Proposed	A	A	A	**F**	A	***A***	A

↑
The proposed strategy does not punish,
because users have been punished in period 3

Figure 4.8: Evolution of states and recommended plans taken in different rating mechanisms.

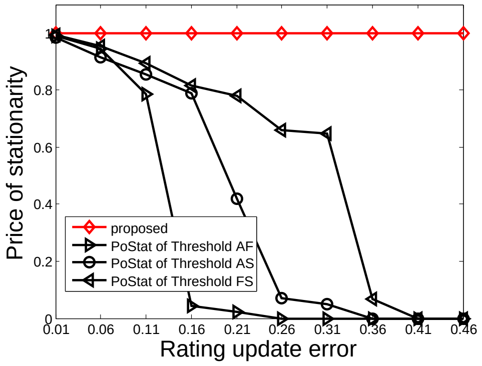


Figure 4.9: Price of stationarity of different stationary rating mechanisms under different rating update errors.

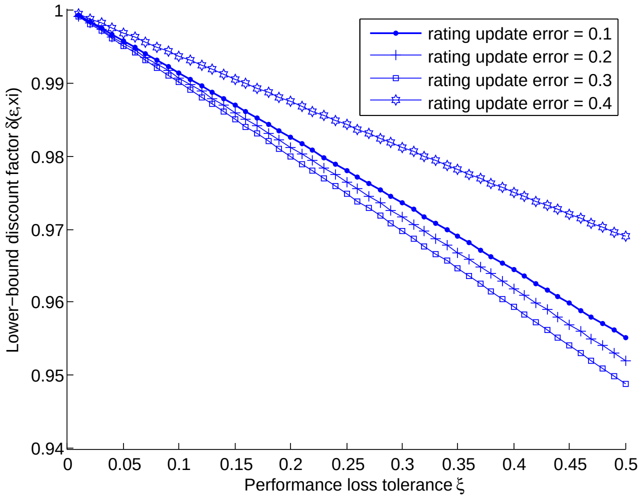


Figure 4.10: Lower-bound discount factors under different performance loss tolerances and rating update errors.

formance loss tolerance becomes larger, the lower-bound discount factor becomes smaller. What is unexpected is how the lower-bound discount factor changes with the rating update error. Specifically, the lower-bound discount factor decreases initially with the increase of the error, and then increases with the error. It is intuitive to see the discount factor increases with the rating update error, because the users need to be more patient when the rating update is more erroneous. The initial decrease of the discount factor in the error can be explained as follows. If the rating update error is extremely small, the punishment for the rating-0 users in the optimal rating update rule needs to be very severe (i.e. a smaller β_0^+ and a larger β_0^-). Hence, once a user is assigned with rating 0, it needs to be more patient to carry out the severe punishment (i.e. weigh the future payoffs more).

Finally, we illustrate the robustness of the proposed mechanisms with respect to the estimation of rating update errors. Suppose that the rating update error is ε. However, the designer cannot accurately measure this error. Under the estimated error $\hat{\varepsilon}$, the rating mechanism will construct another recommended strategy. In Fig. 4.11, we illustrate the performance gain/loss in terms of social welfare under

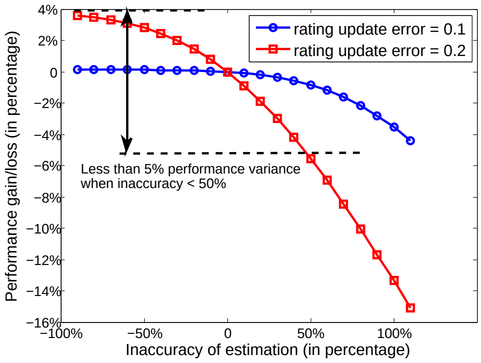


Figure 4.11: Performance gain/loss (in percentage) under different inaccuracy of estimation (in percentage).

the estimated error $\hat{\varepsilon}$, when the rating update error is ε. We can see that there is less than 5% performance variance when the estimation inaccuracy is less than 50%. The performance variance is larger when the rating update error is larger.

4.7 Conclusion

In this paper, we proposed a design framework for simple binary rating mechanisms that can achieve the social optimum in the presence of rating update errors. We provided design guidelines for the optimal rating update rules, and an algorithm to construct the optimal nonstationary recommended strategy. The key design principles that enable the rating mechanism to achieve the social optimum are the differential punishments, and the nonstationary strategies that reduce the performance loss while providing enough incentives. We also reduced the complexity of computing the recommended strategy by proving that using three recommended plans is enough to achieve the social optimum. The proposed rating mechanism is the first one that can achieve the social optimum even when the rating update errors are large. Simulation results demonstrated the significant performance gain of the proposed rating mechanism over the state-of-the-art mechanisms, especially when the rating update error is large.

4.8 Appendix

4.8.1 Proof of Proposition 4

4.8.1.1 The Claim to Prove

In order to prove Proposition 4, we quantify the performance loss of strategies restricted to $A^{\mathrm{a}s}$. The performance loss is determined in the following claim:

Claim: Starting from any initial rating profile $\boldsymbol{\theta}$, the maximum social welfare achievable at the PAE by $(\pi_0, \pi \cdot \mathbf{1}_N) \in \Pi(A^{\mathrm{a}s}) \times \Pi^N(A^{\mathrm{a}s})$ is at most

$$b - c - c \cdot \rho(\boldsymbol{\theta}, \alpha_0^*, S_B^*) \sum_{\boldsymbol{s}' \in S_B^*} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0^*, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \tag{4.15}$$

where α_0^*, the optimal recommended plan, and S_B^*, the optimal subset of rating distributions, are the solutions to the following optimization problem:

$$\begin{aligned} \min_{\alpha_0} \min_{S_B \subset S} \quad & \left\{ \rho(\boldsymbol{\theta}, \alpha_0, S_B) \textstyle\sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \right\} \\ s.t. & \textstyle\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) > \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \ \forall i, \\ & \textstyle\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) > \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \ \forall, \\ & \textstyle\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) > \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \ \forall i, \end{aligned} \tag{4.16}$$

where $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$ is defined as

$$\begin{aligned} & \rho(\boldsymbol{\theta}, \alpha_0, S_B) \triangleq \\ \max_{i \in \mathcal{N}} \max & \left\{ \frac{\frac{s_{\theta_i} - 1}{N-1}}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)}, \right. \\ & \frac{\frac{s_{1-\theta_i}}{N-1}}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)}, \\ & \left. \frac{1}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)} \right\}, \end{aligned} \tag{4.17}$$

where α^0 (resp. α^1) is the plan in which the user does not serve rating-0 (resp. rating-1) users, and α^{01} is the plan in which the user does not serve anyone.

The above claim shows that

$$W(\varepsilon, \delta, A^{\mathrm{a}s}) \leq b - c - c \cdot \rho(\boldsymbol{\theta}, \alpha_0^*, S_B^*) \sum_{\boldsymbol{s}' \in S_B^*} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0^*, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)$$

for any δ. By defining

$$\zeta(\varepsilon) \triangleq c \cdot \rho(\boldsymbol{\theta}, \alpha_0^*, S_B^*) \sum_{\boldsymbol{s}' \in S_B^*} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0^*, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N),$$

we obtain the result in Proposition 1, namely $\lim_{\delta \to 1} W(\varepsilon, \delta, A^{\mathrm{as}}) \leq b - c - \zeta(\varepsilon)$. Note that $\zeta(\varepsilon)$ is indeed a function of the rating update error ε, because ε determines the state transition function $q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0^*, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)$, and thus affects $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$. Note also that $\zeta(\varepsilon)$ is independent of the discount factor δ.

In the expression of $\zeta(\varepsilon)$, $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$ represents the normalized benefit from deviation (normalized by $b - c$). The numerator of $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$ is the probability of a player matched to the type of clients whom it deviates to not serve. The higher this probability, the larger benefit from deviation a player can get. The denominator of $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$ is the difference between the two state transition probabilities when the player does and does not deviate, respectively. When the above two transition probabilities are closer, it is less likely to detect the deviation, which results in a larger $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$. Hence, we can expect that a larger $\rho(\boldsymbol{\theta}, \alpha_0, S_B)$ (i.e. a larger benefit from deviation) will result in a larger performance loss, which is indeed true as will be proved later.

We can also see that $\zeta(\varepsilon) > 0$ as long as $\varepsilon > 0$. The reason is as follows. Suppose that $\varepsilon > 0$. First, from (4.17), we know that $\rho(\boldsymbol{\theta}, \alpha_0, S_B) > 0$ for any $\boldsymbol{\theta}$, α, and $S_B \neq \emptyset$. Second, we can see that $\sum_{\boldsymbol{s}' \in S_B^*} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0^*, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) > 0$ as long as $S_B^* \neq \emptyset$. Since $S_B^* = \emptyset$ cannot be the solution to the optimization problem (4.16) (because $S_B^* = \emptyset$ violates the constraints), we know that $\zeta(\varepsilon) > 0$.

4.8.1.2 Proof of the Claim

We prove that for any self-generating set $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta^N}$, the maximum payoff in $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta^N}$, namely $\max_{\boldsymbol{\theta}\in\Theta^N}\max_{\boldsymbol{v}\in\mathcal{W}^{\boldsymbol{\theta}}}\max_{i\in\mathcal{N}} v_i$, is bounded away from the social optimum $b-c$, regardless of the discount factor. In this way, we can prove that any equilibrium payoff is bounded away from the social optimum. In addition, we analytically quantify the efficiency loss, which is independent of the discount factor.

Since the strategies are restricted on the subset of plans A^{as}, in each period, all the users will receive the same stage-game payoff, either $(b-c)$ or 0, regardless of the matching rule and the rating profile. Hence, the expected discounted average payoff for each user is the same. More precisely, at any given history $\boldsymbol{h}^t = (\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^t)$, we have

$$U_i(\boldsymbol{\theta}^t, \pi_0|_{\boldsymbol{h}^t}, \pi|_{\boldsymbol{h}^t}\cdot\mathbf{1}_N) = U_j(\boldsymbol{\theta}^t, \pi_0|_{\boldsymbol{h}^t}, \pi|_{\boldsymbol{h}^t}\cdot\mathbf{1}_N),\ \forall i,j\in\mathcal{N}, \tag{4.18}$$

for any $(\pi_0, \pi\cdot\mathbf{1}_N)\in\Pi(A^{\mathrm{as}})\times\Pi^N(A^{\mathrm{as}})$. As a result, when we restrict to the plan set A^{as}, the self-generating set $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta^N}$ satisfies for any $\boldsymbol{\theta}$ and any $\boldsymbol{v}\in\mathcal{W}^{\boldsymbol{\theta}}$

$$v_i = v_j,\ \forall i,j\in\mathcal{N}. \tag{4.19}$$

Given any self-generating set $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta^N}$, define the maximum payoff $\bar{v}$ as

$$\bar{v} \triangleq \max_{\boldsymbol{\theta}\in\Theta^N}\max_{\boldsymbol{v}\in\mathcal{W}^{\boldsymbol{\theta}}}\max_{i\in\mathcal{N}} v_i. \tag{4.20}$$

Now we derive the upper bound of $\bar{v}$ by looking at the decomposability constraints.

To decompose the payoff profile $\bar{v}\cdot\mathbf{1}_N$, we must find a recommended plan $\alpha_0\in A^{\mathrm{as}}$, a plan profile $\alpha\cdot\mathbf{1}_N$ with $\alpha\in A^{\mathrm{as}}$, and a continuation payoff function $\boldsymbol{\gamma}:\Theta^N\to\cup_{\boldsymbol{\theta}'\in\Theta^N}\mathcal{W}^{\boldsymbol{\theta}'}$ with $\boldsymbol{\gamma}(\boldsymbol{\theta}')\in\mathcal{W}^{\boldsymbol{\theta}'}$, such that for all $i\in\mathcal{N}$ and for all $\alpha_i\in A$,

$$\begin{aligned}\bar{v} &= (1-\delta)u_i(\boldsymbol{\theta},\alpha_0,\alpha\cdot\mathbf{1}_N) + \delta\sum_{\boldsymbol{\theta}'}\gamma_i(\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha_0,\alpha\cdot\mathbf{1}_N) \qquad (4.21)\\ &\geq (1-\delta)u_i(\boldsymbol{\theta},\alpha_0,\alpha_i,\alpha\cdot\mathbf{1}_{N-1}) + \delta\sum_{\boldsymbol{\theta}'}\gamma_i(\boldsymbol{\theta}')q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha_0,\alpha_i,\alpha\cdot\mathbf{1}_{N-1}).\end{aligned}$$

Note that we do *not* require the users' plan α to be the same as the recommended plan α_0, and that we also do *not* require the continuation payoff function $\boldsymbol{\gamma}$ to be a simple continuation payoff function.

First, the payoff profile $\bar{v} \cdot \mathbf{1}_N$ cannot be decomposed by a recommended plan α_0 and the selfish plan α^{s}. Otherwise, since $\boldsymbol{\gamma}(\boldsymbol{\theta}') \in \mathcal{W}^{\boldsymbol{\theta}'}$, we have

$$\begin{aligned} \bar{v} &= (1-\delta) \cdot 0 + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \\ &\leq \delta \sum_{\boldsymbol{\theta}'} \bar{v}_i \cdot q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \\ &= \delta \cdot \bar{v} < \bar{v}, \end{aligned}$$

which is a contradiction.

Since we must use a recommended plan α_0 and the altruistic plan α^{a} to decompose $\bar{v} \cdot \mathbf{1}_N$, we can rewrite the decomposability constraint as

$$\begin{aligned} \bar{v} &= (1-\delta)(b-c) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \qquad (4.22) \\ &\geq (1-\delta) u_i(\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}). \end{aligned}$$

Since the continuation payoffs under different rating profiles $\boldsymbol{\theta}, \boldsymbol{\theta}'$ that have the same rating distribution $\boldsymbol{s}(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta}')$ are the same, namely $\boldsymbol{\gamma}(\boldsymbol{\theta}) = \boldsymbol{\gamma}(\boldsymbol{\theta}')$, the continuation payoff depends only on the rating distribution. For notational simplicity, with some abuse of notation, we write $\boldsymbol{\gamma}(\boldsymbol{s})$ as the continuation payoff when the rating distribution is $\boldsymbol{s}$, write $q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1})$ as the probability that the next state has a rating distribution $\boldsymbol{s}'$, and write $u_i(\boldsymbol{s}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1})$ as the stage-game payoff when the next state has a rating distribution $\boldsymbol{s}$. Then the decomposability constraint can be rewritten as

$$\begin{aligned} \bar{v} &= (1-\delta)(b-c) + \delta \sum_{\boldsymbol{s}'} \gamma_i(\boldsymbol{s}') q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \qquad (4.23) \\ &\geq (1-\delta) u_i(\boldsymbol{s}, \alpha_0, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) + \delta \sum_{\boldsymbol{s}'} \gamma_i(\boldsymbol{s}') q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}). \end{aligned}$$

Now we focus on a subclass of continuation payoff functions, and derive the maximum payoff $\bar{v}$ achievable under this subclass of continuation payoff functions. Later, we will prove that we cannot increase $\bar{v}$ by choosing other continuation payoff functions. Specifically, we focus on a subclass of continuation payoff functions that satisfy

$$\begin{aligned} \gamma_i(\boldsymbol{s}) &= x_A, \ \forall i \in \mathcal{N}, \ \forall \boldsymbol{s} \in S_A \subset S, \qquad (4.24) \\ \gamma_i(\boldsymbol{s}) &= x_B, \ \forall i \in \mathcal{N}, \ \forall \boldsymbol{s} \in S_B \subset S, \qquad (4.25) \end{aligned}$$

where S_A and S_B are subsets of the set of rating distributions S that have no intersection, namely $S_A \cap S_B = \emptyset$. In other words, we assign the two continuation payoff values to two subsets of rating distributions, respectively. Without loss of generality, we assume $x_A \geq x_B$.

Now we derive the incentive compatibility constraints. There are three plans to deviate to, the plan α^0 in which the user does not serve users with rating 0, the plan α^1 in which the user does not serve users with rating 1, and the plan α^{01} in which the user does not serve anyone. The corresponding incentive compatibility constraints for a user i with rating $\theta_i = 1$ are

$$\begin{aligned} \left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \right] (x_A - x_B) &\geq \frac{1-\delta}{\delta} \frac{s_0}{N-1} c, \\ \left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \right] (x_A - x_B) &\geq \frac{1-\delta}{\delta} \frac{s_1 - 1}{N-1} c, \\ \left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \right] (x_A - x_B) &\geq \frac{1-\delta}{\delta} c. \qquad (4.26) \end{aligned}$$

Similarly, the corresponding incentive compatibility constraints for a user j

with rating $\theta_j = 0$ are

$$\begin{aligned}
\left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_j = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right](x_A - x_B) &\geq \frac{1-\delta}{\delta}\frac{s_0 - 1}{N-1}c, \\
\left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_j = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right](x_A - x_B) &\geq \frac{1-\delta}{\delta}\frac{s_1}{N-1}c, \\
\left[\sum_{\boldsymbol{s}' \in S_A} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_j = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right](x_A - x_B) &\geq \frac{1-\delta}{\delta}c. \quad (4.27)
\end{aligned}$$

We can summarize the above incentive compatibility constraints as

$$x_A - x_B \geq \frac{1-\delta}{\delta}c \cdot \rho(\boldsymbol{\theta}, \alpha_0, S_A), \tag{4.28}$$

where

$$\begin{aligned}
&\rho(\boldsymbol{\theta}, \alpha_0, S_B) \triangleq \\
\max\nolimits_{i \in \mathcal{N}} \max \quad &\left\{ \frac{\frac{s_{\theta_i} - 1}{N-1}}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)}, \right. \\
&\frac{\frac{s_{1-\theta_i}}{N-1}}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)}, \\
&\left. \frac{1}{\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) - q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)} \right\}.
\end{aligned}$$

Since the maximum payoff $\bar{v}$ satisfies

$$\bar{v} = (1-\delta)(b-c) + \delta\left(x_A \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) + x_B \sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right),$$

to maximize $\bar{v}$, we choose $x_B = x_A - \frac{1-\delta}{\delta}c \cdot \rho(\boldsymbol{\theta}, \alpha_0, S_B)$. Since $x_A \geq \bar{v}$, we have

$$\begin{aligned}
\bar{v} &= (1-\delta)(b-c) + \delta\left(x_A - \frac{1-\delta}{\delta}c \cdot \rho(\boldsymbol{\theta}, \alpha_0, S_B) \sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right) \\
&\leq (1-\delta)(b-c) + \delta\left(\bar{v} - \frac{1-\delta}{\delta}c \cdot \rho(\boldsymbol{\theta}, \alpha_0, S_B) \sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)\right),
\end{aligned}$$

which leads to

$$\bar{v} \leq b - c - c \cdot \rho(\boldsymbol{\theta}, \alpha_0, S_B) \sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}'|\boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N). \tag{4.29}$$

Hence, the maximum payoff $\bar{v}$ satisfies

$$\bar{v} \leq b - c - c \cdot \min_{S_B \subset S} \left\{ \rho(\boldsymbol{\theta}, \alpha_0, S_B) \sum_{\boldsymbol{s}' \in S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \right\}, \tag{4.30}$$

where S_B satisfies for all $i \in \mathcal{N}$,

$$\begin{aligned}
\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) &> \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \\
\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) &> \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^1, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N), \\
\sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) &> \sum_{\boldsymbol{s}' \in S \setminus S_B} q(\boldsymbol{s}' | \boldsymbol{\theta}, \alpha_0, \alpha_i = \alpha^{01}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N). \qquad (4.31)
\end{aligned}$$

Following the same logic as in the proof of Proposition 6 in [97], we can prove that we cannot achieve a higher maximum payoff by other continuation payoff functions.

4.8.2 Analytical Expression of The Lower-Bound Discount Factor

The lower-bound discount factor $\underline{\delta}(\varepsilon, \xi)$ is the maximum of three critical discount factors, namely $\underline{\delta}(\varepsilon, \xi) \triangleq \max\{\delta_1(\varepsilon, \xi), \delta_2(\varepsilon, \xi), \delta_3(\varepsilon, \xi)\}$, where

$$\delta_1(\varepsilon, \xi) \triangleq \max_{\theta \in \{0,1\}} \frac{c}{c + (1 - 2\varepsilon)(\beta_\theta^+ - (1 - \beta_\theta^-))(\xi \frac{\kappa_2}{\kappa_1 + \kappa_2})},$$

$$\delta_2(\varepsilon, \xi) \triangleq \max_{s_1 \in \{1, \ldots, N-1\} : \frac{s_1}{N-1} b + \frac{N - s_1}{N-1} c > \xi \frac{\kappa_2}{\kappa_1 + \kappa_2}} \left\{ \frac{\xi \frac{\kappa_2}{\kappa_1 + \kappa_2} - \left(\frac{s_1}{N-1} b + \frac{N - s_1}{N-1} c \right)}{\left(\xi \frac{\kappa_2}{\kappa_1 + \kappa_2}\right) \left(x_{s_1}^+ - x_0^+ \right) - \left(\frac{s_1}{N-1} b + \frac{N - s_1}{N-1} c \right)} \right\},$$

and

$$\delta_3(\varepsilon, \xi) \triangleq \max_{\theta \in \{0,1\}} \frac{b - c + c \frac{x_\theta^+}{(1 - 2\varepsilon)\left[\beta_\theta^+ - (1 - \beta_\theta^-)\right]}}{b - c + \frac{c \cdot x_\theta^+}{(1 - 2\varepsilon)\left[\beta_\theta^+ - (1 - \beta_\theta^-)\right]} - \frac{(1 + \kappa_1)(\xi \frac{\kappa_2}{\kappa_1 + \kappa_2}) - z_2}{\kappa_1} - z_3}, \tag{4.32}$$

where $z_2 \triangleq -\kappa_1(b - c) + \kappa_1(1 - 1/\kappa_2)\xi + \xi \frac{\kappa_1}{\kappa_2} / (1 + \frac{\kappa_2}{\kappa_1})$, and $z_3 \triangleq z_2 / (\kappa_1 + \kappa_2)$. Note that $\frac{(1 + \kappa_1)(\xi \frac{\kappa_2}{\kappa_1 + \kappa_2}) - z_2}{\kappa_1} + z_3 < 0$. We can see from the above expressions that $\delta_1(\varepsilon, \xi) < 1$ and $\delta_2(\varepsilon, \xi) < 1$ as long as $\xi > 0$. For $\delta_3(\varepsilon, \xi)$, simple calculations tell us that ξ appears in the denominator in the form of $-\frac{(2\kappa_1+\kappa_2)\kappa_2}{(\kappa_1+\kappa_2)^2\kappa_1} \cdot \xi$. Since $\kappa_1 > 0$ and $\kappa_2 > 0$, we know that $-\frac{(2\kappa_1+\kappa_2)\kappa_2}{(\kappa_1+\kappa_2)^2\kappa_1} < 0$. Hence, $\delta_3(\varepsilon, \xi)$ is increasing in ξ. As a result, $\delta_3(\varepsilon, \xi) < 1$ as long as ξ is small enough.

Note that all the critical discount factors can be calculated analytically. Specifically, $\delta_1(\varepsilon, \xi)$ and $\delta_3(\varepsilon, \xi)$ are the maximum of two analytically-computed numbers, and $\delta_2(\varepsilon, \xi)$ is the maximum of at most $N-1$ analytically-computed numbers.

4.8.3 Proof of Theorem 7

4.8.3.1 Outline of the proof

We derive the conditions under which the set $(\mathcal{W}^{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\Theta^N}$ is a self-generating set. Specifically, we derive the conditions under which any payoff profile $\boldsymbol{v} \in \mathcal{W}^{\boldsymbol{\theta}}$ is decomposable on $(\mathcal{W}^{\boldsymbol{\theta}'})_{\boldsymbol{\theta}'\in\Theta^N}$ given $\boldsymbol{\theta}$, for all $\boldsymbol{\theta} \in \Theta^N$.

4.8.3.2 When users have different ratings

4.8.3.3 Preliminaries

We first focus on the states $\boldsymbol{\theta}$ with $1 \leq s_1(\boldsymbol{\theta}) \leq N-1$, and derive the conditions under which any payoff profile $\boldsymbol{v} \in \mathcal{W}^{\boldsymbol{\theta}}$ can be decomposed by $(\alpha_0 = \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)$ or $(\alpha_0 = \alpha^{\mathrm{f}}, \alpha^{\mathrm{f}} \cdot \mathbf{1}_N)$. First, $\boldsymbol{v}$ could be decomposed by $(\alpha^{\mathrm{a}}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N)$, if there exists a continuation payoff function $\boldsymbol{\gamma} : \Theta^N \to \cup_{\boldsymbol{\theta}'\in\Theta^N} \mathcal{W}^{\boldsymbol{\theta}'}$ with $\boldsymbol{\gamma}(\boldsymbol{\theta}') \in \mathcal{W}^{\boldsymbol{\theta}'}$, such that for all $i \in \mathcal{N}$ and for all $\alpha_i \in A$,

$$\begin{aligned} v_i &= (1-\delta)u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}} \cdot \mathbf{1}_N) \qquad (4.33) \\ &\geq (1-\delta)u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}). \end{aligned}$$

Since we focus on simple continuation payoff functions, all the users with the same future rating will have the same continuation payoff regardless of the recommended plan α_0, the plan profile $(\alpha_i, \alpha \cdot \mathbf{1}_{N-1})$, and the future state $\boldsymbol{\theta}'$. Hence, we write the continuation payoffs for the users with future rating 1 and 0 as γ^1 and γ^0, respectively. Consequently, the above conditions on decomposability can be simplified to

$$\begin{aligned} v_i &= (1-\delta)\cdot u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N) \qquad (4.34)\\ &+ \delta\left(\gamma^1 \sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N) + \gamma^0 \sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N)\right)\\ &\geq (1-\delta)\cdot u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}}\cdot \mathbf{1}_{N-1})\\ &+ \delta\left(\gamma^1 \sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}}\cdot \mathbf{1}_{N-1}) + \gamma^0 \sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}}\cdot \mathbf{1}_{N-1})\right). \end{aligned}$$

First, consider the case when user i has rating 1 (i.e. $\theta_i = 1$). Based on (**??**), we can calculate the stage-game payoff as $u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N) = b - c$. The term $\sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N)$ is the probability that user i has rating 1 in the next state. Since user i's rating update is independent of the other users' rating update, we can calculate this probability as

$$\begin{aligned} \sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N) &= [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \sum_{m\in M:\theta_{m(i)}=1} \mu(m) (4.35)\\ &+ [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \sum_{m\in M:\theta_{m(i)}=0} \mu(m) (4.36)\\ &= (1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-) = x_1^+. \qquad (4.37) \end{aligned}$$

Similarly, we can calculate $\sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N)$, the probability that user i has rating 0 in the next state, as

$$\begin{aligned} \sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot \mathbf{1}_N) &= [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \sum_{m\in M:\theta_{m(i)}=1} \mu(m) (4.38)\\ &+ [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \sum_{m\in M:\theta_{m(i)}=0} \mu(m) (4.39)\\ &= (1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^- = 1 - x_1^+. \qquad (4.40) \end{aligned}$$

Now we discuss what happens if user i deviates. Since the recommended plan α^{a} is to exert high effort for all the users, user i can deviate to the other three plans, namely "exert high effort for rating-1 users only", "exert high effort for rating-0 users only", "exert low effort for all the users". We can calculate the corresponding stage-game payoff and state transition probabilities under each deviation.

- "exert high effort for rating-1 users only" ($\alpha_i(1,\theta_i)=1, \alpha_i(0,\theta_i)=0$):

$$u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) = b - c \cdot \sum_{m \in M : \theta_{m(i)}=1} \mu(m) = b - c \cdot \frac{s_1(\boldsymbol{\theta})-1}{N-1} \tag{4.41}$$

$$\begin{aligned} & \sum_{\boldsymbol{\theta}' : \theta'_i = 1} q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) \\ = & [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \sum_{m \in M : \theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \sum_{m \in M : \theta_{m(i)}=0} \mu(m) \\ = & [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \frac{s_1(\boldsymbol{\theta})-1}{N-1} + [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \frac{s_0(\boldsymbol{\theta})}{N-1}. \end{aligned} \tag{4.42}$$

$$\begin{aligned} & \sum_{\boldsymbol{\theta}' : \theta'_i = 0} q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) \\ = & [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \sum_{m \in M : \theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \sum_{m \in M : \theta_{m(i)}=0} \mu(m) \\ = & [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \frac{s_1(\boldsymbol{\theta})-1}{N-1} + [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \frac{s_0(\boldsymbol{\theta})}{N-1}. \end{aligned} \tag{4.43}$$

- "exert high effort for rating-0 users only" ($\alpha_i(1,\theta_i)=0, \alpha_i(0,\theta_i)=1$):

$$u_i(\boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) = b - c \cdot \sum_{m \in M : \theta_{m(i)}=0} \mu(m) = b - c \cdot \frac{s_0(\boldsymbol{\theta})}{N-1} \tag{4.44}$$

$$\begin{aligned} & \sum_{\boldsymbol{\theta}' : \theta'_i = 1} q(\boldsymbol{\theta}' | \boldsymbol{\theta}, \alpha^{\mathrm{a}}, \alpha_i, \alpha^{\mathrm{a}} \cdot \mathbf{1}_{N-1}) \\ = & [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \sum_{m \in M : \theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \sum_{m \in M : \theta_{m(i)}=0} \mu(m) \\ = & [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \frac{s_1(\boldsymbol{\theta})-1}{N-1} + [(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)] \frac{s_0(\boldsymbol{\theta})}{N-1}. \end{aligned} \tag{4.45}$$

$$\begin{aligned}
&\sum_{\boldsymbol{\theta}':\theta'_i=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{a}},\alpha_i,\alpha^{\mathrm{a}}\cdot\mathbf{1}_{N-1}) && (4.46)\\
= \;& [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \sum_{m\in M:\theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \sum_{m\in M:\theta_{m(i)}=0} \mu(m)\\
= \;& [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \frac{s_1(\boldsymbol{\theta})-1}{N-1} + [(1-\varepsilon)(1-\beta_1^+) + \varepsilon\beta_1^-] \frac{s_0(\boldsymbol{\theta})}{N-1}.
\end{aligned}$$

- "exert low effort for all the users" ($\alpha_i(1,\theta_i)=0, \alpha_i(0,\theta_i)=0$):

$$u_i(\boldsymbol{\theta},\alpha^{\mathrm{a}},\alpha_i,\alpha^{\mathrm{a}}\cdot\mathbf{1}_{N-1}) \;=\; b \qquad (4.47)$$

$$\begin{aligned}
&\sum_{\boldsymbol{\theta}':\theta'_i=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{a}},\alpha_i,\alpha^{\mathrm{a}}\cdot\mathbf{1}_{N-1}) && (4.48)\\
= \;& [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \sum_{m\in M:\theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+] \sum_{m\in M:\theta_{m(i)}=0} \mu(m)\\
= \;& (1-\varepsilon)(1-\beta_1^-) + \varepsilon\beta_1^+.
\end{aligned}$$

$$\begin{aligned}
&\sum_{\boldsymbol{\theta}':\theta'_i=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{a}},\alpha_i,\alpha^{\mathrm{a}}\cdot\mathbf{1}_{N-1}) && (4.49)\\
= \;& [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \sum_{m\in M:\theta_{m(i)}=1} \mu(m) + [(1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+)] \sum_{m\in M:\theta_{m(i)}=0} \mu(m)\\
= \;& (1-\varepsilon)\beta_1^- + \varepsilon(1-\beta_1^+).
\end{aligned}$$

Plugging the above expressions into (4.34), we can simplify the incentive compatibility constraints (i.e. the inequality constraints) to

$$(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c, \qquad (4.50)$$

under all three deviating plans.

Hence, if user i has rating 1, the decomposability constraints (4.34) reduces to

$$v^1 = (1-\delta)\cdot(b-c) + \delta\cdot\left[x_1^+\gamma^1 + (1-x_1^+)\gamma^0\right], \qquad (4.51)$$

where v^1 is the payoff of the users with rating 1, and

$$(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c. \qquad (4.52)$$

Similarly, if user i has rating 0, we can reduce the decomposability constraints (4.34) to

$$v^0 = (1-\delta)\cdot(b-c) + \delta\cdot\left[x_0^+\gamma^1 + (1-x_0^+)\gamma^0\right], \tag{4.53}$$

and

$$(1-2\varepsilon)\left[\beta_0^+ - (1-\beta_0^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c. \tag{4.54}$$

For the above incentive compatibility constraints (the above two inequalities) to hold, we need to have $\beta_1^+ - (1-\beta_1^-) > 0$ and $\beta_0^+ - (1-\beta_0^-) > 0$, which are part of Condition 1 and Condition 2. Now we will derive the rest of the sufficient conditions in Theorem 7.

The above two equalities determine the continuation payoff γ^1 and γ^0 as below

$$\left\{\begin{array}{l} \gamma^1 = \frac{1}{\delta}\cdot\frac{(1-x_0^+)v^1-(1-x_1^+)v^0}{x_1^+-x_0^+} - \frac{1-\delta}{\delta}\cdot(b-c) \\ \gamma^0 = \frac{1}{\delta}\cdot\frac{x_1^+v^0-x_0^+v^1}{x_1^+-x_0^+} - \frac{1-\delta}{\delta}\cdot(b-c) \end{array}\right. . \tag{4.55}$$

Now we consider the decomposability constraints if we want to decompose a payoff profile $\boldsymbol{v}\in\mathcal{W}^{\boldsymbol{\theta}}$ using the fair plan α^{f}. Since we focus on decomposition by simple continuation payoff functions, we write the decomposition constraints as

$$\begin{aligned} v_i &= (1-\delta)\cdot u_i(\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha^{\mathrm{f}}\cdot\mathbf{1}_N) \qquad (4.56)\\ &+ \delta\left(\gamma^1\sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha^{\mathrm{f}}\cdot\mathbf{1}_N) + \gamma^0\sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha^{\mathrm{f}}\cdot\mathbf{1}_N)\right)\\ &\geq (1-\delta)\cdot u_i(\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha_i,\alpha^{\mathrm{f}}\cdot\mathbf{1}_{N-1})\\ &+ \delta\left(\gamma^1\sum_{\boldsymbol{\theta}':\theta_i'=1} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha_i,\alpha^{\mathrm{f}}\cdot\mathbf{1}_{N-1}) + \gamma^0\sum_{\boldsymbol{\theta}':\theta_i'=0} q(\boldsymbol{\theta}'|\boldsymbol{\theta},\alpha^{\mathrm{f}},\alpha_i,\alpha^{\mathrm{f}}\cdot\mathbf{1}_{N-1})\right). \end{aligned}$$

Due to space limitation, we omit the details and directly give the simplification of the above decomposability constraints as follows. First, the incentive compatibility constraints (i.e. the inequality constraints) are simplified to

$$(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c, \tag{4.57}$$

and

$$(1-2\varepsilon)\left[\beta_0^+ - (1-\beta_0^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c, \tag{4.58}$$

under all three deviating plans. Note that the above incentive compatibility constraints are the same as the ones when we want to decompose the payoffs using the altruistic plan α^{a}.

Then, the equality constraints in (4.56) can be simplified as follows. For the users with rating 1, we have

$$v^1 = (1-\delta)\cdot\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) + \delta\cdot\left[x_{s_1(\boldsymbol{\theta})}^+\cdot\gamma^1 + (1-x_{s_1(\boldsymbol{\theta})}^+)\cdot\gamma^0\right], \tag{4.59}$$

where

$$x_{s_1(\boldsymbol{\theta})} \triangleq \left[(1-\varepsilon)\frac{s_1(\boldsymbol{\theta})-1}{N-1} + \frac{s_0(\boldsymbol{\theta})}{N-1}\right]\beta_1^+ + \left(\varepsilon\frac{s_1(\boldsymbol{\theta})-1}{N-1}\right)(1-\beta_1^-). \tag{4.60}$$

For the users with rating 0, we have

$$v^0 = (1-\delta)\cdot\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right) + \delta\cdot\left[x_0^+\gamma^1 + (1-x_0^+)\gamma^0\right]. \tag{4.61}$$

The above two equalities determine the continuation payoff γ^1 and γ^0 as below

$$\begin{cases} \gamma^1 = \frac{1}{\delta}\cdot\frac{(1-x_0^+)v^1-(1-x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} - \frac{1-\delta}{\delta}\cdot\frac{\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)(1-x_0^+)-\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)(1-x_{s_1(\boldsymbol{\theta})}^+)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} \\ \gamma^0 = \frac{1}{\delta}\cdot\frac{x_{s_1(\boldsymbol{\theta})}^+v^0-x_0^+v^1}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} - \frac{1-\delta}{\delta}\cdot\frac{\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)x_0^+-\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)x_{s_1(\boldsymbol{\theta})}^+}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} \end{cases} \tag{4.62}$$

4.8.3.4 Sufficient conditions

Now we derive the sufficient conditions under which any payoff profile $\boldsymbol{v} \in \mathcal{W}^{\boldsymbol{\theta}}$ can be decomposed by $(\alpha_0 = \alpha^{\mathrm{a}}, \alpha^{\mathrm{a}}\cdot\mathbf{1}_N)$ or $(\alpha_0 = \alpha^{\mathrm{f}}, \alpha^{\mathrm{f}}\cdot\mathbf{1}_N)$. Specifically, we will derive the conditions such that for any payoff profile $\boldsymbol{v} \in \mathcal{W}^{\boldsymbol{\theta}}$, at least one of the two decomposability constraints (4.34) and (4.56) is satisfied. From the preliminaries, we know that the incentive compatibility constraints in (4.34) and (4.56) can be simplified into the same constraints:

$$(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right](\gamma^1-\gamma^0) \geq \frac{1-\delta}{\delta}\cdot c, \tag{4.63}$$

and

$$(1-2\varepsilon)\left[\beta_0^+ - (1-\beta_0^-)\right](\gamma^1 - \gamma^0) \geq \frac{1-\delta}{\delta}\cdot c. \tag{4.64}$$

The above constraints impose the constraint on the discount factor, namely

$$\delta \geq \max_{\theta\in\Theta} \frac{c}{c + (1-2\varepsilon)\left[\beta_\theta^+ - (1-\beta_\theta^-)\right](\gamma^1-\gamma^0)}. \tag{4.65}$$

Since γ_1 and γ_0 should satisfy $\gamma^1 - \gamma^0 \geq \epsilon_0 - \epsilon_1$, the above constraints can be rewritten as

$$\delta \geq \max_{\theta\in\Theta} \frac{c}{c + (1-2\varepsilon)\left[\beta_\theta^+ - (1-\beta_\theta^-)\right](\epsilon_0-\epsilon_1)}, \tag{4.66}$$

where is part of Condition 3 in Theorem 7.

In addition, the continuation payoffs γ_1 and γ_0 should satisfy the constraints of the self-generating set, namely

$$\begin{aligned}
\gamma^1 - \gamma^0 &\geq \epsilon_0 - \epsilon_1, && (4.67)\\
\gamma^1 + \frac{c}{(N-1)b}\cdot\gamma^0 &\leq z_2 \triangleq (1+\frac{c}{(N-1)b})(b-c) - \frac{c}{(N-1)b}\epsilon_0 - \epsilon_1, && (4.68)\\
\gamma^1 - \frac{b}{\frac{N-2}{N-1}b-c}\cdot\gamma^0 &\leq z_3 \triangleq -\frac{\frac{b}{\frac{N-2}{N-1}b-c}-1}{1+\frac{c}{(N-1)b}}\cdot z_2. && (4.69)
\end{aligned}$$

We can plug the expressions of the continuation payoffs γ_1 and γ_0 in (4.55) and (4.62) into the above constraints. Specifically, if a payoff profile $\boldsymbol{v}$ is decomposed by the altruistic plan, the following constraints should be satisfied for the continuation payoff profile to be in the self-generating set: (for notational simplicity, we define $\kappa_1 \triangleq \frac{b}{\frac{N-2}{N-1}b-c} - 1$ and $\kappa_2 \triangleq 1 + \frac{c}{(N-1)b}$)

$$\frac{1}{\delta}\cdot\frac{v^1 - v^0}{x_1^+ - x_0^+} \geq \epsilon_0 - \epsilon_1, \tag{α^{a}-1}$$

$$\frac{1}{\delta}\cdot\left\{\frac{(1-\kappa_2 x_0^+)v^1 - (1-\kappa_2 x_1^+)v^0}{x_1^+ - x_0^+} - \kappa_2\cdot(b-c)\right\} \leq z_2 - \kappa_2\cdot(b-c), \tag{α^{a}-2}$$

$$\frac{1}{\delta}\cdot\left\{\frac{(1+\kappa_1 x_0^+)v^1 - (1+\kappa_1 x_1^+)v^0}{x_1^+ - x_0^+} + \kappa_1\cdot(b-c)\right\} \leq z_3 + \kappa_1\cdot(b-c). \tag{α^{a}-3}$$

The constraint (α^{a}-1) is satisfied for all v^1 and v^0 as long as $x_1^+ > x_0^+$, because $v^1 - v^0 > \epsilon_0 - \epsilon_1$, $|x_1^+ > x_0^+| < 1$, and $\delta < 1$.

Since both the left-hand side (LHS) and the right-hand side (RHS) of (α^{a}-2) are smaller than 0, we have

$$(\alpha^{\mathrm{a}}\text{-}2) \Leftrightarrow \delta \leq \frac{\frac{(1-\kappa_2 x_0^+)v^1-(1-\kappa_2 x_1^+)v^0}{x_1^+-x_0^+} - \kappa_2 \cdot (b-c)}{z_2 - \kappa_2 \cdot (b-c)} \tag{4.70}$$

The RHS of (α^{a}-3) is larger than 0. Hence, we have

$$(\alpha^{\mathrm{a}}\text{-}3) \Leftrightarrow \delta \geq \frac{\frac{(1+\kappa_1 x_0^+)v^1-(1+\kappa_1 x_1^+)v^0}{x_1^+-x_0^+} + \kappa_1 \cdot (b-c)}{z_3 + \kappa_1 \cdot (b-c)}. \tag{4.71}$$

If a payoff profile $\boldsymbol{v}$ is decomposed by the fair plan, the following constraints should be satisfied for the continuation payoff profile to be in the self-generating set:

$$\frac{1}{\delta} \cdot \left\{ \frac{v^1 - v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} - \frac{\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \right\} \geq \epsilon_0 - \epsilon_1 - \frac{\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}, \tag{α^{f}-1}$$

$$\frac{1}{\delta} \cdot \left\{ \frac{(1-\kappa_2 x_0^+)v^1 - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} - \frac{(1-\kappa_2 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \right.$$

$$\leq z_2 - \frac{(1-\kappa_2 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}, \tag{α^{f}-2}$$

$$\frac{1}{\delta} \cdot \left\{ \frac{(1+\kappa_1 x_0^+)v^1 - (1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} - \frac{(1+\kappa_1 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \right.$$

$$\leq z_3 - \frac{(1+\kappa_1 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}. \tag{α^{f}-3}$$

Since $\frac{v^1-v^0}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} > \epsilon_0 - \epsilon_1$, the constraint ($\alpha^{\mathrm{f}}$-1) is satisfied for all v^1 and v^0 if $v^1 - v^0 \geq \frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c$. Hence, the constraint (α^{f}-1) is equivalent to

$$\delta \geq \frac{v^1 - v^0 - \left(\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c\right)}{(\epsilon^0 - \epsilon^1)(x_{s_1(\boldsymbol{\theta})}^+ - x_0^+) - \left(\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c\right)},\ \textit{for } \boldsymbol{\theta} \textit{ s.t. } \frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c \geq v^1 - v^0 \tag{4.72}$$

For (α^{f}-2), we want to make the RHS have the same (minus) sign under any state $\boldsymbol{\theta}$, which is true if

$$1-\kappa_2 x_0^+ > 0,\ 1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+ < 0,\ \frac{1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+}{1-\kappa_2 x_0^+} \geq -(\kappa_2-1),\ s_1(\boldsymbol{\theta}) = 1,\ldots,N-1, \tag{4.73}$$

which leads to

$$x_{s_1(\boldsymbol{\theta})}^+ > \frac{1}{\kappa_2},\ x_0^+ < \frac{1}{\kappa_2},\ x_0^+ < \frac{1-x_{s_1(\boldsymbol{\theta})}^+}{1-\kappa_2},\ s_1(\boldsymbol{\theta}) = 1,\ldots,N-1, \tag{4.74}$$

$$\Leftrightarrow\ \frac{N-2}{N-1}x_1^+ + \frac{1}{N-1}\beta_1^+ > \frac{1}{\kappa_2},\ x_0^+ < \min\left\{\frac{1}{\kappa_2},\ \frac{1-\beta_1^+}{1-\kappa_2}\right\}. \tag{4.75}$$

Since the RHS of (α^{f}-2) is smaller than 0, we have

$$(\alpha^{\mathrm{f}}\text{-}2) \Leftrightarrow \delta \leq \frac{\frac{(1-\kappa_2 x_0^+)v^1-(1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} - \frac{(1-\kappa_2 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}}{z_2 - \frac{(1-\kappa_2 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}} \tag{4.76}$$

For (α^{f}-3), since $\frac{1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+}{1+\kappa_1 x_0^+} < 1+\kappa_1$, the RHS is always smaller than 0. Hence, we have

$$(\alpha^{\mathrm{f}}\text{-}3) \Leftrightarrow \delta \leq \frac{\frac{(1+\kappa_1 x_0^+)v^1-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} - \frac{(1+\kappa_1 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}}{z_3 - \frac{(1+\kappa_1 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}} \tag{4.77}$$

We briefly summarize what requirements on δ we have obtained now. To make the continuation payoff profile in the self-generating under the decomposition of α^{a}, we have one upper bound on δ resulting from (α^{a}-2) and one lower bound on δ resulting from (α^{a}-3). To make the continuation payoff profile in the self-generating under the decomposition of α^{f}, we have two upper bounds on δ resulting from (α^{f}-2) and (α^{f}-3), and one lower bound on δ resulting from (α^{f}-1). First, we want to eliminate the upper bounds, namely make the upper bounds larger than 1, such that δ can be arbitrarily close to 1.

To eliminate the following upper bound resulting from (α^{a}-2)

$$\delta \leq \frac{\frac{(1-\kappa_2 x_0^+)v^1-(1-\kappa_2 x_1^+)v^0}{x_1^+-x_0^+} - \kappa_2\cdot(b-c)}{z_2 - \kappa_2\cdot(b-c)}, \tag{4.78}$$

we need to have (since $z_2 - \kappa_2 \cdot (b-c) < 0$)

$$\frac{(1-\kappa_2 x_0^+)v^1 - (1-\kappa_2 x_1^+)v^0}{x_1^+ - x_0^+} \le z_2, \ \forall v^1, v^0. \tag{4.79}$$

The LHS of the above inequality is maximized when $v^0 = \frac{z_2 - z_3}{\kappa_1+\kappa_2}$ and $v^1 = v^0 + \frac{\kappa_1 z_2 + \kappa_2 z_3}{\kappa_1+\kappa_2}$. Hence, the above inequality is satisfied if

$$\frac{(1-\kappa_2 x_0^+)\left(\frac{z_2-z_3}{\kappa_1+\kappa_2} + \frac{\kappa_1 z_2+\kappa_2 z_3}{\kappa_1+\kappa_2}\right) - (1-\kappa_2 x_1^+)\frac{z_2-z_3}{\kappa_1+\kappa_2}}{x_1^+ - x_0^+} \le z_2 \tag{4.80}$$

$$\Leftrightarrow \quad \left(\frac{1-x_1^+ - x_0^+(\kappa_2-1)}{x_1^+ - x_0^+}\frac{\kappa_1}{\kappa_1+\kappa_2}\right) z_2 \le -\frac{1-x_1^+ - x_0^+(\kappa_2-1)}{x_1^+ - x_0^+}\frac{\kappa_2}{\kappa_1+\kappa_2} z_3. \tag{4.81}$$

Since $x_0^+ < \frac{1-\beta_1^+}{1-\kappa_2} < \frac{1-x_1^+}{1-\kappa_2}$, we have

$$z_2 \le -\frac{\kappa_2}{\kappa_1} z_3. \tag{4.82}$$

To eliminate the following upper bound resulting from (α^{f}-2)

$$\delta \le \frac{\frac{(1-\kappa_2 x_0^+)v^1 - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} - \frac{(1-\kappa_2 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}}{z_2 - \frac{(1-\kappa_2 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}}, \tag{4.83}$$

we need to have (since $z_2 - \frac{(1-\kappa_2 x_0^+)\left(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c\right)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} < 0$)

$$\frac{(1-\kappa_2 x_0^+)v^1 - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \le z_2, \ \forall v^1, v^0. \tag{4.84}$$

Similarly, the LHS of the above inequality is maximized when $v^0 = \frac{z_2 - z_3}{\kappa_1+\kappa_2}$ and $v^1 = v^0 + \frac{\kappa_1 z_2 + \kappa_2 z_3}{\kappa_1+\kappa_2}$. Hence, the above inequality is satisfied if

$$\frac{(1-\kappa_2 x_0^+)\left(\frac{z_2-z_3}{\kappa_1+\kappa_2} + \frac{\kappa_1 z_2+\kappa_2 z_3}{\kappa_1+\kappa_2}\right) - (1-\kappa_2 x_{s_1(\boldsymbol{\theta})}^+)\frac{z_2-z_3}{\kappa_1+\kappa_2}}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \le z_2 \tag{4.85}$$

$$\Leftrightarrow \quad \left(\frac{1-x_{s_1(\boldsymbol{\theta})}^+ - x_0^+(\kappa_2-1)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}\frac{\kappa_1}{\kappa_1+\kappa_2}\right) z_2 \le -\frac{1-x_{s_1(\boldsymbol{\theta})}^+ - x_0^+(\kappa_2-1)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}\frac{\kappa_2}{\kappa_1+\kappa_2} z_3, \tag{4.86}$$

Since $x_0^+ < \frac{1-\beta_1^+}{1-\kappa_2} < \frac{1-x_{s_1(\boldsymbol{\theta})}^+}{1-\kappa_2}$, we have

$$z_2 \le -\frac{\kappa_2}{\kappa_1} z_3. \tag{4.87}$$

To eliminate the following upper bound resulting from (α^{f}-3)

$$\delta \leq \frac{\frac{(1+\kappa_1 x_0^+)v^1-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} - \frac{(1+\kappa_1 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}}{z_3 - \frac{(1+\kappa_1 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+}}, \quad (4.88)$$

we need to have (since $z_3 - \frac{(1+\kappa_1 x_0^+)\left(b-\frac{s_1(\boldsymbol{\theta})-1}{N-1}c\right)-(1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\left(\frac{s_0(\boldsymbol{\theta})-1}{N-1}b-c\right)}{x_{s_1(\boldsymbol{\theta})}^+-x_0^+} < 0$)

$$\frac{(1+\kappa_1 x_0^+)v^1 - (1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \leq z_3, \ \forall v^1, v^0. \quad (4.89)$$

Again, the LHS of the above inequality is maximized when $v^0 = \frac{z_2-z_3}{\kappa_1+\kappa_2}$ and $v^1 = v^0 + \frac{\kappa_1 z_2+\kappa_2 z_3}{\kappa_1+\kappa_2}$. Hence, the above inequality is satisfied if

$$\frac{(1+\kappa_1 x_0^+)\left(\frac{z_2-z_3}{\kappa_1+\kappa_2} + \frac{\kappa_1 z_2+\kappa_2 z_3}{\kappa_1+\kappa_2}\right) - (1+\kappa_1 x_{s_1(\boldsymbol{\theta})}^+)\frac{z_2-z_3}{\kappa_1+\kappa_2}}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} \leq z_3 \quad (4.90)$$

$$\Leftrightarrow \ \left(\frac{1-x_{s_1(\boldsymbol{\theta})}^+ + x_0^+(\kappa_1+1)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}\frac{\kappa_1}{\kappa_1+\kappa_2}\right) z_2 \leq -\frac{1-x_{s_1(\boldsymbol{\theta})}^+ + x_0^+(\kappa_1+1)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}\frac{\kappa_2}{\kappa_1+\kappa_2} z_3 \quad (4.91)$$

Since $1 - x_{s_1(\boldsymbol{\theta})}^+ + x_0^+(\kappa_1+1) > 0$, we have

$$z_2 \leq -\frac{\kappa_2}{\kappa_1} z_3. \quad (4.92)$$

In summary, to eliminate the upper bounds on δ, we only need to have $z_2 \leq -\frac{\kappa_2}{\kappa_1}z_3$, which is satisfied since we define $z_3 \triangleq -\frac{\kappa_1}{\kappa_2}z_2$.

Now we derive the analytical lower bound on δ based on the lower bounds resulting from (α^{a}-3) and (α^{f}-1):

$$(\alpha^{\mathrm{a}}\text{-}3) \Leftrightarrow \delta \geq \frac{\frac{(1+\kappa_1 x_0^+)v^1-(1+\kappa_1 x_1^+)v^0}{x_1^+-x_0^+} + \kappa_1\cdot(b-c)}{z_3+\kappa_1\cdot(b-c)}, \quad (4.93)$$

and

$$\delta \geq \frac{v^1 - v^0 - \left(\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c\right)}{(\epsilon^0-\epsilon^1)(x_{s_1(\boldsymbol{\theta})}^+ - x_0^+) - \left(\frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c\right)}, \ \textit{for}\ \boldsymbol{\theta}\ \textit{s.t.}\ \frac{s_1(\boldsymbol{\theta})}{N-1}b + \frac{s_0(\boldsymbol{\theta})}{N-1}c \geq v^1 - v^0 \quad (4.94)$$

We define an intermediate lower bound based on the latter inequality along with the inequalities resulting from the incentive compatibility constraints:

$$\underline{\delta}' = \max\left\{ \max_{s_1 \in \{1,\ldots,N-1\}: \frac{s_1}{N-1}b + \frac{N-s_1}{N-1}c > \epsilon_0 - \epsilon_1} \frac{\epsilon_0 - \epsilon_1 - \left(\frac{s_1}{N-1}b + \frac{N-s_1}{N-1}c\right)}{(\epsilon_0 - \epsilon_1)\left(\frac{N-s_1}{N-1}\beta_1^+ + \frac{s_1-1}{N-1}x_1^+\right) - \left(\frac{s_1}{N-1}b + \frac{N-s_1}{N-1}c\right)}, \right.$$
$$\left. \max_{\theta \in \{0,1\}} \frac{c}{c + (1-2\varepsilon)(\beta_\theta^+ - (1-\beta_\theta^-))(\epsilon_0 - \epsilon_1)} \right\} \quad (4.95)$$

Then the lower bound can be written as $\underline{\delta} = \max\{\underline{\delta}', \underline{\delta}''\}$, where $\underline{\delta}''$ is the lower bound that we will derive for the case when the users have the same rating. If the payoffs v^1 and v^0 satisfy the constraint resulting from (α^{a}-3), namely satisfy

$$\frac{(1+\kappa_1 x_0^+)v^1 - (1+\kappa_1 x_1^+)v^0}{x_1^+ - x_0^+} \leq \underline{\delta} z_3 - (1-\underline{\delta})\kappa_1 \cdot (b-c), \quad (4.96)$$

then we use α^{a} to decompose v^1 and v^0. Otherwise, we use α^{f} to decompose v^1 and v^0

4.8.3.5 When the users have the same rating

Now we derive the conditions under which any payoff profile in $\mathcal{W}^{\mathbf{1}_N}$ and $\mathcal{W}^{\mathbf{0}_N}$ can be decomposed.

If all the users have rating 1, namely $\boldsymbol{\theta} = \mathbf{1}_N$, to decompose $\boldsymbol{v} \in \mathcal{W}^{\mathbf{1}_N}$, we need to find a recommended plan α_0 and a simple continuation payoff function $\boldsymbol{\gamma}$ such that for all $i \in \mathcal{N}$ and for all $\alpha_i \in A$,

$$\begin{aligned} v_i &= (1-\delta)u_i(\boldsymbol{\theta}, \alpha_0, \alpha_0 \cdot \mathbf{1}_N) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha_0 \cdot \mathbf{1}_N) & (4.97) \\ &\geq (1-\delta)u_i(\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha_0 \cdot \mathbf{1}_{N-1}) + \delta \sum_{\boldsymbol{\theta}'} \gamma_i(\boldsymbol{\theta}') q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \alpha_0, \alpha_i, \alpha_0 \cdot \mathbf{1}_{N-1}). \end{aligned}$$

When all the users have the same rating, the altruistic plan α^{a} is equivalent to the fair plan α^{f}. Hence, we use the altruistic plan and the selfish plan to decompose the payoff profiles.

If we use the altruistic plan α^{a} to decompose a payoff profile $\boldsymbol{v}$, we have

$$v^1 = (1-\delta)(b-c) + \delta(x_1^+ \gamma^1 + (1-x_1^+)\gamma^0), \quad (4.98)$$

and the incentive compatibility constraint

$$(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right](\gamma^1 - \gamma^0) \geq \frac{1-\delta}{\delta}c. \tag{4.99}$$

Setting $\gamma^1 = \gamma^0 + \frac{1-\delta}{\delta}\frac{c}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]}$ and noticing that $\gamma^0 \in \left[\frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1}, \frac{\kappa_1 z_2 + (\kappa_2-1)z_3}{\kappa_1+\kappa_2}\right]$, we get an lower bound on v^1 that can be decomposed by α^{a}

$$\begin{aligned} v^1 &= (1-\delta)(b-c) + \delta\left(\gamma^0 + x_1^+ \frac{1-\delta}{\delta}\frac{c}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]}\right) & (4.100)\\ &\geq (1-\delta)\left(b - c + c\frac{x_1^+}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]}\right) + \delta\frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1} & (4.101)\end{aligned}$$

If we use the selfish plan α^{s} to decompose a payoff profile $\boldsymbol{v}$, we have

$$v^1 = \delta(x_1^+\gamma^1 + (1-x_1^+)\gamma^0). \tag{4.102}$$

Since the selfish plan is NE of the stage game, the incentive compatibility constraint is satisfied as long as we set $\gamma^1 = \gamma^0$. Hence, we have $v^1 = \delta\gamma^0$. Again, noticing that $\gamma^0 \in \left[\frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1}, \frac{\kappa_1 z_2 + (\kappa_2-1)z_3}{\kappa_1+\kappa_2}\right]$, we get an upper bound on v^1 that can be decomposed by α^{s}

$$v^1 = \delta\gamma^0 \leq \delta\frac{\kappa_1 z_2 + (\kappa_2-1)z_3}{\kappa_1+\kappa_2}. \tag{4.103}$$

In order to decompose any payoff profile $\boldsymbol{v} \in \mathcal{W}^{\mathbf{1}_N}$, the lower bound on v^1 that can be decomposed by α^{a} must be smaller than the upper bound on v^1 that can be decomposed by α^{s}, which leads to

$$\begin{aligned} &(1-\delta)\left(b - c + c\frac{x_1^+}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]}\right) + \delta\frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1} \leq \delta\frac{\kappa_1 z_2 + (\kappa_2-1)z_3}{\kappa_1+\kappa_2} \\ &\Rightarrow \delta \geq \frac{b-c+c\frac{x_1^+}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]}}{b-c+c\frac{x_1^+}{(1-2\varepsilon)\left[\beta_1^+ - (1-\beta_1^-)\right]} + \frac{\kappa_1 z_2 + (\kappa_2-1)z_3}{\kappa_1+\kappa_2} - \frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1}}. \qquad (4.104)\end{aligned}$$

Finally, following the same procedure, we derive the lower bound on δ when all the users have rating 0, namely $\boldsymbol{\theta} = \mathbf{0}_N$. Similarly, in this case, the altruistic plan α^{a} is equivalent to the fair plan α^{f}. Hence, we use the altruistic plan and the selfish plan to decompose the payoff profiles.

If we use the altruistic plan α^{a} to decompose a payoff profile $\boldsymbol{v}$, we have

$$v^0 = (1-\delta)(b-c) + \delta(x_0^+\gamma^1 + (1-x_0^+)\gamma^0), \tag{4.105}$$

and the incentive compatibility constraint

$$(1-2\varepsilon)\left[\beta_0^+ - (1-\beta_0^-)\right](\gamma^1 - \gamma^0) \geq \frac{1-\delta}{\delta}c. \tag{4.106}$$

If we use the selfish plan α^{s} to decompose a payoff profile $\boldsymbol{v}$, we have

$$v^1 = \delta(x_0^+\gamma^1 + (1-x_0^+)\gamma^0). \tag{4.107}$$

Note that when $\boldsymbol{\theta} = \mathbf{0}_N$, if we substitute β_0^+, β_0^-, x_0^- with β_1^+, β_1^-, x_1^-, respectively, the decomposability constraints become the same as those when $\boldsymbol{\theta} = \mathbf{1}_N$. Hence, we derive a similar lower bound on δ

$$\delta \geq \frac{b-c+c\frac{x_0^+}{(1-2\varepsilon)\left[\beta_0^+-(1-\beta_0^-)\right]}}{b-c+c\frac{x_0^+}{(1-2\varepsilon)\left[\beta_0^+-(1-\beta_0^-)\right]} + \frac{\kappa_1 z_2+(\kappa_2-1)z_3}{\kappa_1+\kappa_2} - \frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1}}. \tag{4.108}$$

Finally, we can obtain the lower bound on δ when the users have the same rating as

$$\underline{\delta}'' = \max_{\theta\in\{0,1\}} \frac{b-c+c\frac{x_\theta^+}{(1-2\varepsilon)\left[\beta_\theta^+-(1-\beta_\theta^-)\right]}}{b-c+c\frac{x_\theta^+}{(1-2\varepsilon)\left[\beta_\theta^+-(1-\beta_\theta^-)\right]} + \frac{\kappa_1 z_2+(\kappa_2-1)z_3}{\kappa_1+\kappa_2} - \frac{(1+\kappa_1)(\epsilon_0-\epsilon_1)-z_3}{\kappa_1}}. \tag{4.109}$$

Together with the lower bound $\underline{\delta}'$ derived for the case when the users have different ratings, we can get the lower bound $\underline{\delta}$ specified in Condition 3 of Theorem 7.

4.8.4 Complete Description of the Algorithm

Table 4.7: The algorithm of constructing the equilibrium strategy by the rating mechanism.

```
Require: $b$, $c$, $\varepsilon$, $\xi$; $\tau(\varepsilon)$, $\delta \geq \underline{\delta}(\varepsilon, \xi)$; $\boldsymbol{\theta}^0$
Initialization: $t = 0$, $\epsilon_0 = \xi$, $\epsilon_1 = \epsilon_0/(1 + \frac{\kappa_2}{\kappa_1})$, $v^\theta = b - c - \epsilon_\theta$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.
repeat
    if $s_1(\boldsymbol{\theta}) = 0$ then
        if $v^0 \geq (1-\delta)\left[b - c + \frac{(1-\varepsilon)\beta_0^+ + \varepsilon(1-\beta_0^-)}{(1-2\varepsilon)(\beta_0^+ - (1-\beta_0^-)} c\right] + \delta \frac{\epsilon_0 - \epsilon_1 - z_3}{\kappa_1}$ then
            $\alpha_0^t = \alpha^{\mathrm{a}}$
            $v^0 \leftarrow \frac{v^0}{\delta} - \frac{1-\delta}{\delta}\left[b - c + \frac{(1-\varepsilon)\beta_0^+ + \varepsilon(1-\beta_0^-)}{(1-2\varepsilon)(\beta_0^+ - (1-\beta_0^-)} c\right]$,$v^1 \leftarrow v^0 + \frac{1-\delta}{\delta}\left[\frac{1}{(1-2\varepsilon)(\beta_0^+ - (1-\beta_0^-)} c\right]$
        else
            $\alpha_0^t = \alpha^{\mathrm{s}}$
            $v^0 \leftarrow \frac{v^0}{\delta}, \ v^1 \leftarrow v^0$
        end
    elseif $s_1(\boldsymbol{\theta}) = N$ then
        if $v^1 \geq (1-\delta)\left[b - c + \frac{(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)}{(1-2\varepsilon)(\beta_1^+ - (1-\beta_1^-)} c\right] + \delta \frac{\epsilon_0 - \epsilon_1 - z_3}{\kappa_1}$ then
            $\alpha_0^t = \alpha^{\mathrm{a}}$
            $v^1 \leftarrow \frac{v^1}{\delta} - \frac{1-\delta}{\delta}\left[b - c + \frac{(1-\varepsilon)\beta_1^+ + \varepsilon(1-\beta_1^-)}{(1-2\varepsilon)(\beta_1^+ - (1-\beta_1^-)} c\right]$, $v^0 \leftarrow v^1 - \frac{1-\delta}{\delta}\left[\frac{1}{(1-2\varepsilon)(\beta_1^+ - (1-\beta_1^-)} c\right]$
        else
            $\alpha_0^t = \alpha^{\mathrm{s}}$
            $v^1 \leftarrow \frac{v^1}{\delta}, \ v^0 \leftarrow v^1$
        end
    else
        if $\frac{1+\kappa_1 x_0^+}{x_1^+ - x_0^+} v^1 - \frac{1+\kappa_1 x_1^+}{x_1^+ - x_0^+} v^0 \leq \delta z_3 - (1-\delta)\kappa_1(b-c)$ then
            $\alpha_0^t = \alpha^{\mathrm{a}}$
            $v^{1\prime} \leftarrow \frac{1}{\delta} \frac{(1-x_0^+)v^1 - (1-x_1^+)v^0}{x_1^+ - x_0^+} - \frac{1-\delta}{\delta}(b-c)$, $v^{0\prime} \leftarrow \frac{1}{\delta} \frac{x_1^+ v^0 - x_0^+ v^1}{x_1^+ - x_0^+} - \frac{1-\delta}{\delta}(b-c)$
            $v^1 \leftarrow v^{1\prime}, \ v^0 \leftarrow v^{0\prime}$
        else
            $\alpha_0^t = \alpha^{\mathrm{f}}$
            $v^{1\prime} \leftarrow \frac{1}{\delta} \frac{(1-x_0^+)v^1 - (1-x_{s_1(\boldsymbol{\theta})}^+)v^0}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+} - \frac{1-\delta}{\delta} \frac{(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c)(1-x_0^+) - (\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c)(1-x_{s_1(\boldsymbol{\theta})}^+)}{x_{s_1(\boldsymbol{\theta})}^+ - x_0^+}$
            $v^{1\prime} \leftarrow \frac{1}{\delta} \frac{x_0^+ v^1 - x_{s_1(\boldsymbol{\theta})}^+ v^0}{x_0^+ - x_{s_1(\boldsymbol{\theta})}^+} - \frac{1-\delta}{\delta} \frac{(b - \frac{s_1(\boldsymbol{\theta})-1}{N-1}c)x_0^+ - (\frac{s_0(\boldsymbol{\theta})-1}{N-1}b - c)x_{s_1(\boldsymbol{\theta})}^+}{x_0^+ - x_{s_1(\boldsymbol{\theta})}^+}$
            $v^1 \leftarrow v^{1\prime}, \ v^0 \leftarrow v^{0\prime}$
        end
    end
    $t \leftarrow t + 1$, determine the rating profile $\boldsymbol{\theta}^t$, set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^t$
until $\varnothing$
```

CHAPTER 5

Concluding Remarks

In this thesis, we have studied three classes of multi-agent sequential decision problems: resource sharing with imperfect monitoring, resource sharing with decentralized information, and resource exchange with imperfect monitoring. We have derived optimal solutions for all the above problems, and have demonstrated the significant peroformance gains over state-of-the-art solutions.

There are many future research directions. All the problems studied in this thesis have focused on certain features of the problems while making simplifying assumptions on the other aspects of the problem. It is of great importance and interest, although challenging, to study more general problems that include the problems studied here as special cases. We mention some particularly interesting directions as follows:

- *Extentions on the agents' interaction:* The problems studied in this thesis model the agents' interaction by two extremes: either each player interacts with every else, or each player interacts with only one other player uniformly randomly. We can consider more general models for the interaction. For example, we can have a underlying graph that represents how the agents are connected, under which an agent interacts with its neighbors only. We can also have different matching rules that are not uniformly random. It is especially interesting when the designer can design the network topology or the matching rules. In this case, the optimal design of the agents' interaction is crucial.

- *Extensions on the state dynamics:* The problems studied in this thesis also model the state dynamics by two extremes: either there is a public state known to every agent, or each agent has a private state that is independent of the others' states and actions. We can consider more general models for the state dynamics. The agents can have both the public state and private states. The agents' state transitions can depend on the others' states and actions.

References

[1] M. van der Schaar, Y. Xiao, and W. Zame, “Designing efficient resource sharing for impatient players using limited monitoring,” *Working Paper*, 2014. http://arxiv.org/abs/1309.0262

[2] Y. Xiao and M. van der Schaar, “Dynamic spectrum sharing among repeatedly interacting selfish users with imperfect monitoring,” *IEEE J. Sel. Areas Commun., Special issue on Cognitive Radio Series*, vol. 30, no. 10, pp. 1890–1899, 2012.

[3] Y. Xiao and M. van der Schaar, “Energy-efficient nonstationary spectrum sharing,” *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 810–821, 2014.

[4] L. Song, Y. Xiao, and M. van der Schaar, “Demand side management in smart grids using a repeated game framework,” *IEEE J. Sel. Areas Commun., Special issue on Smart Grid Communications*, 2014. http://arxiv.org/abs/1311.1887

[5] Y. Xiao and M. van der Schaar, “Optimal foresighted multi-user wireless video,” Accepted with minor revision by *IEEE J. Sel. Topics Signal Process., Special issue on Visual Signal Processing for Wireless Networks.* http://arxiv.org/abs/1311.4227

[6] Y. Xiao and M. van der Schaar, “Foresighted Demand Side Management,” *Technical Report*, 2013. http://arxiv.org/abs/1401.2185

[7] Y. Xiao and M. van der Schaar, “Socially-optimal design of service exchange platforms with imperfect monitoring,” *Technical Report*, 2013. http://arxiv.org/abs/1310.2323

[8] G. Mailath and L. Samuelson, *Repeated Games and Reputations: Long-run Relationships.* Oxford, U.K.: Oxford University Press, 2006.

[9] D. Fudenberg, D. K. Levine, and E. Maskin, “The folk theorem with imperfect public information,” *Econometrica*, vol. 62, no. 5, pp. 997–1039, 1994.

[10] S. Athey and K. Bagwell, “Optimal collusion with private information,” *RAND Journal of Economics*, vol. 32, no. 3, pp. 428–465, 2001.

[11] E. J. Green and R. H. Porter, “Noncooperative collusion under imperfect price information,” *Econometrica*, vol. 52, no. 1, pp. 87–100, 1984.

[12] D. Fudenberg and D. K. Levine, “Efficiency and observability with long-run and short-run players,” *Journal of Economic Theory*, vol. 62, pp. 103–135, 1994.

[13] M. Kandori, "The use of information in repeated games with imperfect monitoring," *Review of Economic Studies*, vol. 59, pp. 581–593, 1992.

[14] D. Abreu, D. Pearce, and E. Stacchetti, "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*, vol. 58, no. 5, pp. 1041–1063, 1990.

[15] K. L. Judd, S. Yeltekin, and J. Conklin, "Computing supergame equilibria," *Econometrica*, vol. 71, no. 4, pp. 1239–1254, 2003.

[16] S. Goldlücke and S. Kranz, "Infinitely repeated games with public monitoring and monetary transfers," *Journal of Economic Theory*, vol. 147, no. 3, pp. 1191–1221, 2012.

[17] G. Mailath, I. Obara, and T. Sekiguchi, "The maximum efficient equilibrium payoff in the repeated prisoners' dilemma," *Games and Economic Behavior*, vol. 40, no. 1, pp. 99–122, 2002.

[18] Y. Sannikov, "Games with imperfectly observable actions in continuous time," *Econometrica*, vol. 75, no. 5, pp. 1285–1329, 2007.

[19] P. Jehiel, B. Moldovanu, and E. Stacchetti, "How (not) to sell nuclear weapons," *American Economic Review*, vol. 86, no. 4, pp. 814–829, 1996.

[20] M. Guo, V. Conitzer, and D. M. Reeves, "Competitive repeated allocation without payments," in *Proceedings of the 5th International Workshop on Internet and Network Economics*, WINE '09, pp. 244–255, 2009.

[21] D. Fudenberg, D. K. Levine, and S. Takahashi, "Perfect public equilibrium when players are patient," *Games and Economic Behavior*, vol. 61, no. 1, pp. 27 – 49, 2007.

[22] K. Bharath-Kumar and J. M. Jaffe, "A new approach to performance-oriented flow control," *IEEE Transactions on Communications*, vol. 29, no. 4, pp. 427–435, 1981.

[23] J. Bochnak, M. Coste, and M.-F. Roy, *Real algebraic geometry.* Springer, 1998.

[24] L. E. Blume and W. R. Zame, "The algebraic geometry of perfect and sequential equilibrium," *Econometrica*, vol. 62, no. 4, pp. 783–794, 1994.

[25] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201-220, Feb. 2005.

[26] X. Kang, R. Zhang, Y.-C. Liang, and H. K. Garg, "Optimal power allocation strategies for fading cognitive radio channels with primary user outage constraint," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 374-383, Feb. 2011.

[27] C. W. Tan and S. H. Low, “Spectrum management in multiuser cognitive wireless networks: Optimality and algorithm,” *IEEE J. Sel. Areas Commun.*, vol. 29, no. 2, pp. 421-430, Feb. 2011.

[28] J. Huang, R. A. Berry, and M. L. Honig, “Distributed interference compensation for wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 1074-1084, May 2006.

[29] Y. Xing, C. N. Mathur, M. A. Haleem, R. Chandramouli, and K. P. Subbalakshmi, “Dynamic spectrum access with QoS and interference temperature constraints,” *IEEE Trans. Mobile Comput.*, vol. 6, no. 4, pp. 423-433, Apr. 2007.

[30] L. B. Le and E. Hossain, “Resource allocation for spectrum underlay in cognitive radio networks,” *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5306-5315, Dec. 2008.

[31] N. Gatsis, A. G. Marques, G. B. Giannakis, “Power control for cooperative dynamic spectrum access networks with diverse QoS constraints,” *IEEE Trans. Commun.*, vol. 58, no. 3, pp. 933-944, Mar. 2010.

[32] O. Ileri, D. Samardzija, and N. Mandayam, “Demand responsive pricing and competitive spectrum allocation via a spectrum server,” in *Proc. IEEE DySPAN*, Baltimore, MD, Nov. 2005, pp. 194C202.

[33] J. Acharya and R. Yates, “A framework for dynamic spectrum sharing between cognitive radios,” in *Proc. IEEE ICC*, Glasgow, Scotland, Jun. 2007, pp. 5166C5171.

[34] S. Sharma and D. Teneketzis, “An externalities-based decentralized optimal power allocation algorithm for wireless networks,” *IEEE/ACM Trans. Netw.*, vol. 17, no. 6, pp. 1819–1831, Dec. 2009.

[35] S. Sorooshyari, C. W. Tan, M. Chiang, “Power control for cognitive radio networks: Axioms, algorithms, and analysis,” To appear in *IEEE/ACM Trans. Netw.*, 2011.

[36] J. Huang, R. A. Berry, and M. L. Honig, “Auction-based spectrum sharing,” *Mobile Networks and Applications*, vol. 11, pp. 405-418, 2006.

[37] S. Sharma and D. Teneketzis, “A game-theoretic approach to decentralized optimal power allocation for cellular networks,” *Telecommunication Systems*, pp. 1–16, 2010.

[38] Y. Xiao, J. Park, and M. van der Schaar, “Intervention in power control games with selfish users,” *IEEE J. Sel. Topics Signal Process., Special issue on Game Theory in Signal Processing*, vol. 6, no. 2, pp. 165–179, Apr. 2012.

[39] R. Etkin, A. Parekh, and D. Tse, "Spectrum sharing for unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 517–528, Apr. 2007.

[40] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "Repeated open spectrum sharing game with cheat-proof strategies," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1922-1933, 2009.

[41] M. Le Treust and S. Lasaulce, "A repeated game formulation of energy-efficient decentralized power control," *IEEE Trans. on Wireless Commun.*, vol. 9, no. 9, pp. 2860–2869, september 2010.

[42] Y. Xiao, J. Park, and M. van der Schaar, "Repeated games with intervention: Theory and applications in communications," To appear in *IEEE Trans. Commun.*. Available: "http://arxiv.org/abs/1111.2456".

[43] S. Stańczak and H. Boche, "On the convexity of feasible QoS regions," *IEEE Trans. Inf. Theory*, vol. 53, no. 2, Feb. 2007.

[44] J. Park and M. van der Schaar, "Cognitive MAC protocols using memory for distributed spectrum sharing under limited spectrum sensing," *IEEE Trans. Commun.*, vol. 59, no. 9, pp. 2627-2637, Sep. 2011.

[45] C. Cordeiro and K. Challapali, "C-MAC: A cognitive MAC protocol for multi-channel wireless networks," in *Proc. Symposium on Dynamic Spectrum Access Networks (DySPAN'07)*, pp. 147–157, Apr. 2007.

[46] M. Timmers, S. Pollin, A. Dejonghe, L. van der Perre, and F. Catthoor, "A distributed multichannel MAC protocol for multihop cognitive radio networks, *IEEE Trans. Veh. Technol.*, vol. 59, no. 1, pp. 446–459, Jan. 2010.

[47] A. De Domenico, E. C. Strinati, and M. G. Di Benedetto, "A survey on MAC strategies for cognitive radio networks," *IEEE Commun. Surveys Tutorials*, vol. 14, no. 1, pp. 21–44, 2012.

[48] N. Gatsis, A. G. Marques, G. B. Giannakis, "Power control for cooperative dynamic spectrum access networks with diverse QoS constraints," *IEEE Trans. Commun.*, vol. 58, no. 3, pp. 933–944, Mar. 2010.

[49] L. Zheng and C. W. Tan, "Cognitive radio network duality and algorithms for utility maximization," *IEEE J. on Sel. Areas Commun.*, Vol. 31, No. 3, pp. 500–513, Mar. 2013.

[50] C. W. Tan, M. Chiang, and R. Srikant, "Fast algorithms and performance bounds for sum rate maximization in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 706–719, Jun. 2013.

[51] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 291–303, Feb. 2002.

[52] G. He, S. Lasaulce, and Y. Hayel, "Stackelberg games for energy-efficient power control in wireless networks," *Proc. IEEE INFOCOM'2011*, pp. 591–595, 2011.

[53] R. Xie, F. R. Yu, and H. Ji, "Energy-efficient spectrum sharing and power allocation in cognitive radio femtocell networks," *Proc. IEEE INFOCOM'2012*, pp. 1665–1673, 2012.

[54] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 1341–1347, 1995.

[55] N. Bambos, S. Chen, and G. Pottie, "Channel access algorithms with active link protection for wireless communication networks with power control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 583–597, Oct. 2000.

[56] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power control scheme in wireless cellular systems," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 210–221, Apr. 2003.

[57] E. Altman and Z. Altman, "S-modular games and power control in wireless networks," *IEEE Trans. Autom. Control*, vol. 48, no. 5, pp. 839–842, May 2003.

[58] P. Hande, S. Rangan, M. Chiang, and X. Wu, "Distributed uplink power control for optimal SIR assignment in cellular data networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1420–1433, Dec. 2008.

[59] S. M. Perlaza, H. Tembine, S. Lasaulce, and M. Debbah, "Quality-of-service provisioning in decentralized networks: A satisfaction equilibrium approach," *IEEE J. Sel. Topics Signal Process., Special issue on Game Theory in Signal Processing*, vol. 6, pp. 104–116, Apr. 2012.

[60] C. W. Tan, D. P. Palomar, and M. Chiang, "Energy-robustness tradeoff in cellular network power control," *IEEE/ACM Trans. Netw.*, vol. 17, no. 3, pp. 912–925, Jun. 2009.

[61] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 589–600, 2007.

[62] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Proc.*, vol. 58, no. 11, pp. 5667-5681, Nov., 2010.

[63] K. Liu, Q. Zhao, and B. Krishnamachari, "Dynamic multichannel access with imperfect channel state detection," *IEEE Trans. Signal Proc.*, vol. 58, no. 5, May 2010.

[64] K. Liu and Q. Zhao, "Cooperative game in dynamic spectrum access with unknown model and imperfect sensing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, Apr. 2012.

[65] H. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 320–331, 2011.

[66] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," *Proc. IEEE Power and Energy Society General Meeting*, 2011.

[67] L. Jiang and S. H. Low, "Multi-period optimal procurement and demand responses in the presence of uncrtain supply," *Proc. IEEE Conference on Decision and Control (CDC)*, Dec. 2011.

[68] B.-G. Kim, S. Ren, M. van der Schaar, and J.-W. Lee, "Bidirectional energy trading and residential load scheduling with electric vehicles in the smart grid," *IEEE J. Sel. Areas Commun., Special issue on Smart Grid Communications Series*, vol. 31, no. 7, pp. 1219–1234, Jul. 2013.

[69] A. Malekian, A. Ozdaglar, E. Wei, "Competitive equilibrium in electricity markets with heterogeneous users and ramping constraints," *Proc. IEEE Allerton Conference*, 2013.

[70] K. M. Chandy, S. H. Low, U. Topcu, and H. Xu, "A simple optimal power flow model with energy storage," *Proc. IEEE Conference on Decision and Control (CDC)*, Dec. 2010.

[71] Italo Atzeni, Luis G. Ordóñez, Gesualdo Scutari, Daniel P. Palomar, and Javier R. Fonollosa, "Noncooperative and cooperative optimization of distributed energy generation and storage in the demand-side of the smart grid," *IEEE Trans. on Signal Process.*, vol. 61, no. 10, pp. 2454–2472, May 2013.

[72] Italo Atzeni, Luis G. Ordóñez, Gesualdo Scutari, Daniel P. Palomar, and Javier R. Fonollosa, "Demand-side management via distributed energy generation and storage optimization," *IEEE Trans. on Smart Grids*, vol. 4, no. 2, pp. 866–876, June 2013.

[73] L. Jia and L. Tong, "Optimal pricing for residential demand response: A stochastic optimization approach," *Proc. IEEE Allerton Conference*, 2012.

[74] L. Jia, L. Tong, and Q. Zhao, "Retail pricing for stochastic demand with unknown parameters: An online machine learning approach," *Proc. IEEE Allerton Conference*, 2013.

[75] L. Huang, J. Walrand and K. Ramchandran, "Optimal demand response with energy storage management," *Technical Report.* Available: "http://arxiv.org/abs/1205.4297".

[76] L. Huang, J. Walrand, and K. Ramchandran, "Optimal power procurement and demand response with quality-of-usage guarantees," *Proc. IEEE Power and Energy Society General Meeting*, Jul. 2012.

[77] Y. Zhang and M. van der Schaar, "Structure-aware stochastic load management in smart grids," accepted and to appear in *Infocom 2014*.

[78] J. Hawkins, "A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications," PhD Dissertation, MIT, Cambridge, MA, 2003.

[79] F. Fu and M. van der Schaar, "A systematic framework for dynamically optimizing multi-user video transmission," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 308–320, Apr. 2010.

[80] E. Altmam, K. Avrachenkov, N. Bonneau, M. Debbah, R. El-Azouzi, D. S. Menasche, "Constrained cost-coupled stochastic games with independent state processes," *Technical Report.* Available: http://www-sop.inria.fr/members/Konstantin.Avratchenkov/pubs/ConstrGame.pdf

[81] J. Hörner, T. Sugaya, S. Takahashi, and N. Vielle, "Recursive methods in discounted stochastic games: An algorithm for $\delta \to 1$ and a folk theorem," *Econometrica*, vol. 79, no. 4, pp. 1277–1318, 2011.

[82] J. Yao, I. Adler, and S. S. Oren, "Modeling and computing two-settlement oligopolistic equilibrium in a congested electricity network," *Operations Research*, vol. 56, no. 1, pp. 34–47, 2008.

[83] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 645–658, Dec. 2011.

[84] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Trans. Veh. Tech.*, vol. 58, no. 4, pp. 1904–1919, May 2009.

[85] V. Borkar and S. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optimization*, vol. 28, pp. 447–469, 1999.

[86] R. Sutton and A. Barto, "Reinforcement learning: An introduction," MIT Press, 1998.

[87] "Power Systems Test Case Archive," Available: http://www.ee.washington.edu/research/pstca/

[88] E. B. Fisher, R. P. O'Neill, and M. C. Ferris, "Optimal transmission switching," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1346–1355, Aug. 2008.

[89] C. Zhao, U. Topcu, and S. H. Low, "Optimal load control via frequency measurement and neighborhood area communication," *IEEE Trans. on Power Syst.*, vol. 28, no. 4, pp. 3576–3587, Nov. 2013.

[90] A. R. Borden, D. K. Molzahn, B. C. Lesieutre, and P. Ramanathan, "Power system structure and confidentiality preserving transformation of optimal power flow model," in *Proc. 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013.

[91] H. Kushner and G. Yin, Stochastic Approximation and Recursive Algorithms and Applications. Springer, 2003.

[92] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[93] Y. Zhang and M. van der Schaar, "Rating protocols for online communities," *ACM Transactions on Economics and Computation*, 2013.

[94] Y. Zhang and M. van der Schaar, "Peer-to-peer multimedia sharing based on social norms," *Elsevier Journal Signal Processing: Image Communication Special Issue on "Advances in video streaming for P2P networks"*, vol. 27, no. 5, pp. 383–400, May 2012.

[95] A. Blanc, Y.-K. Liu, and A. Vahdat, "Designing incentives for peer-to-peer routing," in *Proceedings of the IEEE INFOCOM 2005*, 2005.

[96] M. Feldman, K. Lai, I. Stoica, and J. Chuang, "Robust incentive techniques for peer-to-peer networks," in *Proceedings of ACM Conference on Electronic Commerce*, 2004.

[97] C. Dellarocas, "Reputation mechanism design in online trading environments with pure moral hazard," *Information Systems Research*, vol. 16, no. 2, pp. 209–230, 2005.

[98] M. Kandori, "Social norms and community enforcement," *Review of Economic Studies*, vol. 59, no. 1, pp. 63 – 80, 1992.

[99] M. Okuno-Fujiwara and A. Postlewaite, "Social norms and random matching games," *Games and Economic Behaviors*, vol. 9, no. 1, pp. 79 – 109, 1993.

[100] P. Dal Bó, "Social norms, cooperation and inequality," *Economic Theory*, vol. 30, pp. 89 – 105, 2007.

[101] K. Hasker, "Social norms and choice: A general folk theorem for repeated matching games with endogenous matching rules," *International Journal of Game Theory*, vol. 36, pp. 137–146, 2007.

[102] S. Takahashi, "Community enforcement when players observe partners' past play," *Journal of Economic Theory*, vol. 145, no. 1, pp. 42–62, 2010.

[103] J. Deb, "Cooperation and community responsibility: A folk theorem for random matching games with names," *Revise and Resubmit at the Review of Economic Studies*, 2013.

[104] G. Ellison, "Cooperation in the prisoner's dilemma with anonymous random matching," *Review of Economic Studies*, vol. 61, no. 3, pp. 567 – 588, 1994.

[105] R. Izhak-Ratzin, H. Park, and M. van der Schaar, "Reinforcement learning in BitTorrent systems," in *Proc. IEEE Infocom 2011*, 2011.

[106] H. Park and M. van der Schaar, "Evolution of resource reciprocation strategies in P2P networks," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1205–1218, Mar. 2010.