Dynamic decision problems with cooperative and strategic agents and asymmetric information

by

Deepanshu Vasal

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in the University of Michigan
2016

Doctoral Committee:

Associate Professor Achilleas Anastasopoulos, Chair
Professor Mingyan Liu
Assistant Professor Grant Schoenebeck
Associate Professor Vijay G. Subramanian
Professor Rakesh V. Vohra, University of Pennsylvania

2016

To my parents and grandparents

Acknowledgments

Foremost, I would like to express my deepest gratitude to my advisor Prof. Achilleas Anastasopoulos for his invaluable guidance, patience and support during my graduate studies. Over the years, I have learned a lot from him, in problem solving and conducting meaningful research, writing good papers and giving clear presentations. It is absolutely amazing how closely he has worked with me on problems, allowing me time to make mistakes and learn from them. He has always been approachable, friendly, responsive and enthusiastic towards working on problems, which kept me motivated through the years. I feel very fortunate to have been given an opportunity to work on hard, conceptual problems that are also relevant.

I would then like to thank my committee members for their participation in my defense presentations. I would especially like to thank Prof. Vijay Subramanian for his guidance in research and constructive feedback on my dissertation, which helped me greatly, in improving the content and presentation of my thesis. I would like to extend my heartfelt gratitude to many professors whom I have interacted with through courses or otherwise, and have enriched my experience in graduate school, including Prof. Sandeep Pradhan, Prof. Demosthenis Teneketzis and Prof. Kim Winick. I consider myself fortunate to attend a great institution like Michigan, where I learnt immensely from excellent courses in EECS and Math.

I want to thank my friends for their support, without whom I may have graduated earlier, but may not have been this fun. I would especially like to thank Arun for being a great friend and a mentor. He pulled me into studying topics in math on fine summer days, and on the way, instilled in me a deep appreciation for the same. Over the years, he exposed me to many beautiful mathematical ideas, and I learnt a lot from him about working on hard probability problems, and life in general. I also want to thank Farhad, Mohsen, Mike and James for fun discussions. I would like to deeply thank Mich Haus and its inhabitants, including Chelsea, Ryan, Swiz, David and Mikhail, for the whacky times, and for providing open, cooperative environment, helping me grow personally. Bike rides with Julien, Aniket, Pallavi and Bafna would be one of the sweet, memorable times for me. As will be the fun times with Kartiki and gang.

Finally, I would like to express my deepest gratitude to my family, for providing unconditional support to pursue my academic goals. Their support and values, helped me go through some tough times. Their insistence on sincerity and dedication has been instrumental in helping me achieve my goals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Dynamic decision problems with cooperative and strategic agents and asymmetric information

by

Deepanshu Vasal

Chair: Achilleas Anastasopoulos

There exist many real world situations involving multiple decision makers with asymmetric information, such as communication systems, social networks, economic markets and many others. Through this dissertation, we attempt to enhance the conceptual understanding of such systems and provide analytical tools to characterize the optimum or equilibrium behavior.

Specifically, we study four discrete time, decentralized decision problems in stochastic dynamical systems with cooperative and strategic agents. The first problem we consider is a relay channel where nodes' queue lengths, modeled as conditionally independent Markov chains, are nodes' private information, whereas nodes' actions are publicly observed. This results in non-classical information pattern. Energy-delay tradeoff is studied for this channel through stochastic control techniques for cooperative agents. Extending this model for strategic users, in the second problem we study a general model with N strategic players having conditionally independent, Markovian types and publicly observed actions. This results in a dynamic game with asymmetric information. We present a forward/backward sequential decomposition algorithm to find a class of perfect Bayesian equilibria of the game. Using this methodology, in the third problem, we study a general two player dynamic LQG

game with asymmetric information, where players' types evolve as independent, controlled linear Gaussian processes and players incur quadratic instantaneous costs. We show that under certain conditions, players' strategies that are linear in their private types, together with Gaussian beliefs, form a perfect Bayesian equilibrium (PBE) of the game. Finally, we consider two sub problems in decentralized Bayesian learning in dynamic games. In the first part, we consider an ergodic version of a sequential buyers game where strategic users sequentially make a decision to buy or not buy a product. In this problem, we design incentives to align players' individual objectives with the team objective. In the second part, we present a framework to study learning dynamics and especially informational cascades for decentralized dynamic games. We first generalize our methodology to find PBE to the case when players do not perfectly observe their types; rather they make independent, noisy observations. Based on this, we characterize informational cascades for a specific learning model.

CHAPTER 1

Introduction

1.1 Background

Dynamic decision problems are ubiquitous in real life situations and are studied in many academic disciplines such as communication systems, industrial engineering, computer science, economics, and many many more. Some examples include inventory control, iterative decoding, resource allocation, finding minimum spanning tree, traffic management, shortest path algorithms, computing equilibria for markets, control of queues, sequential hypothesis testing, and the list is unending. Such problems involve a single or multiple decision makers (also referred to as players, agents, users or controllers) who make observations and take actions throughout the duration of the process and also receive rewards (or incur costs). Each player wants to maximize its total reward, which may or may not align with other players rewards. If the players' rewards are aligned i.e. when all players have the same objective, then we refer to such problems as team problems. If the players have different objectives, we refer to such problems as game problems. In this thesis, we study scenarios of decision makers with different information sets in a dynamic setting and provide tools to analyze such systems, and present structural results for optimum or equilibrium strategies.

We start by describing a simple, canonical example on inventory control to highlight some key ideas from stochastic control theory for a problem with classical information structure i.e. with a single controller and with perfect recall. Suppose there is a gasoline seller who makes a decision everyday on the quantity of gasoline she buys to maintain a stock, based on the demand and her stock capacity. Let $x_t \in \mathcal{X}$ be the stock of gasoline she has at the starting of the day t, $u_t \in \mathcal{U}$ be the amount she buys and $w_t \in \mathcal{W}$ is the random demand she receives during the day, whose statistics she knows. Then, the next day, her stock is given by

$$x_{t+1} = x_t + u_t - w_t$$

Her everyday reward depends on the amount of gasoline she sells that day, which is $min(x_t + u_t, w_t)$, and she wants to maximize her cumulative rewards over the duration of T days. Till day t, she has made the observations $x_1, x_2, \ldots x_t$, and thus, her decision action on that day, u_t, is some function of this information available, i.e. $u_t = g_t(x_1, \ldots x_t)$, where g_t is her strategy. Thus, she wants to find the best set of strategies $(g_1, g_2 \ldots, g_T)$ that maximizes her total reward for T days.

Assuming the action and space of stock, $\mathcal{U}$ and $\mathcal{X}$, are finite, then there exist $|\mathcal{U}|^{|\mathcal{X}|^t}$ possible strategies g_t, at time t. For any finite duration T, the complexity of the last day dominates and thus the space of optimization for the problem is of the order of $|\mathcal{U}|^{|\mathcal{X}|^T}$. Since the space of optimization increases double exponentially in T, it renders the problem practically intractable for any reasonable time duration. This curse of dimensionality, as presented in this canonical example, represents a fundamental issue in dynamical optimization problems. However, many a times there exists more structure to the problem, for example a concept of 'state' of system, which could be exploited to mitigate this issue. For example, in the problem described before, if it were known that w_t are i.i.d. random variables, then $(X_t, U_t)_t$ can be shown to be a controlled Markov process and results from classical stochastic control theory provide structural results for the optimal policies. Specifically, these results show that there exist an optimal strategy at time t, which depends only on the current state x_t, i.e. $u_t = g_t^*(x_t)$, and thus the optimal strategy g_t^* could be found in the space of $|\mathcal{U}|^{|\mathcal{X}|}$ functions. Moreover, there exists a backward recursive, dynamic programming methodology to find optimal strategies, which further reduces the space of optimization at time t to $|\mathcal{X}||\mathcal{U}|$. Thus, for a problem with a finite horizon T, this methodology reduces the complexity of the optimization from double exponential in T to linear in T, which highlights the power and usefulness of this technique.

Such problems, for which the state of the system is perfectly observed, are called Markov decision problems (MDPs). If the state is not observed perfectly, rather independent, noisy observations are made, then such problems are called partially observed Markov decision problems (POMDPs), which are also MDPs with posterior beliefs as perfectly observed state (for a precise and elaborate description, see a standard text on stochastic control e.g. [31]).

These problems consists of

- State update function: $x_{t+1} = f(x_t, u_t, w_t)$
- Observation function (for POMDP): $y_t = h(x_t, v_t)$
- Actions as function of information (MDP or POMDP) : $u_t = g_t(x_1, \ldots, x_t)$ or $u_t = g_t(y_1, \ldots, y_t)$

- Instantaneous reward (or cost): $R_t(x_t, u_t)$
- All basic random variables $(x_1, w_1, w_2, \ldots, v_1, v_2, \ldots)$ are mutually independent
- Objective: $\max_g(\text{ or } \min_g)\, \mathbb{E}^g\{\sum_{t=1}^{T} R_t(X_t, U_t)\}$

Dynamic programming is used profusely in many dynamic optimization problems to find analytical and numerical, optimal or near-optimal solutions. One such case that has been extensively considered in literature, for its virtue of being analytical and for the appeal for its ease of implementation, is linear quadratic Gaussian (LQG) control. In the LQG model for perfectly observed states, the state update is linear, the instantaneous cost is quadratic in state and control, and all basic random variables are i.i.d. Gaussian. The optimal strategies are linear function of the state with coefficients as Kalman gains. If the state is not observed perfectly, but through a linear, independent observation kernel, then it is shown that strategies that are linear function of estimate of the state are optimal, with same coefficients as in the case of perfectly observable state. This substitution of estimate of state for the state itself, in the optimal control policies, is also referred to as certainty equivalence [31] or separation of estimation and control [60].

The problem described above with a single controller with perfect recall (i.e. with access to all past observations) is called a stochastic control problem with classical information pattern. The problem becomes considerably more difficult for non-classical information pattern i.e. when there are multiple controllers with different information sets, or without perfect recall, or both. For example, it is shown in Witsenhausen's counterexample [60] that even for a very simple two-stage LQG system without perfect recall, linear strategies are not optimal, and moreover, the optimization problem is non-convex.

In this thesis, we always assume perfect recall and refer to problems with multiple controllers with different information but same objective, as decentralized team problems. There are specifically two key line of thoughts in the literature to find structural properties of the optimal control policies (which we discuss more in chapter 2, where we deal with a decentralized team problem). The first approach, which is called agent-by-agent approach [18], works as follows. It is shown that for any fixed strategy of the other players, player i faces an MDP with an appropriately defined state, and thus can restrict its search over Markov policies that are function of that state. Since each player can do the same, using this approach, one can show that there exist optimal policies for the players that are functions of a considerably smaller set of players' available information. The second approach, which is referred to as common-agent approach [43], works as follows. It assumes that there is a fictitious common agent who, at each time t, observes the common information of the players at that time, and take actions that are prescription functions for the

players. Each player uses that prescription function on its private information to generate its action. Using this approach, the decentralized problem is shown to be equivalent to a centralized problem with only one decision maker, the common agent[1]. Then it is shown that common agent's problem is a POMDP, thus there exists a dynamic programming equation to find its optimal policies. Its optimal policies are Markov in nature and are functions of the posterior belief on the state of the system and players' private information conditioned on the common information. The optimal policies of the common agent can easily be translated to decentralized optimal policies of the players.

When players are strategic and information is perfect and complete, the appropriate notions of equilibria are sub-game perfect equilibrium (SPE) and Markov perfect equilibrium (MPE) [38, 45]. These equilibria can be found through backward induction by computing Nash equilibria for every subgame, for every history or every state, respectively. When players are strategic and information is asymmetric (although complete), such games are called dynamic games with asymmetric information and appropriate notions of equilibria include perfect Bayesian equilibrium (PBE), sequential equilibrium (SE), trembling hand equilibrium (THE). In such games, for every time t, for every history of the game h_t, player i observes only part of it, say h_t^i. For the part that it does not observe i.e. $h_t \backslash h_t^i$, it puts a belief on it, in order to calculate its future reward from that time on. Thus the equilibrium notion consists of a strategy and a belief profile for the players for all private histories. The strategies satisfy sequential rationality conditions (i.e. no player gains by unilateral deviation in strategies, for every subgame) using equilibrium beliefs and the beliefs are found using equilibrium strategies and Bayes' rule (with some other refinements). Thus there is a circular argument for finding equilibrium strategy and belief profiles, and there does not exist any dynamic programming like backward recursive methodology to find such equilibria for such games in general. This remains a bottleneck in studying many real-life situations that involve strategic agents in a dynamical system with different information sets, for instance social networks, markets etc.

1.2 Problems considered

In this thesis, we consider four problems of stochastic systems with asymmetric information pattern. A common thread in these problems is that they involve multiple decision makers with different information sets with common and private components. There is an

[1]These problems are equivalent for total reward in expectation but not for every realization. This point becomes crucial and hinders this approach from being utilized directly for decentralized dynamic games, as discussed in chapter 3 in section 3.4.2.

underlying discrete time dynamical system that obeys controlled Markov dynamics. Players are cooperative or strategic, and incur cost or rewards in each period that are additive over a time horizon.

In the first problem, described in chapter 2, we study node cooperation in a wireless network from the multiple access control (MAC) layer perspective. A simple relay channel with a source, a relay and a destination node is considered, where the source and the relay nodes have packets arriving as Bernoulli arrival processes. The source can transmit a packet directly to the destination or transmit through the relay. The tradeoff between average energy and delay is studied by posing the problem as a stochastic dynamical optimization problem. The following two cases are considered: (a) nodes are cooperative and information is decentralized; (b) nodes are strategic and information is centralized.

With decentralized information and cooperative nodes, a structural result is proven that the optimal policy is the solution of a Bellman-type fixed-point equation over a time invariant state space. For specific cost functions reflecting transmission energy consumption and average delay, numerical results are presented showing that a policy found by solving this fixed-point equation outperforms conventionally used time-division multiple access (TDMA) and random access (RA) policies.

When nodes are strategic and information is common knowledge, it is shown that cooperation can be induced by exchange of payments between the nodes, imposed by the network designer such that the socially optimal Markov policy corresponding to the centralized solution is the unique subgame perfect equilibrium of the resulting dynamic game.

Taking motivation from the previous model, we then consider in chapter 3, a finite horizon dynamic game with N selfish players, who observe their types privately and take actions, which are publicly observed. Players' types evolve as conditionally independent Markov processes, conditioned on their current actions. Their actions and types jointly determine their instantaneous rewards. Since each player has a different information set, this is a dynamic game with asymmetric information, and in general, there is no known methodology to find perfect Bayesian equilibria (PBE) for such games. In this chapter, for a specific class of such games with independent types, we develop a methodology to obtain a class of PBE using a belief state based on players' common information. We first show that any expected reward profile that can be achieved by any general strategy profile can also be achieved by a policy based on players' private information and this belief state. With this structural result as our motivation, we develop our main result that provides a two-step backward-forward recursive algorithm to find a class of PBE of this game that are based on this belief state. We refer to such equilibria as *structured Bayesian perfect equilibria* (SPBE). The backward recursive part of this algorithm defines an equilibrium

generating function. Each period in the backward recursion involves solving a fixed point equation on the space of probability simplexes for every possible belief on types. Using this function, equilibrium strategies and beliefs are generated through a forward recursion.

In chapter 4, we then consider a finite horizon dynamic game with two players who observe their types privately and take actions, which are publicly observed. Players' types evolve as independent, controlled linear Gaussian processes and players incur quadratic instantaneous costs. This forms a dynamic linear quadratic Gaussian (LQG) game with asymmetric information. We show that under certain conditions, players' strategies that are linear in their private types, together with Gaussian beliefs form an SPBE of the game. Furthermore, it is shown that this is a signaling equilibrium due to the fact that future beliefs on players' types are affected by the equilibrium strategies. We provide a backward-forward algorithm to find SPBEs. Each step of the backward algorithm reduces to solving an algebraic matrix equation for every possible realization of the state estimate covariance matrix. The forward algorithm consists of Kalman filter recursions, where state estimate covariance matrices depend on equilibrium strategies. As a result, unlike the case of classical stochastic control or LQG games with non-signaling equilibria, the beliefs are strategy dependent.

In Chapter 5, we study two problems that relate to decentralized Bayesian learning in dynamical systems with strategic agents. In the first problem, we consider the problem of how strategic users with asymmetric information can learn an underlying time-varying state in a sequential buyers game. The exogenously selected strategic users sequentially make a decision to buy or not to buy a product, which is either good or bad, based on their private observation and publicly available information about decision of the past users. There is interesting literature on this problem, on occurrence of informational cascades under certain conditions where a user would discard its private information and base its decision on previous users' actions. This leads to its actions being uninformative for future users, and learning stops for the team as a whole. Every future player repeats the same action and users are said to be in a cascade. For a social objective, it is desirable to avoid to bad cascades. We consider an ergodic version of this problem where users who observe private signals about the state, sequentially make a decision about buying a product whose value varies with time via an ergodic process. We formulate the team problem as an instance of decentralized stochastic control and characterize its optimal policies. With strategic users, we design incentives such that users reveal their true private signals, so that the gap between the strategic and team objective is small, and the overall expected incentive payments are also small.

In the second part, we study a more general model for decentralized Bayesian learn-

ing in a dynamical system involving strategic agents with asymmetric information, where players participate (take actions and receive rewards) for the whole duration of the game, and cases where an internal process selects which subset of players will act at each time instance. The proposed methodology hinges on a sequential decomposition for finding perfect Bayesian equilibria (PBE) of a general class of dynamic games with asymmetric information, where users' types evolve as conditionally independent Markov process and users make independent noisy observations of their types. Based on this methodology, we study a specific scenario of Bayesian learning where we characterize informational cascades for the truly dynamic game considered.

1.2.1 Contributions

In this thesis, we make contributions to the theory of dynamic games with asymmetric information and provide insights into specific decentralized problems considered.

In chapter 2, we utilize two keys ideas in the literature of decentralized team problems to provide structural results of the optimum policies in energy-delay tradeoff in a relay channel. Based on these structural results, we find two, potentially suboptimal policies and show that they perform better than standard TDMA (time division multiple access) and RA (random access) policies. Furthermore, for strategic users with complete and perfect information, we show existence of incentives to align the team objective with the social goal. This is one of the very few works that considers the stochastic arrival nature of the packets in a relay channel, and studies the problem from the stochastic control perspective. In general, the structural results presented motivate design or redesign of optimal and suboptimal policies for cooperative communication in decentralized network systems.

In chapter 3, we consider a class of dynamic games with asymmetric information with conditionally independent Markovian types. We provide a sequential decomposition methodology to find a class of PBE of the game, based on common belief state, where there does not exist such a methodology for such games in general. We illustrate this methodology for a public goods example in this chapter and for models considered in chapter 4 and chapter 5. Before this, a common approach to find a PBE was to guess the solution, if possible, and prove that it satisfies the equilibrium conditions. This, of course ,could be done for simpler systems, and in general, PBE remained an elusive concept mainly for theoretical understanding and interest. This methodology provides a fundamental result in the theory of dynamic games, which opens up the door to study many problems in economic markets, social networks, auctions and more, and to provide analytical and numerical solutions that were not tractable before. The existence of a solution of the fixed point equation

in the backward recursion remains an open problem. In general, this work inspires new research directions such as

(a) Proving existence of solution for the fixed point equation for general or a class of games;

(b) Finding such decomposition for other classes of dynamic games with asymmetric information;

(c) Finding structural properties of equilibria in specific games;

(d) Extension of the methodology for infinite time horizon games;

(e) Dynamic mechanism design for dynamic games with asymmetric information.

In chapter 4, we study dynamic LQG games with asymmetric information, where we use the methodology developed in chapter 3 to show that under certain conditions, signaling strategies that are linear in players' private information, in conjunction with Gaussian beliefs, form a PBE of the game. This result is an important result in the theory of the dynamic LQG games and extends the models considered in the literature thus far. We also provide algorithmic sufficient conditions for a solution to exist for scalar actions. However, there remains a possibility of finding more general conditions.

In chapter 5, we consider the problem of decentralized Bayesian learning in dynamic games through two specific models. In the first problem, we highlight the significance of certain infrequent histories of the game that play a very crucial role for the learning of the players as a whole. Specifically, we show that by incentivizing players to report their observations at these histories, and thus contributing to the learning of the team, the resulting game objective is close to the social objective. And since such histories are infrequent, the expected payout is small. In the second problem, we consider a more general model than the one considered in the informational cascades literature for the Bayesian learning with strategic agents, where players participate in the game for the whole duration. We first extend the methodology developed in chapter 3 to find PBE for the case where users' do not observe their types, but make independent, noisy observations. We propose this as a framework to study informational cascades for a more dynamic set-up than the one studied in the literature. Based on this methodology, for a specific learning model, we characterize sets of common beliefs as "cascades", such that if these beliefs occur at any point in the game, then players repeat the cascading actions for the rest of the game. In general this framework presents a vast unexplored landscape to study learning dynamics and informational cascades. Some important research directions include

(a) Characterization of cascades for specific classes of models;

(b) Studying convergent learning behavior in such games including the probability and the rate of "falling" into a cascade;

(c) Incentive or mechanism design to avoid bad cascades.

1.3 Notation

We use the following notation throughout this thesis, however, some chapter specific notation is provided in at the end of introductions of the respective chapters. A random variable is denoted by an upper case letter and its realization by the corresponding lower case letter. Subscripts denote time indices, such that $X_{a:b}$ is a short hand for the vector $(X_a, X_{a+1}, \ldots, X_b)$, if $a > b$, then $X_{a:b}$ is empty. Superscripts denote agents' identities, such that U_t^i, is a quantity relevant to agent i. We use notation $-i$ to represent all players other than player i i.e. $-i = \{1, 2, \ldots i-1, i+1, \ldots, N\}$. We use A_t^{-i} to mean $(A_t^1, A_t^2, \ldots, A_t^{i-1}, A_t^{i+1} \ldots, A_t^N)$. We remove superscripts or subscripts if we want to represent the whole vector, for example A_t represents $(A_t^1, \ldots, A_t^N)$. In a similar vein, for any collection of sets $(\mathcal{X}^i)_{i\in\mathcal{N}}$, we denote $\times_{i\in\mathcal{N}}\mathcal{X}^i$ by $\mathcal{X}$. We denote the indicator function of any set A by $I_A(\cdot)$. For any finite set $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ represents the space of probability measures on $\mathcal{S}$ and $|\mathcal{S}|$ represents its cardinality. We denote by P^g (or $\mathbb{E}^g$) the probability measure generated by (or expectation with respect to) strategy profile g. We denote the set of real numbers by $\mathbb{R}$. All equalities and inequalities involving random variables are to be interpreted in the *a.s.* sense, unless otherwise specified.

The proofs of theorems, lemmas and claims in each chapter, are provided in the appendices at the end of that chapter.

CHAPTER 2

Stochastic control of relay channel

2.1 Introduction

In a wireless network, energy efficiency is an important criterion due to battery constraints. Traditionally in a network, nodes communicate directly to the base station. However, the presence of other nodes in the network can lead to more energy efficient systems through node cooperation. This is because the presence of other nodes in the network can provide alternate routes with possibly less transmission energy costs. On the other hand, this also increases the delay in the system as such cooperation requires successful transmission from the source to the relay node, and then from the relay node to the destination node. Thus there is a tradeoff between the energy cost for successfully routing a packet and the corresponding delay cost. The study of this tradeoff in the case of cooperative or strategic users with decentralized information can lead to interesting insights for the design of future cooperative communication systems.

The relay channel is the simplest model and a building block for user cooperation in a network. It has been and is currently being studied extensively from the perspective of information theory (see for instance [10, 12, 54] and references therein), where theoretically achievable rates and practically implementable codes are investigated. Since information theory is an asymptotic theory, it does not capture directly the delay requirements that are important for many communication applications. In addition, information theoretic formulations cannot capture the dynamical aspect of a relay network that may be crucial when studying the behavior of higher layers in the network hierarchy. Finally, since in practice wireless devices are operated by humans, selfish behavior needs to be taken into account for cooperation to be successful.

Recently game theory has been used as a tool to study strategic behavior of nodes participating in a communication network (see [20, 22, 29, 36, 37, 39, 47, 48, 53, 61] and references therein). The work in [61] studies a source-relay channel with non-cooperative

nodes in fading and non-fading channel as finite and infinite repeated games, respectively. The works in [37, 47] propose schemes for multi-hop routing based on a reputation system to punish non-cooperative nodes in wireless ad hoc networks. Evolutionary game theory is used in [29] to punish selfish nodes that do not cooperate to forward packets. The works in [20, 22, 36, 48] adopt pricing mechanisms in wireless ad hoc networks to foster cooperation among non-cooperative nodes. Two auction mechanisms for a relay network are proposed in [20] (SNR auction and power auction) that determine relay selection and relay power allocation in a distributed fashion. The research reported in [22] introduces a "bribery" mechanism that fosters cooperation using a microeconomic framework based on game theory that encourages forwarding among selfish nodes by reimbursing forwarding. Finally, [53] proposes a distributed and scalable acceptance algorithm for nodes to decide whether to accept or reject a relay request. Their algorithm results in a Nash equilibrium, and it is proven that the system converges to the rational and optimal operating point.

Most of the works [20, 22, 29, 36, 48, 53, 61] assume that sources always have data to send, and thus do not consider random traffic arrival processes at a node as may be the case in a network. Others [29, 53, 61] model the communication as repeated games (for example iterative prisoner's dilemma where Tit-for-Tat strategy induces cooperative behavior), whereas [20,22,36,48] use prices to incentivize nodes to cooperate. This chapter considers independent random arrival processes at the source and the relay node which results in a dynamic system (either a dynamic team for cooperative nodes, or a dynamic game [45] for strategic users).

In this chapter the MAC layer of the relay channel is studied as a stochastic control problem of optimally routing packets to the destination. The model assumes a half duplex relay channel with a source, a relay and a destination node with incoming traffic at both the source and the relay node. Also it assumes generic cost functions at time t reflecting packet delay (through the backlog in each agent's queue) and transmission energy. Among other possible formulations, stochastic control of a relay channel can be studied as a static or a dynamic optimization problem, or when information (such as queue backlog) is centralized or decentralized, or when users are non-strategic (cooperative) or strategic, or with complete or incomplete information (of utilities and system parameters). This chapter focuses on complete-information of utilities and system parameters and studies two cases:

(a) dynamic, non-strategic (cooperative) players with decentralized information (Section 2.3);

(b) dynamic, strategic players with centralized information (Section 2.4).

In Section 2.3 the relay channel is studied in the case of cooperative users. The case

where information such as the backlog of each user is known to everybody (the centralized problem) is well known, and can be formulated and solved as a standard Markov Decision Process (MDP) problem. The case is more interesting when information about agents' backlogs is not available to everyone. This problem is studied in Section 2.3.2, where it is formulated as a decentralized dynamic team problem, where users cooperate to minimize expected energy and average delay of the entire network. The key feature of the model is the non-classical information structure, that is, the source and the relay nodes do not observe each other's queue lengths, but through feedback, learn each other's previous actions. Since there is no single controller, rather both source and relay nodes are controllers with linked stochastic control problems, this set-up does not fit into the standard framework of MDP theory [6, 31]. Similar decentralized control problems were studied in [35, 42, 44]. Utilizing an approach similar to [42] of viewing the decentralized system from the point of view of a fictitious coordinator, a structural result is proven which shows that there exists an optimal policy that is the solution of a Bellman-type fixed point equation where the optimization is done over a fixed state space as opposed to an ever-increasing state-space in general. The contribution in this part is to show that the optimal decentralized control strategies are based on the pair of marginal distributions of the queue lengths that the source and relay agents are storing and updating in real time. Inspired by the optimal solution found above, suboptimal decentralized strategies are investigated and their performance is compared (using numerical analysis and simulation) to standard transmission strategies such as time-division multiple access (TDMA) and other random access (RA) protocols.

In Section 2.4 this communication setup is studied with the assumption that users are strategic in nature and the following question is asked: can the socially optimal policy (obtained by a centralized controller) be implemented by strategic users. The relay channel was also studied for strategic users in [49]. In that paper, authors pose the problem as a static game, cooperation is induced using a reward mechanism, and strategies are analyzed in Nash equilibrium. The present work takes into account the dynamic aspect of the problem and poses it as a sequential game of complete information and simultaneous moves [45]. The starting point here is the observation that when users are strategic, there are more than one equilibria that may not coincide with the socially optimum solution. The contribution in this part is to show (through an explicit construction) that the network designer can impose payments to be exchanged between source and relay nodes, such that the resulting dynamic game has the social optimal policy as the unique subgame perfect equilibrium.

The two parts of the chapter broadly address the issue of stochastic control of the relay

channel, and have a dynamic flavor due to the random incoming traffic model. This induces the solution concept for both cases to be within the sequential framework: dynamic programming for the former and subgame perfect equilibrium for the latter.

The remainder of this chapter is structured as follows. Section 2.2 presents the model. In Section 2.3, the team problem is studied when users are cooperative and minimize a common cost criterion. The centralized and decentralized problems are posed and structural results are given for the decentralized problem in Section 2.3.2. In Section 2.3.3, numerical results are presented comparing the performance of a suboptimal decentralized policy with TDMA and RA. In Section 2.4, the case when users are strategic is discussed. Section 2.5 presents the conclusion.

2.2 Model

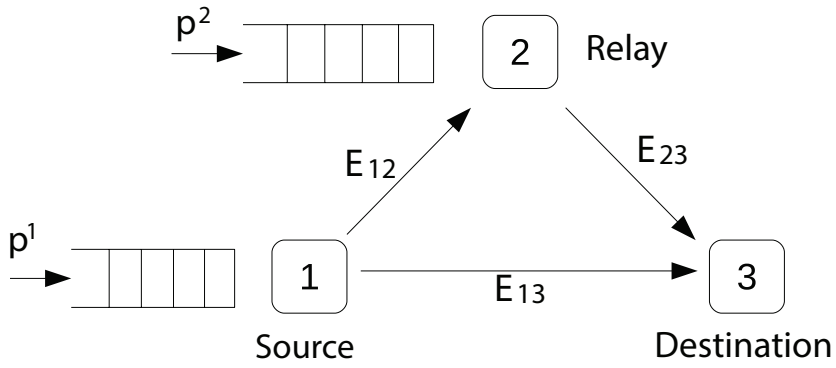


Figure 2.1: A simple relay channel model with simultaneous incoming traffic to source and relay.

The model of the system studied in this chapter, as shown in Fig. 2.1, consists of a source node (node 1), a relay node (node 2) and a destination node (node 3). The time is discretized into slots and Bernoulli[1] packet arrival processes $\{P_t^i\}_{t=1}^{\infty},\ i = 1, 2$ are assumed at node i, with the probability of arrival of a packet in any slot being $p^i \in [0, 1],\ i = 1, 2$. This model can be considered as a prototype for a larger network, where each source node can also act as a relay for other source nodes, thereby potentially minimizing total cost in the network. This view justifies the above assumption of independent arrival processes. Both nodes 1 and 2 have queues of size N. The number of packets at time t in the queue of node i is denoted by $x_t^i,\ i = 1, 2$. The source has to send the packets in its queue to the destination, and it has a choice to either transmit them directly to the destination or transmit

[1]In Section 2.3.2 we discuss extensions for the case of first-order Markov arrival processes.

them through the relay or not transmit at all. At time $t \in \{1, 2, \ldots\}$, node $i,\ i = 1, 2$ takes action u_t^i, as a function of all the information gathered till time t. In subsequent sections we will study different scenarios with different available information to the agents at time t. The possible actions for node 1 are wait (W), transmit to node 2 (T_{12}) and transmit to node 3 (T_{13}); and possible actions for node 2 are wait (W) and transmit to node 3 (T_{23}), thus having

$$u_t^1 \in \mathcal{U}^1 = \{W, T_{12}, T_{13}\}, \tag{2.1a}$$

$$u_t^2 \in \mathcal{U}^2 = \{W, T_{23}\}. \tag{2.1b}$$

It is assumed that simultaneous transmissions from both node 1 and node 2 lead to unsuccessful reception (collision) at the receiver. It is further assumed that even at the event of a collision the packet headers can be decoded at node 3[2]. Under these assumptions, the system evolution can be described by the following set of equations, where the queue length of a node at time t is given by the minimum of (a) its queue size N and (b) its queue length at time $t-1$ plus 1 if there was an arrival in time slot t, minus 1 if a packet was successfully transmitted in $t-1$ slot

$$x_t^1 = \min\left\{x_{t-1}^1 + p_t^1 - \mathbf{1}_{\{T_{12},T_{13}\}}(u_{t-1}^1)\mathbf{1}_{\{W\}}(u_{t-1}^2),\ N\right\}, \tag{2.2a}$$

$$x_t^2 = \min\left\{x_{t-1}^2 + p_t^2 - \mathbf{1}_{\{T_{23}\}}(u_{t-1}^2)\mathbf{1}_{\{W\}}(u_{t-1}^1) + \mathbf{1}_{\{T_{12}\}}(u_{t-1}^1)\mathbf{1}_{\{W\}}(u_{t-1}^2),\ N\right\}, \tag{2.2b}$$

for $t \in \{2, 3, \ldots\}$, where $\mathbf{1}_A(\cdot)$ is the indicator function of the set A.

At the end of time slot t, nodes 1 and 2 receive noiseless feedback $w_t \in \{0, 1, 2, e_1, e_2\}$ from the destination node stating if the destination node successfully received the transmission from node 1 ($w_t = 1$); if the destination node successfully received the transmission from node 2 ($w_t = 2$); if the destination node didn't receive any transmission destined to it ($w_t = 0$); if there was a collision with two packets destined to it ($w_t = e_1$); or, finally, if there was a collision at node 3 due to a simultaneous transmission from node 1 to 2 and node 2 to 3 ($w_t = e_2$). Thus each node $i = 1, 2$ at time t can determine $u_{t-1} = (u_{t-1}^1, u_{t-1}^2)$ from the feedback w_t and it own action u_t^i. This implies that part of the system information (in this case the agents' actions) is shared between the agents with unit delay, while information about the queue lengths is not shared (please refer to [42] for a general discussion on delay-shared patterns). Throughout this chapter, it is also assumed that all controllers have perfect recall.

[2]This is possible with sufficiently strong error correction coding of the headers.

Generic instantaneous cost functions $g_t^i(x_t, u_t)$ are defined for node $i = 1, 2$ as functions of queue lengths $x_t = (x_t^1, x_t^2)$ and actions u_t of both the nodes. To quantify the energy-delay tradeoff the following costs are assumed. The energy cost of transmissions are defined by functions $e^1 : \mathcal{U}^1 \to \{0, E_{12}, E_{13}\}, e^2 : \mathcal{U}^2 \to \{0, E_{23}\}$ such that energy cost for transmission from node 1 to node 3, $e^1(T_{13}) = E_{13}$, node 1 to node 2 is $e^1(T_{12}) = E_{12}$, and that for node 2 to node 3 is $e^2(T_{23}) = E_{23}$. The energy cost for waiting is 0, i.e. $e^1(W) = e^2(W) = 0$. The instantaneous delay cost at time t for node $i = 1, 2$ is assumed to be equal to the total number of packets waiting in the queue of node i plus cost c_d for each dropped packet. One instance of such cost function that captures both energy and average delay is $g_t^i(x_t, u_t) = x_t^i + e^i(u_t^i) + p^i c_d \mathbf{1}_{A^i}(x_t, u_t)$, where A^1, A^2 represent the events that the next arrived packet is dropped for node 1 and 2 respectively,

$$\begin{aligned} A^1 = \ & \{x_t^1 = N, u_t \in \{WW, WT_{23}, T_{12}T_{23}, T_{13}T_{23}\}\} \\ & \bigcup \{x_t^1 = N, x_t^2 = N, u_t = T_{12}W\}, \\ A^2 = \ & \{x_t^2 = N - 1, u_t = T_{12}W\} \\ & \bigcup \{x_t^2 = N, u_t \in \{WW, T_{12}W, T_{12}T_{23}, T_{13}W, T_{13}T_{23}\}\}. \end{aligned}$$

All costs are additive and costs for future slots (or epochs) are discounted by a discount factor $\lambda, (0 < \lambda < 1)$. The tuple $(p^1, p^2, E_{13}, E_{23}, E_{12}, \lambda, N, c_d)$ summarizes the basic parameters of the system.

2.3 Cooperative users

In this section the team problem is studied, under the assumption that both nodes act cooperatively i.e. have the same objective. In particular, in Section 2.3.1 the centralized problem is defined, where there is a single controller observing all relevant information and has perfect recall. The solution of this problem is not studied in this chapter, since it is well-known and can be found as the solution of a dynamic program, either analytically or numerically. The result is briefly stated for completeness, and since it serves as the baseline solution with which all other solutions will be compared. In Section 2.3.2 the decentralized problem is discussed, where there are two controllers with different information sets, though with perfect recall. In this case, the nodes cannot observe each other's queues. Their own queue history is their private information, and feedback history is the common information. Decentralized problems with non-classical information structure are notoriously hard [59]. One of the contributions of this work is to show that the optimum strategy can be found as

solution of a dynamic program over a large but time-invariant state space. Finally, numerical results are presented in Section 2.3.3 that compare the performance of a suboptimal decentralized policy (inspired by the optimal one found in Section 2.3.2) with TDMA and RA policies.

2.3.1 Centralized control problem

At time t, the common knowledge of nodes 1 and 2 (or centralized controller) is $u_{1:t-1}, x_{1:t}$. Thus the control action at time t, $u_t \in \mathcal{U}^1 \times \mathcal{U}^2$, can (in general) be a function of all the information available till that time

$$u_t = \hat{\psi}_t(x_{1:t}, u_{1:t-1}) = \psi_t(x_{1:t}). \tag{2.3}$$

A given policy $\psi = (\psi_1, \psi_2, \psi_3, \ldots)$ induces a total discounted cost over the horizon T equal to

$$J^\psi := \mathbb{E}^\psi \left[\sum_{t=1}^{T} \lambda^{t-1} \left\{ g_t^1(X_t, U_t) + g_t^2(X_t, U_t) \right\} \right]. \tag{2.4}$$

The centralized problem is defined as follows

Problem 2.1. Find the centralized policy ψ^* that achieves the optimum cost,

$$J^* := \min_{\psi} J^\psi, \tag{2.5}$$

with J^ψ defined in (2.4), system update equation given by (2.2) and control actions u_t as in (2.3).

The solution of this problem is well-known and hinges on the following lemma.

Lemma 2.1. The process $\{(X_t, U_t);\ t = 0, 1, \ldots\}$ is a controlled Markov process with state X_t, control U_t, and instantaneous cost $g_t^1(X_t, U_t) + g_t^2(X_t, U_t)$, i.e.,

$$P(x_{t+1}|x_{1:t}, u_{1:t}) = P(x_{t+1}|x_t, u_t) \tag{2.6}$$

Proof. This is trivially true due to the system evolution given in (2.2), the independence of the basic random variables $(X_1^1, X_1^2, P_1^1, P_1^2, P_2^1, P_2^2, \ldots)$ and the instantaneous cost being a function of only the current state and action pair (X_t, U_t).

Thus by the theory of MDPs [6, 31], there exists a Markov policy of the form $u_t = (u_t^1, u_t^2) = \psi_t^*(x_t)$ that achieves the optimum cost J^* in (2.5). Moreover this optimal cost

can be found using dynamic programming. The reader is referred to the authors' report [55] for more details about this centralized problem and its solution.

2.3.2 Decentralized control problem

In this section, a more practical case is considered whereby users cannot observe each other's queues. This is an instance of decentralized information as the information sets of the two nodes are not the same.

At time t, information available to node k is $(x^k_{1:t}, u^k_{1:t-1}, w_{1:t-1})$ which is equivalent to $(x^k_{1:t}, u_{1:t-1})$, since as mentioned earlier, knowledge of one's own actions and the feedback reveals the actions of the other node. If ϕ^k_t is a decentralized action policy of node k at time t, then control actions can be defined as follows

$$u^k_t = \hat{\phi}^k_t(x^k_{1:t}, u^k_{1:t-1}, w_{1:t-1}) = \phi^k_t(x^k_{1:t}, u_{1:t-1}), \quad k = 1, 2. \tag{2.7}$$

In the remaining of this section, the infinite horizon problem is considered for expositional simplicity (the proposed solution also applies to the finite-horizon case). The instantaneous cost functions are assumed time invariant, i.e., $g^k_t = g^k$. Let the combined cost be defined as $g(x_t, u_t) := g^1(x_t, u_t) + g^2(x_t, u_t)$. If ϕ^k is any strategy of node k i.e., $\phi^k = (\phi^k_1, \phi^k_2, \ldots)$ where $k \in \{1, 2\}$ then $\phi = (\phi^1, \phi^2)$ is the combined strategy of both the nodes and the corresponding discounted cost J^ϕ is given by

$$J^\phi = \mathbb{E}^\phi\{\sum_{t=1}^{\infty} \lambda^{t-1} g(X_t, U_t)\}. \tag{2.8}$$

The decentralized control problem can now be stated as

Problem 2.2. Find the optimal decentralized policy ϕ^* that achieves the optimal cost

$$J^* := \inf_{\phi} J^\phi \tag{2.9}$$

with J^ϕ defined in (2.8), system update equation given by (2.2), and control actions u_t as in (2.7).

The controls actions, as given in (2.7), are functions of an ever increasing space. In this section, we seek to simplify the domain of these functions to a succinct, fixed space. To that effect, a structural result is proved for the optimal decentralized policy, and shown that it can be found as a solution of a Bellman-type fixed-point equation. It is first proven that there exist optimal control actions of a node that depend only on its current queue length and

the entire control history of both the nodes i.e., $(x_t^k, u_{1:t-1})$. In the second simplification step, it is shown that there exists an optimal policy that depends on the current queue length x_t^k and the posterior on x_t conditioned on the control history $u_{1:t-1}$. In the final simplification step it is shown that the aforementioned posterior distribution on x_t can be substituted by the pair of marginal distributions over x_t^k for $k = 1, 2$.

In this decentralized case, at time t, $x_{1:t}^k$ is the private information of node k and $u_{1:t-1}$ is the common information available to both nodes. The following lemma proves that given the common information, the private information of the two nodes is independent.

Lemma 2.2. For any fixed strategy $\boldsymbol{\phi}$, random variables $X_{1:t}^1$ and $X_{1:t}^2$ are conditionally independent given the control history till time t, $U_{1:t-1}$ i.e.,

$$P^{\boldsymbol{\phi}}(x_{1:t}|u_{1:t-1}) = P^{\boldsymbol{\phi}^1}(x_{1:t}^1|u_{1:t-1}) P^{\boldsymbol{\phi}^2}(x_{1:t}^2|u_{1:t-1}) \tag{2.10}$$

Proof. The causal decomposition of $P^{\boldsymbol{\phi}}(x_{1:t}, u_{1:t-1})$ gives,

$$\begin{aligned} P^{\boldsymbol{\phi}}(x_{1:t}, u_{1:t-1}) =& P(x_1^1) \prod_{i=1}^{t-1} P(x_{i+1}^1|x_i^1, u_i) P^{\boldsymbol{\phi}^1}(u_i^1|x_{1:i}^1, u_{1:i-1}) \\ & \times P(x_1^2) \prod_{j=1}^{t-1} P(x_{j+1}^2|x_j^2, u_j) P^{\boldsymbol{\phi}^2}(u_j^2|x_{1:j}^2, u_{1:j-1}) \end{aligned} \tag{2.11a}$$

Thus,

$$\begin{aligned} P^{\boldsymbol{\phi}}(x_{1:t}|u_{1:t-1}) =& \frac{P(x_1^1) \prod_{i=1}^{t-1} P(x_{i+1}^1|x_i^1, u_i) P^{\boldsymbol{\phi}^1}(u_i^1|x_{1:i}^1, u_{1:i-1})}{\sum_{x_{1:t}^1} P(x_1^1) \prod_{i=1}^{t-1} P(x_{i+1}^1|x_i^1, u_i) P^{\boldsymbol{\phi}^1}(u_i^1|x_{1:i}^1, u_{1:i-1})} \\ & \times \frac{P(x_1^2) \prod_{j=1}^{t-1} P(x_{j+1}^2|x_j^2, u_j) P^{\boldsymbol{\phi}^2}(u_j^2|x_{1:j}^2, u_{1:j-1})}{\sum_{x_{1:t}^2} P(x_1^2) \prod_{j=1}^{t-1} P(x_{j+1}^2|x_j^2, u_j) P^{\boldsymbol{\phi}^2}(u_j^2|x_{1:j}^2, u_{1:j-1})} \end{aligned} \tag{2.11b}$$

$$= P^{\boldsymbol{\phi}^1}(x_{1:t}^1|u_{1:t-1}) P^{\boldsymbol{\phi}^2}(x_{1:t}^2|u_{1:t-1}). \tag{2.11c}$$

We now proceed to show that each node can summarize its private information to only the current queue state without loss of optimality. For this the following lemma is required.

Lemma 2.3. For any given fixed strategy $\boldsymbol{\phi}^2$ of node 2, the process $\{(X_t^1, U_{1:t-1}, U_t^1); t = 1, 2, \ldots\}$ is a controlled Markov process with state $(X_t^1, U_{1:t-1})$ and control input U_t^1 i.e.,

$$P^{\phi^2}(x^1_{t+1}, u_{1:t}|x^1_{1:t}, u_{1:t-1}, u^1_{1:t}) = P^{\phi^2}(x^1_{t+1}, u_{1:t}|x^1_t, u_{1:t-1}, u^1_t) \tag{2.12}$$

$$\begin{aligned}\mathbb{E}^{\phi^2}\{g(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_{1:t}, u_{1:t-1}, u^1_{1:t}\} &= \mathbb{E}^{\phi^2}\{g(x^1_t, x^2_t, u^1_t, u^2_t)|x^1_t, u_{1:t-1}, u^1_t\} \\ &= \hat{g}(x^1_t, u_{1:t-1}, u^1_t)\end{aligned} \tag{2.13}$$

Proof. See Appendix A

As a consequence of the MDP structure of the problem (given a fixed strategy ϕ^2 of node 2), the optimal policy by node 1 can be a Markov policy [31]. Since this is true for any fixed strategy of node 2, it is also true for the optimal strategy of the node 2. A similar result can be obtained by interchanging the roles of node 1 and 2, thus the optimal decentralized policy can be of the form below without loss of optimality

$$u^k_t = \phi^k_t(x^k_t, u_{1:t-1}), \quad k = 1, 2. \tag{2.14}$$

Even with the above simplification, Problem 2.2 reduces to two linked stochastic control problems for which a solution is not readily available. We proceed to the second simplification step by reexamining this problem from the perspective of a fictitious coordinator [42], who observes, at time t, the feedback w_t or equivalently u_{t-1} (common information) but does not observe $x^k_t, k \in \{1, 2\}$ (private information). This fictitious coordinator (which can be replicated at both nodes) generates partial functions $\gamma_t = \gamma_t^{1:2} = (\gamma^1_t, \gamma^2_t)$ as its control output, where $\gamma^k_t : \mathcal{N} \to \mathcal{U}^k, k \in \{1, 2\}$. Based upon these coordinator control outputs, node k, $k \in \{1, 2\}$ computes its action by operating these partial functions on its private information x^k_t, as shown in Fig. 2.2. In particular, denoting the coordinator strategy at time t by $\Psi_t = (\Psi^1_t, \Psi^2_t)$ we can write

$$(\gamma^1_t, \gamma^2_t) = \Psi_t(u_{1:t-1}) = (\Psi^1_t(u_{1:t-1}), \Psi^2_t(u_{1:t-1})), \tag{2.15}$$

and node k action will be given by

$$u^k_t = \gamma^k_t(x^k_t) \tag{2.16a}$$

$$= \Psi^k_t(u_{1:t-1})(x^k_t) \tag{2.16b}$$

$$= \phi^k_t(x^k_t, u_{1:t-1}). \tag{2.16c}$$

In the following we show that the coordinator strategy can be simplified by summarizing the history of the common information into a sufficient statistic with time-invariant domain.

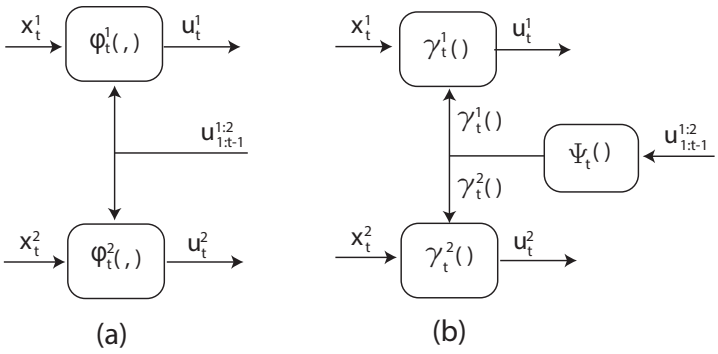


Figure 2.2: Control by the fictitious coordinator: (a) original control action generation; (b) equivalent control action generation through intermediate coordinator actions.

In particular we show that belief on x_t given the observation history $u_{1:t-1}$ and control history $\gamma_{1:t-1}$ till time t, forms a sufficient state for the coordinator's problem. We define the random variable $\Pi_t \in \mathcal{P}(\mathcal{N}^2)$ as the posterior pmf of X_t conditioned on $U_{1:t}, \Gamma_{1:t-1}$ i.e.,

$$\Pi_t(x_t) = P(X_t = x_t | U_{1:t-1}, \Gamma_{1:t-1}). \tag{2.17}$$

The next lemma shows that this quantity can be recursively updated by the coordinator in a deterministic fashion.

Lemma 2.4. There exists a deterministic update function F, independent of the coordinator's policy Ψ , that updates the state π_t given control γ_t and variable u_t .

$$\pi_{t+1} = F(\pi_t, \gamma_t, u_t) \tag{2.18}$$

Proof. See Appendix B.

The following theorem establishes that the coordinator's problem is an MDP.

Theorem 2.1. The process $\{(\Pi_t, \Gamma_t);\ t = 1, 2, ...\}$ is a controlled Markov process with state Π_t and control Γ_t, i.e.,

$$P(\pi_{t+1} | \pi_{1:t}, \gamma_{1:t}) = P(\pi_{t+1} | \pi_t, \gamma_t) \tag{2.19}$$

$$\mathbb{E}(g(x_t, u_t) | \pi_{1:t}, \gamma_{1:t}) = \mathbb{E}(g(x_t, u_t) | \pi_t, \gamma_t) \tag{2.20}$$

$$=: \hat{g}(\pi_t, \gamma_t) \tag{2.21}$$

Proof. See Appendix C

Since $\{(\Pi_t, \Gamma_t);\ t = 1, 2, \dots\}$ is a controlled Markov process the optimal output functions can be given by Markov policies [31] $(\gamma_t^1, \gamma_t^2) = \psi_t(\pi_t)$. Thus optimal action by node

k can be written (with some notational abuse) as

$$u_t^k = \gamma_t^k(x_t) = \psi_t^k(\pi_t)(x_t) = \phi_t^k(x_t^k, \pi_t). \tag{2.22}$$

Furthermore, the optimal actions for the coordinator are minimizers of the fixed-point equation

$$V(\pi) = \inf_{\gamma} [\hat{g}(\pi, \gamma) + \lambda \mathbb{E}\{V(\pi')|\pi, \gamma\}], \tag{2.23}$$

where the expectation is with respect to the conditional probability induced by the update function F and u_t as random variable (noise), in accordance with [42].

Finally, in the remaining of this section we proceed with the last simplification step and show that, due to the specific nature of our problem, instead of the joint probability Π_t on the queue length of the two nodes, individual marginals form a sufficient state. To that effect, we define the random variable $\Xi_t^k \in \mathcal{P}(\mathcal{N})$ as the posterior pmf of X_t^k conditioned on $U_{1:t-1}, \Gamma_{1:t-1}$ i.e.,

$$\Xi_t^k(x_t^k) = P(X_t^k = x_t^k | U_{1:t-1}, \Gamma_{1:t-1}), \quad k = 1, 2 \tag{2.24}$$

and show that $\{(\Xi_t, \Gamma_t);\ t = 1, 2, ...\}$ is controlled Markov process (where $\Xi_t = \Xi_t^{1:2} = (\Xi_t^1, \Xi_t^2)$). This gives a significant reduction in the size of the state space over which the optimal policies are defined, since π is defined over a space of $\mathcal{P}(\mathcal{N}^2)$ while ξ is defined over $\mathcal{P}(\mathcal{N}) \times \mathcal{P}(\mathcal{N})$. For a finite queue length N, $\mathcal{P}(\mathcal{N}^2)$ has dimensionality N^2, and thus grows super exponentially in N as $\mathbb{R}^{N^2}$, whereas $\mathcal{P}(\mathcal{N}) \times \mathcal{P}(\mathcal{N})$ has dimensionality $2N$, and grows exponentially in N as $\mathbb{R}^{2N}$.

The above statement is made precise in the following lemma and subsequent theorem.

Lemma 2.5. There exist deterministic update functions $G^k,\ k \in \{1, 2\}$, independent of the policy Ψ, that update the state ξ_t^k given control γ_t^k and actions u_t as

$$\xi_{t+1}^k = G^k(\xi_t^k, \gamma_t^k, u_t), \qquad k \in \{1, 2\}. \tag{2.25}$$

Proof. See Appendix D.

Theorem 2.2. The process $\{(\Xi_t, \Gamma_t);\ t = 1, 2, ...\}$ is a controlled Markov process with

state Ξ_t and controls Γ_t, i.e.,

$$P^{\phi}(\xi_{t+1}|\xi_{1:t},\gamma_{1:t}) = P(\xi_{t+1}|\xi_t,\gamma_t) \tag{2.26}$$

$$\begin{aligned}\mathbb{E}(g(x_t,u_t)|\xi_{1:t},\gamma_{1:t}) &= \mathbb{E}(g(x_t,u_t)|\xi_t,\gamma_{1:t}) \\ &:= \tilde{g}(\xi_t,\gamma_t)\end{aligned} \tag{2.27}$$

Proof. See Appendix E.

Since $\{(\Xi_t,\Gamma_t); t = 1,2,...\}$ is a controlled Markov process, the optimal output functions can be given by Markov policies $\gamma_t = \psi_t(\xi_t)$, and can be derived as minimizers of the fixed-point equation

$$V(\xi) = \inf_{\gamma}[\tilde{g}(\xi,\gamma) + \lambda\mathbb{E}\{V(\bar{\xi})|\xi,\gamma\}], \tag{2.28}$$

where the expectation is with respect to the conditional probability induced by the update functions (G^1, G^2) and u_t as random variable (noise). In summary, the optimal control actions are of the form

$$u_t^k = \phi_t^k(x_t^k,\xi_t), \quad k = 1,2, \tag{2.29}$$

and in the on-line operation of the system are generated as follows: at time t, each node (source and relay) first updates the quantities ξ_{t-1} as dictated by the recursion (2.25), and based upon ξ_t they find the corresponding action γ_t as dictated by the (off-line) solution of (2.28). Finally they generate their action u_t^k by evaluating γ_t^k on their private information x_t^k i.e., $u_t^k = \gamma_t^k(x_t^k)$.

We conclude this section by observing that in this model we assumed two independent Bernoulli arrival processes for ease of exposition. The analysis can be easily extended to the case of independent first-order Markov arrival processes. This can be achieved by defining $Z_t^k = (X_t^k, P_t^k),\ k = 1,2$, where $\{P_t^k\}_t,\ k = 1,2$ are independent Markov arrival processes of node 1 and 2 respectively. Since in our model both X_t^k, P_t^k are observable, so is Z_t^k; thus similar results can be easily derived for Z_t^k in place of $X_t^k, k \in \{1,2\}$. When arrival processes (P_t^1, P_t^2) follow joint Markov process, the analysis can be extended in a similar way as for two independent Markov processes, till equation (2.23) but the simplification in Theorem 2.2 does not follow.

2.3.3 Numerical results

In this section, we compare the performance of a suboptimal decentralized policy obtained from our analysis in the previous section, with standard TDMA and RA policies (which themselves are decentralized policies). We assume the cost function of the form $g(x_t, u_t) = x_t^1 + x_t^2 + e^1(u_t^1) + e^2(u_t^2)$.

Regarding the decentralized policy, we chose to solve the fixed-point equation (2.23) instead of the simpler one in (2.28). Although (28) is an important theoretic simplification of the state space, the reason for this choice is that equation (2.23) resembles a Bellman-type fixed-point equation for a POMDP (partially observed MDP), and thus can be solved using standard POMDP solvers if the underlying state and action space is finite. However, although equation (2.28) has significantly smaller space of optimization, it cannot be solved using standard POMDP solvers as it does not have the linear structure required by standard POMDP solvers.

For a finite maximum queue length N, there are $|\mathcal{U}^1|^N \cdot |\mathcal{U}^2|^N$ possible γ functions. We expect the optimum $\gamma^{1:2,*}$ functions to be of "threshold-type" [17, 51], i.e. the domain $\{0, 1, \dots, N\}$ of each γ^i to be partitioned into contiguous regions, one corresponding to each action in $\mathcal{U}^i$. The threshold nature of policies is proved for queueing and other problems in [7, 17] and generally requires proving properties like concavity, supermodularity, superconcavity etc. of the cost-to-go function. It is usually hard to show that some set of properties of the cost-to-go function propagate through the dynamic programming equation, more so for POMDPs where state space is a connected simplex. We do not prove the optimality of threshold policies, however, inspired by the solution of centralized policies, we find a suboptimal decentralized policy by solving (2.23) over threshold policies defined as all possible policies of the following form parametrized by (α, β, θ). We call this policy ψ^*.

$$\gamma^1(x^1) = \begin{cases} W, & x^1 < \alpha \\ T_{12}, & \alpha \le x^1 < \beta \\ T_{13}, & \beta \le x^1 \end{cases} \tag{2.30a}$$

$$\gamma^2(x^2) = \begin{cases} W, & x^2 < \theta \\ T_{23}, & \theta \le x^2. \end{cases} \tag{2.30b}$$

For the numerical analysis, we choose $N = 2$, which is a compromise between accuracy of results and complexity of solving the fixed-point equation. For a given policy

that results in a system where queues are expected to get empty frequently (i.e., $(0,0)$ is expected to be a recurrent state), the stationary distribution of the system is expected to have significant weight around the queue sizes $(0,0)$ and negligible weight for larger queue lengths. Thus such a small value for N is a justified approximation for lightly loaded systems. We numerically solve equation (2.23) over the set of threshold policies using a POMDP solver for $E_{12} = 0.1, E_{23} = 0.2, E_{13} = 10$ and $p^1, \in \{0.1, 0.2, 0.4\}, p^2 \in \{0.01, 0.1, 0.2, 0.4, 0.6\}, N = 2, c_d = 5, \lambda = 0.99$. Figures 2.3 compare their costs for different traffic parameters.

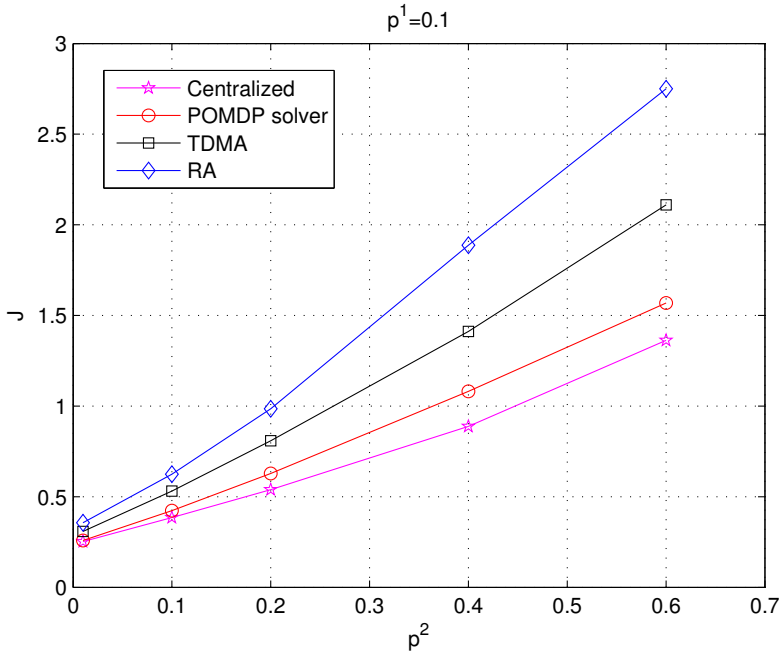
Both TDMA and RA policies prescribe allocation of slots to users. In each allocated slot, a user chooses the optimal action that minimizes the cost to go. For example when the slot is allocated to the source, it takes one of the following actions $\{W, T_{12}, T_{13}\}$ that is optimal for the current state. The numerical results are obtained by numerically analyzing the steady state distribution of the Markov process induced by implementing the optimal policy, which is found by enumerating all threshold policies. The TDMA policy assumes that users access the channel as dictated by a commonly observed binary random variable, the distribution of which is optimized for each pair of (p^1, p^2). As a result the channel is not allocated equally to source and relay, but optimally as dictated by their traffic loads. The RA policy assumes that users access the channel as dictated by private random back-off times whose distribution is geometric and optimized for their traffic loads. The corresponding results are obtained by numerically analyzing the steady state distribution of the induced Markov process. In both TDMA and RA, when a slot is available to a node, that node plays its optimal action. Thus for e.g. in TDMA, relaying is achieved as follows: when node 1 is given a slot, node 1 sends a packet to node 2, and when node 2 is given a slot, node 2 it sends its packet to the receiver. Finally, the solution of the centralized problem mentioned in Section 2.3.1 is shown, which serves as a lower bound for all decentralized policies.

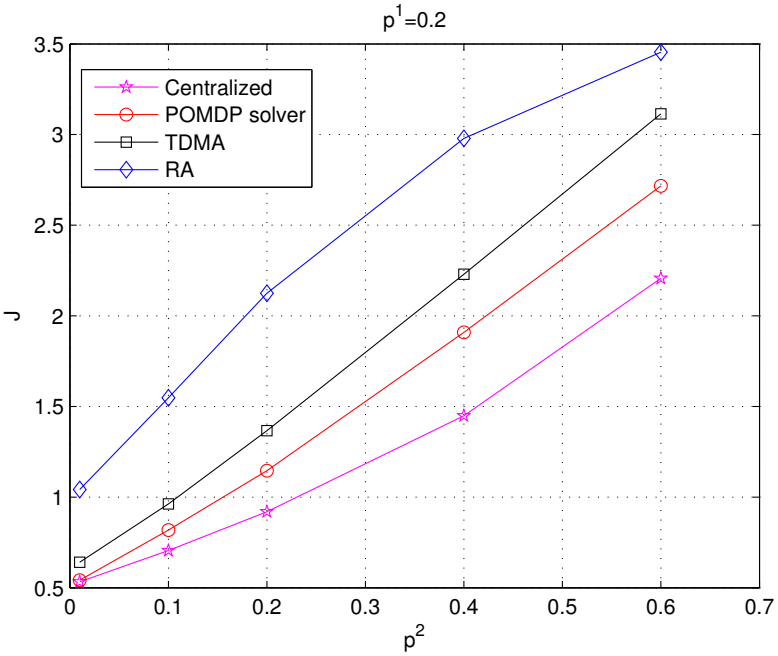
We make the following observations regarding the results shown in Fig. 2.3. The decentralized policy ψ^* outperforms both RA and TDMA policies for all given $p^1,\ p^2$. To take an instance, for $p^1 = 0.1,\ p^2 = 0.1$, the costs obtained for policies ψ^*, RA, TDMA and optimum centralized are $0.4230,\ 0.6241,\ 0.5307,\ 0.3845$, respectively.

2.4 Strategic users

This section considers the case when users are strategic i.e., when they have potentially different cost criteria, and the information is perfect and complete[3]. For problems with

[3]This is a preliminary analysis of the system with strategic agents. We develop tools in the next chapter to relax the condition of perfect information and study dynamic games with asymmetric information; however,


p1=0.1
Centralized
POMDP solver
TDMA
RA
J
p2
0
0.5
1
1.5
2
2.5
3
0.1
0.2
0.3
0.4
0.5
0.6
0.7


p1=0.2
Centralized
POMDP solver
TDMA
RA
J
p2
0.5
1
1.5
2
2.5
3
3.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

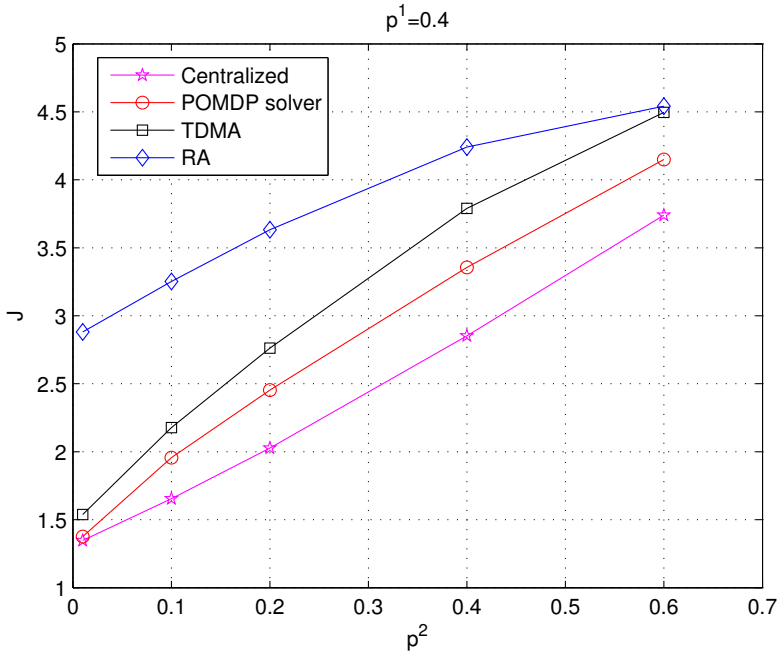


Figure 2.3: Numerical and simulation results for the decentralized policy, TDMA, and RA. The baseline case (centralized solution) is also shown for comparison.

multiple agents as decision makers, having their individual objective functions, equilibrium is an appropriate solution concept [45], where equilibrium is loosely defined as set of policies for each agent such that no agent has an incentive to unilaterally deviate from its policy.

With strategic nodes in our previous model, both nodes have linked stochastic processes such that their actions affect each other's current and future costs, so they determine their actions in order to minimize their own average total cost over the given time horizon. In this section we consider a centralized information model where all system variables are perfectly observed by both agents (as was the case in Section 2.3.1). It is also assumed that cost functions and system parameters are common knowledge. In this case, the system evolves as a dynamic game of perfect information and simultaneous moves [45]. In such a case, the relay node may not want to accept packets from the source node as that will result in larger delay (longer backlog for his own queue) as well as extra transmission energy. Thus, the socially optimal solution may not be implementable in an equilibrium concept. One possible way to induce relay cooperation is for the source to pay the relay node for relaying the packet (and similarly, for the relay to pay the source for backing

we do not do mechanism or incentive design. The problem with imperfect or incomplete information, is an interesting problem that comes under the purview of dynamic mechanism design for games with asymmetric information, and is not considered in this thesis.

off from transmission so that it can transmit its own packets). This entails the question of whether, through such payments, an equilibrium solution would lead to the socially optimum solution.

The main result in this section is that there exist payment transfers to be imposed by the network designer such that the optimum Markov policy of the centralized problem (as discussed in Section 2.3.1) is also the unique subgame perfect equilibrium of the dynamic game.

The model for the dynamic game with finite horizon T is now formalized. When nodes are strategic, it is intuitive to think that the relay node would not want to accept source's packets since that increases it's queue length. Thus the strategic users may never achieve socially optimal cost when (T_{12}, W) is an optimum action. To capture this behavior of node 2, we enhance its available actions as

$$u_t^2 \in \mathcal{U}^2 = \{W_a, W_r, T_{23}\}, \tag{2.31}$$

where possible actions for node 2 are wait while accepting the packet from node 1 (W_a), wait while rejecting the packet (W_r), or transmit to node 3 (T_{23}). The queue length of a node at time t is given by the minimum of (a) its queue size N and (b) its queue length at time $t-1$ plus 1 if there was an arrival in time slot t , minus 1 if a packet was successfully transmitted in $t-1$ slot, i.e.,

$$\begin{aligned} x_t^1 = &\min\big\{p_t^1 + [x_{t-1}^1 - \mathbf{1}_{\{T_{12},T_{13}\}}(u_{t-1}^1)\mathbf{1}_{\{W_a\}}(u_{t-1}^2) \\ &-\mathbf{1}_{\{T_{13}\}}(u_{t-1}^1)\mathbf{1}_{\{W_r\}}(u_{t-1}^2)]^+,\ N\big\}, \end{aligned} \tag{2.32a}$$

$$\begin{aligned} x_t^2 = &\min\big\{p_t^2 + [x_{t-1}^2 - \mathbf{1}_{\{T_{23}\}}(u_{t-1}^2)\mathbf{1}_{\{W\}}(u_{t-1}^1) \\ &+\mathbf{1}_{\{T_{12}\}}(u_{t-1}^1)\mathbf{1}_{\{W_a\}}(u_{t-1}^2)]^+,\ N\big\}, \end{aligned} \tag{2.32b}$$

with the meaning of these variables being exactly the same as in (2.2). Let $\mathcal{H}$ denote the set of all histories that are all possible sequences of actions taken by the nodes and nature (arrival processes $p_t = (p_t^1, p_t^2) \in \{0,1\}^2$ are viewed as actions by nature) i.e.,

$$\mathcal{H} := \{h_t = (p_{1:t}, u_{1:t-1}, x_{1:t}) | \quad \forall t \in \{0, 1, \dots T\}\}, \tag{2.33}$$

where state sequences have also been added to the histories since they can be determined through (2.32) by the action and arrival histories. With $N = 2$ users, set of histories $\mathcal{H}$, generic instantaneous cost functions $g_t^i(x_t, u_t),\ i = 1, 2$, and simultaneous actions taken at each time t based on the history h_t, a dynamic game of perfect information with simultaneous moves is defined [4, 45], henceforth referred to as $\mathcal{G}$. The goal is to design a payment

exchange mechanism[4] such that the optimal solution of the centralized problem discussed in Section 2.3.1 is achieved as the subgame perfect equilibrium of the dynamic game. To that effect functions $C_t : \mathcal{H} \times \mathcal{U}^1 \times \mathcal{U}^2 \to \mathbb{R}$ are defined as the payment made[5] by node 1 to node 2 at time t. With these payment exchanges, let the new game $\tilde{\mathcal{G}}$, be defined as before with instantaneous cost functions

$$\hat{g}_t^1(h, u_t) = g_t^1(x_t, u_t) + C_t(h, u_t), \tag{2.34a}$$
$$\hat{g}_t^2(h, u_t) = g_t^2(x_t, u_t) - C_t(h, u_t), \tag{2.34b}$$

for node 1 and 2 respectively. Note that although the payment functions are history dependent, and thus their domain is time-varyting, it will be shown that due to the structure of this problem, they need only be state dependent.

It is easy to see that with so many degrees of freedom, a game can be constructed such that socially optimum solution is a (not necessarily unique) subgame perfect equilibrium of the game as shown by the following construction. Indeed, for $\alpha \in (0, 1)$, let payments be designed as

$$C_t(h, u) = C_t(x_t, u) = \alpha g_t^2(x_t, u) - (1 - \alpha) g_t^1(x_t, u).$$

Then the instantaneous costs of user 1 and 2 are

$$\hat{g}_t^1(h, u_t) = \alpha \left(g_t^1(x_t, u) + g_t^2(x_t, u) \right) \tag{2.35a}$$
$$\hat{g}_t^2(h, u_t) = (1 - \alpha) \left(g_t^1(x_t, u) + g_t^2(x_t, u) \right). \tag{2.35b}$$

In this case, the objectives of both users are aligned with the social objective and the socially optimal solution is also a Nash Equilibrium. The problem with this construction is that, in general, there may exist additional Nash equilibria that may not achieve socially optimal solution.

In the remaining part of this section, a new construction of payment function is provided and sufficient conditions are found for the socially optimal strategy to be the unique subgame perfect Nash equilibrium of game $\tilde{\mathcal{G}}$. In the following lemma, payments are constructed for a static game, and later this idea is extended to dynamic games.

Lemma 2.6. Consider a strategic game of two players with finite action sets $(\mathcal{U}^i)_{i=1,2}$ and costs $(g^i)_{i=1,2}$. Fix an action profile $(a, b) \in \mathcal{U}^1 \times \mathcal{U}^2$ (not necessarilly an equillibrium).

[4]Note that this is not a "mechanism" in the sense of Implementation Theory [9] since utilities and system parameters are common knowledge.

[5]This payment can be positive or negative, as considered in many other problems [15, 21, 23, 27, 28, 50].

There exists a payment function $C : \mathcal{U}^1 \times \mathcal{U}^2 \to \mathbb{R}$ such that (a, b) is the unique strictly dominant strategy equilibrium of the new strategic game with the updated costs.

Proof. For (a, b) to be the strictly dominant strategy equilibrium, the following conditions need to be satisfied

$$g^1(a, u^2) + C(a, u^2) < g^1(u^1, u^2) + C(u^1, u^2) \qquad \forall u^1 \in \mathcal{U}^1 \backslash \{a\}, u^2 \in \mathcal{U}^2 \tag{2.36a}$$

$$g^2(u^1, b) - C(u^1, b) < g^2(u^1, u^2) - C(u^1, u^2) \qquad \forall u^1 \in \mathcal{U}^1, u^2 \in \mathcal{U}^2 \backslash \{b\}. \tag{2.36b}$$

Let

$$r^1(u^1, u^2) := g^1(u^1, u^2) - g^1(a, u^2) \tag{2.37a}$$

$$r^2(u^1, u^2) := g^2(u^1, u^2) - g^2(u^1, b) \tag{2.37b}$$

Equations (2.36) for $u^2 = b, u^1 = a$ are equivalent to, respectively,

$$C(a, b) < r^1(u^1, b) + C(u^1, b) \qquad \forall u^1 \in \mathcal{U}^1 \backslash \{a\} \tag{2.38a}$$

$$C(a, u^2) < r^2(a, u^2) + C(a, b) \qquad \forall u^2 \in \mathcal{U}^2 \backslash \{b\} \tag{2.38b}$$

and for $u^1 \neq a, u^2 \neq b$, are equivalent to

$$C(a, u^2) - r^1(u^1, u^2) < C(u^1, u^2) < C(u^1, b) + r^2(u^1, u^2) \qquad \forall u^1 \in \mathcal{U}^1 \backslash \{a\}, u^2 \in \mathcal{U}^2 \backslash \{b\}. \tag{2.39}$$

Since action sets are finite, there exists payments $C(u^1, u^2)$ with $C(u^1, b)$ large enough and $C(a, u^2)$ small enough such that inequalities (2.38)-(2.39) are satisfied.

The above lemma shows that the cost variables $\{C(u^1, u^2), (u^1, u^2) \in \mathcal{U}^1 \times \mathcal{U}^2\}$ need to satisfy linear inequalities (2.38)-(2.39), i.e. can be chosen from a polytope that is always non-empty. A feasible solution can be found by first applying a vertex enumeration algorithm [2] which finds the vertices, rays and linearities of the polytope described by the linear inequalities (2.38)-(2.39), and then an arbitrary point is selected as the sum of a convex combination of the vertices, a conic combination of the rays and a linear combination of the linearities.

Note that the above lemma provides a very strong type of equilibrium design, i.e., strictly dominant strategy equilibrium. Also, the above construction of the game allows for

negative payments by either user. Even though negative payments are commonly used to induce users to behave in a way desired by the designer (see for e.g. [15,21,26,27,50]), the above construction can be proved with positive payments only[6].

This idea is now extended to dynamic games in the following theorem.

Theorem 2.3. Fix a Markov policy $\psi^* = (\psi_t^*)_{t=1}^T$ for the original centralized stochastic control problem (not necessarily the optimal one). Consider the dynamic game $\mathcal{G}$ defined above. There exists payments $C_t : \mathcal{N} \times \mathcal{N} \times \mathcal{U}^1 \times \mathcal{U}^2 \to \mathbb{R}$ which are exchanged between nodes 1 and 2 at each time $t \in \{1, 2, \dots T-1\}$ such that ψ^* is the unique subgame perfect equilibrium of the resulting dynamic game $\tilde{\mathcal{G}}$.

Proof. For game $\tilde{\mathcal{G}}$ at any time $t \in \{1, 2, \dots T\}$, after history $h \in \mathcal{H}$ and action profile $u \in \mathcal{U}^1 \times \mathcal{U}^2$, the cost-to-go functions for both nodes are

$$V_t^1(h,u) := g_t^1(x_t,u) + C_t(h,u) + \lambda \mathbb{E}\left[V_{t+1}^1(h')|h,u\right] \tag{2.40a}$$

$$V_t^2(h,u) := g_t^2(x_t,u) - C_t(h,u) + \lambda \mathbb{E}\left[V_{t+1}^2(h')|h,u\right], \tag{2.40b}$$

where $V_{T+1}^i(\cdot,\cdot) = 0$ and $x_t = (x_t^1, x_t^2)$ are queue lengths of user 1 and 2 corresponding to history h.

Let $\sigma = (\sigma_t)_{t=1}^T$ be a Markov policy of the dynamic game $\tilde{\mathcal{G}}$ such that for all histories $h \in \mathcal{H}$ of length t, $\sigma(h) = \psi_t^*(x_t^1, x_t^2)$. The payments $C_t(h, u^1, u^2)$ are constructed backward recursively as follows. Let $V_T^{i,\sigma}(x_T^1, x_T^2) = g_T^i(x_T^1, x_T^2)$ for user i, $i = 1, 2$. Then for any $t \in \{T-1, T-2 \dots 1\}$, let $C_t(h, u^1, u^2)$ be constructed as in Lemma 2.6 for the game with instantaneous costs being the cost-to-go functions $g_t^i(x_t, u) + \lambda \mathbb{E}[V_{t+1}^{i,\sigma}(h')|h, u]$ for each user i, such that action profile $\sigma(h)$ is the strictly dominant strategy equilibrium. Then,

$$V_t^{1,\sigma}(h) = g_t^1(x_t, \sigma(h)) + C_t(h, \sigma(h)) + \lambda \mathbb{E}[V_{t+1}^{1,\sigma}(h')|h, \sigma(h)] \tag{2.41}$$

$$V_t^{2,\sigma}(h) = g_t^2(x_t, \sigma(h)) - C_t(h, \sigma(h)) + \lambda \mathbb{E}[V_{t+1}^{2,\sigma}(h')|h, \sigma(h)]. \tag{2.42}$$

The above implies that for each subgame $\tilde{\mathcal{G}}(h)$, the policy $\sigma|_h$ strictly dominates every other policy σ' in $\tilde{\mathcal{G}}(h)$ that differs from $\sigma|_h$ only in the action it prescribes after the initial history of $\tilde{\mathcal{G}}(h)$. By the one-shot deviation property [45, Lemma 98.2] and strict dominance, the policy σ is the unique subgame perfect equilibrium of the game. By construction, it

[6] Due to the finiteness of the action sets, there always exists a $c > 0$ such that $r^k(u_1, u_2) \geq -c \quad \forall u_1, u_2, k = 1, 2$, then choosing $C(u^1, b) = 4c \ \forall u^1 \neq a, C(a, u^2) = c \ \forall u^2 \neq b, C(a, b) = 2c, C(u^1, u^2) = 2.5c \ \forall u^1 \neq a, u^2 \neq b$ will satisfy inequalities of Lemma 5 with all payments being positive.

achieves a socially optimal solution.

Since $\sigma(h)$ is a Markov policy, $C_t(\cdot,\cdot)$ and $V_t^{i,\sigma}(\cdot,\cdot)$ can be reduced to functions of queue lengths x_t^1, x_t^2 corresponding to history h, instead of h. This follows from induction. For $t = T, C_T(h,\sigma(h)) = 0$ and $V_T^{i,\sigma}(x_T^1, x_T^2) = g_T^i(x_T^1, x_T^2)$ which establishes the base case. Now since $V_{t+1}^{i,\sigma}(\cdot,\cdot)$ depends on history h only through the queue lengths x_{t+1}, and since $\sigma(h) = \psi_t^*(x_t^1, x_t^2)$ depends on x_t, it follows that $C_t(h, u^1, u^2)$ constructed as in Lemma 2.6 for the game with costs $g_t^i(x_t, \sigma(h)) + \lambda \mathbb{E}[V_{t+1}^{i,\sigma}(h')|h,u]$ for user i also depends on history h through queue lengths x_t. Thus from equations (2.41)-(2.42), $V_t^{i,\sigma}(\cdot,\cdot)$ also depends only on x_t. This completes the induction step.

The above result gives a constructive proof of existence of payment transfers such that any Markov policy can be implemented, and thus the socially optimum Markov policy can also be implemented as the unique subgame perfect equilibrium of the dynamic game.

Tables 2.1 and 2.2 show examples of construction of payments based on Proposition 2.3 for arbitrary payments and positive payments respectively. The system with parameters $E_{12} = 0.1, E_{23} = 0.2, E_{13} = 10, \lambda = 0.99, p^1 = 0.1, p^2 = 0.1, c_d = 5, N = 10$ is considered, and an arbitrary state $(x_t^1, x_t^2) = (3,3)$ is chosen for demonstrating the resulting cost-to-go functions and payments. Tables 2.1(a) and 2.2(a) give the cost-to-go functions $V_t^{1,\sigma}(x_t)$ for the game at time t obtained using policy iteration such that the socially optimal Markov policy (for the centralized problem) is implemented (i.e. is the dominant strategy) for all times after t. Due to the time-invariance of costs, policy iteration converges to a stationary solution, which, for this set of parameters is the policy (W, T_{23}) when at the chosen state $(x_t^1, x_t^2) = (3,3)$. Note that from the values of the cost-to-go functions in Tables 2.1(a), 2.2(a) it can be deduced that both actions (W, T_{23}) and (T_{13}, W_r) are Nash equilibria, which is not desired. Tables 2.1(b), 2.2(b) give the cost-to-go functions $V_t^{1,\sigma}(x_t)$ for the corresponding game $\tilde{\mathcal{G}}$ (i.e., with payment transfers), where transfers for this instant t are calculated according to Lemma 2.6. Note that with the introduction of payments, action (W, T_{23}) is the dominant strategy equilibrium.

In this game, the total expected cost for a user by participating in the game should be less than the expected cost they incur by not participating in the game. It is assumed that by non-participation, the users are not able to transmit their packets, and thus their queues increase to capacity, whereas by participating in the game they do strictly better. Thus for reasonable cost functions (for e.g. $g^i(x,u) \sim x^i$) which are monotonously increasing and unbounded with the queue length, and for large enough N and λ close to 1, the users will always have an incentive to participate in the game.

Table 2.1: Costs for subgame at time t. Parameters are $E_{12} = 0.1, E_{23} = 0.2, E_{13} = 10, \lambda = 0.99, p^1 = 0.1, p^2 = 0.1, c_d = 5, N = 10$. States are $(x, y) = (3, 3)$.

(a) Without payments

		W_a	W_r	T_{23}
W	$V_t^1(x,y)$	0.4586	0.4586	0.4103
	$V_t^2(x,y)$	0.3970	0.3970	0.3616
T_{13}	$V_t^1(x,y)$	0.4484	0.4484	0.5693
	$V_t^2(x,y)$	0.3842	0.3842	0.3992
T_{12}	$V_t^1(x,y)$	0.3743	0.4597	0.4597
	$V_t^2(x,y)$	0.4337	0.3970	0.3992

(b) With real-valued payments

		W_a	W_r	T_{23}
W	$V_t^1(x,y)$	0.4129	0.4462	0.4103
	$V_t^2(x,y)$	0.4427	0.4094	0.3616
	$C_t(x,y)$	-0.0438	-0.0124	0
T_{13}	$V_t^1(x,y)$	0.4472	0.4472	0.5841
	$V_t^2(x,y)$	0.3854	0.3854	0.3844
	$C_t(x,y)$	-0.0012	-0.0012	0.0148
T_{12}	$V_t^1(x,y)$	0.4139	0.4597	0.4658
	$V_t^2(x,y)$	0.3941	0.3970	0.3931
	$C_t(x,y)$	0.0397	0	0.0061

Table 2.2: Costs for subgame at time t. Parameters are $E_{12} = 0.1, E_{23} = 0.2, E_{13} = 10, \lambda = 0.99, p^1 = 0.1, p^2 = 0.1, c_d = 5, N = 10$. States are $(x, y) = (3, 3)$.

(a) Without payments

		W_a	W_r	T_{23}
W	$V_t^1(x,y)$	0.6258	0.6258	0.5782
	$V_t^2(x,y)$	0.2298	0.2298	0.1938
T_{13}	$V_t^1(x,y)$	0.6028	0.6028	0.7364
	$V_t^2(x,y)$	0.2298	0.2298	0.2320
T_{12}	$V_t^1(x,y)$	0.5282	0.6269	0.6269
	$V_t^2(x,y)$	0.2798	0.2298	0.2320

(b) With positive payments

		W_a	W_r	T_{23}
W	$V_t^1(x,y)$	0.6258	0.6258	0.5782
	$V_t^2(x,y)$	0.2298	0.2298	0.1938
	$C_t(x,y)$	0	0	0
T_{13}	$V_t^1(x,y)$	0.6258	0.6258	0.7617
	$V_t^2(x,y)$	0.2068	0.2068	0.2068
	$C_t(x,y)$	0.0230	0.0230	0.0252
T_{12}	$V_t^1(x,y)$	0.6258	0.6269	0.6768
	$V_t^2(x,y)$	0.1822	0.2298	0.1822
	$C_t(x,y)$	0.0977	0	0.0499

2.5 Conclusion

This chapter studies energy and delay tradeoff in cooperative communication through a simple relay channel, when the source and the destination node are cooperative and when they are strategic. When users are cooperative, by posing it as a decentralized, infinite horizon decentralized stochastic control problem, a structural result is proven stating that the optimal policy can be found by solving a Bellman-type fixed-point equation and optimal control can be given as $u_t^k = g_t^k(x_t^k, \xi_t^1, \xi_t^2)$. The domain of optimization is the space of the pair of marginal probability mass functions on the integers $\mathcal{P}(\mathcal{N}) \times \mathcal{P}(\mathcal{N})$. Numerical results are presented to compare the performance of a suboptimal policy from our analysis with standard TDMA and RA policies. Future research directions include the unveiling of additional structural properties of the optimal strategy (e.g., threshold strategies), as well as designing optimal and efficient suboptimal strategies. This problem can be extended to the case of multiple source/relay nodes. For this, the model needs to be enriched so that each collision also contains the information regarding which nodes transmitted. This can be achieved if each node transmits a "signature" waveform along with the data waveform such that the signature waveforms of all users are mutually orthogonal and orthogonal to the data (e.g., in frequency). The optimal decentralized solution scales exponentially with the number of nodes i, since, as shown in this work, a sufficient state for control is the set of marginal distributions $\{\xi^k(x_t^k); k \in \{1, 2, \dots i\}\}$ on the queue of size N, which grows as $\mathbb{R}^{iN}$ instead of the joint distribution $\pi(x_t^1, \dots x_t^i)$ which grows double exponentially in i as $\mathbb{R}^{N^i}$.

The second part of this chapter studies the relay channel with strategic source and relay nodes that minimize their individual expected energy and average delay. It is shown that there exist transfer payment functions $C_t(\cdot, \cdot)$ such that implementing socially optimal Markov policy is also the unique subgame perfect equilibrium of the dynamic game. In this work an important assumption is made that all information including states and utility functions are known to everybody. The decentralized setup for strategic users i.e. when states are not known, would be a challenging and interesting problem to consider and to the best knowledge of the authors, dynamic games with decentralized information are not well studied [25, 41]. If the assumption of known utilities is relaxed, then the problem becomes significantly harder and comes under the purview of mechanism design and Implementation Theory for dynamic games [5, 24].

2.6 Appendix A (Proof of Lemma 2.3)

Proof.

$$
\begin{aligned}
P^{\phi^2}(x^1_{t+1}, u_{1:t} | x^1_{1:t}, u_{1:t-1}, u^1_{1:t}) =& P^{\phi^2}(x^1_{t+1} | x^1_{1:t}, u_{1:t}) P^{\phi^2}(u_{1:t} | x^1_{1:t}, u_{1:t-1}, u^1_t) && \text{(2.43a)}\\
=& P(x^1_{t+1} | x^1_t, u_t) P^{\phi^2}(u^2_t | x^1_{1:t}, u_{1:t-1}, u^1_t) && \text{(2.43b)}\\
=& P(x^1_{t+1} | x^1_t, u_t) P^{\phi^2}(u^2_t | u_{1:t-1}) && \text{(2.43c)}\\
=& P^{\phi^2}(x^1_{t+1}, u_{1:t} | x^1_t, u_{1:t-1}, u^1_t) && \text{(2.43d)}
\end{aligned}
$$

where (2.43b) follows since $x^1_{t+1} = f_t(x^1_t, p^1_{t+1}, u_t)$ where f_t is as defined in (2.2) and by independence of basic random variables, and (2.43c) follows since U^2_t is a function of $X^2_{1:t}$, U^1_t is a function of $X^1_{1:t}$ and $X^1_{1:t}, X^2_{1:t}$ are conditionally independent given $U_{1:t-1}$ (Lemma 2.2).

For the second part,

$$
\begin{aligned}
\mathbb{E}^{\phi^2}\{g(x_t, u_t) | x^1_{1:t}, u_{1:t-1}, u^1_{1:t}\} =& \sum_{x_t, u_t} g(x_t, u_t) P^{\phi^2}(x_t, u_t | x^1_{1:t}, u_{1:t-1}, u^1_{1:t}) && \text{(2.44a)}\\
=& \sum_{x^2_t, u^2_t} g(x_t, u_t) P^{\phi^2}(x^2_t, u^2_t | x^1_{1:t}, u_{1:t-1}, u^1_{1:t}) && \text{(2.44b)}\\
=& \sum_{x^2_t, u^2_t} g(x_t, u_t) P^{\phi^2}(x^2_t, u^2_t | u_{1:t-1}) && \text{(2.44c)}\\
=& \mathbb{E}^{\phi^2}\{g(x_t, u_t) | x^1_t, u_{1:t-1}, u^1_t\} && \text{(2.44d)}\\
=& \hat{g}(x^1_t, u_{1:t-1}, u^1_t) && \text{(2.44e)}
\end{aligned}
$$

where (2.44c) follows from Lemma 2.2.

2.7 Appendix B (Proof of Lemma 2.4)

Proof. Fix ψ

$$
\begin{aligned}
\pi_{t+1}(x_{t+1}) =& P^{\psi}(X_{t+1} = x_{t+1} | u_{1:t}, \gamma_{1:t}) && \text{(2.45a)}\\
=& \sum_{x_t} P^{\psi}(x_{t+1}, x_t | u_{1:t}, \gamma_{1:t}) && \text{(2.45b)}\\
=& \sum_{x_t} P^{\psi}(x_t | u_{1:t}, \gamma_{1:t}) . P(x_{t+1} | x_t, u_t) && \text{(2.45c)}
\end{aligned}
$$

where (2.45c) is true by Markov property and the fact that $\gamma_{1:t}$ is a function of $u_{1:t}$. Now,

$$P^{\psi}(x_t|u_{1:t},\gamma_{1:t}) = \frac{P^{\psi}(x_t,u_t|u_{1:t-1},\gamma_{1:t})}{\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t,u_t|u_{1:t-1},\gamma_{1:t})} \tag{2.46a}$$

$$= \frac{P^{\psi}(x_t|u_{1:t-1},\gamma_{1:t})P^{\psi}(u_t|u_{1:t-1},\gamma_{1:t},x_t)}{\sum_{\hat{x}_t} P(\hat{x}_t,u_t|u_{1:t-1},\gamma_{1:t})} \tag{2.46b}$$

$$= \frac{P^{\psi}(x_t|u_{1:t-1},\gamma_{1:t-1})\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)}{\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t|u_{1:t-1},\gamma_{1:t-1})\mathbf{1}_{\{\gamma_t(\hat{x}_t)\}}(u_t)} \tag{2.46c}$$

where first part in numerator in (2.46c) is true since given policy ψ, γ_t can be computed as $\gamma_t = \psi_t(u_{1:t-1})$.

We conclude that

$$P(x_t|u_{1:t},\gamma_{1:t}) = \frac{\pi_t(x_t)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)}{\sum_{\hat{x}_t} \pi_t(\hat{x}_t)\mathbf{1}_{\{\gamma_t(\hat{x}_t)\}}(u_t)}, \tag{2.47}$$

thus,

$$\pi_{t+1} = F(\pi_t,\gamma_t,u_t) \tag{2.48}$$

where F is independent of policy ψ.

2.8 Appendix C (Proof of Theorem 2.1)

Proof.

$$P(\pi_{t+1}|\pi_{1:t},\gamma_{1:t}) = \sum_{u_t} P(\pi_{t+1},u_t|\pi_{1:t},\gamma_{1:t}) \tag{2.49a}$$

$$= \sum_{u_t} \mathbf{1}_{\{F(\pi_t,\gamma_t,u_t)\}}(\pi_{t+1})P(u_t|\pi_{1:t},\gamma_{1:t}) \tag{2.49b}$$

$$= \sum_{u_t,x_t} \mathbf{1}_{\{F(\pi_t,\gamma_t,u_t)\}}(\pi_{t+1})\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)P(x_t|\pi_{1:t},\gamma_{1:t}) \tag{2.49c}$$

$$= \sum_{u_t,x_t} \pi_t(x_t)\mathbf{1}_{\{F(\pi_t,\gamma_t,u_t)\}}(\pi_{t+1})\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t) \tag{2.49d}$$

$$= P(\pi_{t+1}|\pi_t,\gamma_t) \tag{2.49e}$$

$$\mathbb{E}(g(x_t,u_t)|\pi_{1:t},\gamma_{1:t}) = \sum_{x_t,u_t} g(x_t,u_t)P(x_t,u_t|\pi_{1:t},\gamma_{1:t}) \tag{2.50a}$$

$$= \sum_{x_t,u_t} g(x_t,u_t)P(x_t|\pi_{1:t},\gamma_{1:t})\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t) \tag{2.50b}$$

$$= \sum_{x_t,u_t} g(x_t,u_t)\pi_t(x_t)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t) \tag{2.50c}$$

$$= \hat{g}(\pi_t,\gamma_t) \tag{2.50d}$$

2.9 Appendix D (Proof of Lemma 2.5)

Proof. For any fixed coordinator strategy ψ,

$$\xi^1_{t+1}(x^1_{t+1}) = P^\psi(x^1_{t+1}|u_{1:t},\gamma_{1:t}) \tag{2.51a}$$

$$= \sum_{x_t} P^\psi(x^1_{t+1},x_t|u_{1:t},\gamma_{1:t}) \tag{2.51b}$$

$$= \sum_{x_t} P^\psi(x_t|u_{1:t},\gamma_{1:t}).P(x^1_{t+1}|x^1_t,u_t) \tag{2.51c}$$

Now,

$$P^\psi(x_t|u_{1:t},\gamma_{1:t}) = \frac{P^\psi(x_t,u_t|u_{1:t-1},\gamma_{1:t})}{\sum_{\hat{x}_t} P^\psi(\hat{x}_t,u_t|u_{1:t-1},\gamma_{1:t})} \tag{2.52a}$$

$$= \frac{P^\psi(x_t|u_{1:t-1},\gamma_{1:t})P^\psi(u_t|u_{1:t-1},\gamma_{1:t},x_t)}{\sum_{\hat{x}_t} P^\psi(\hat{x}_t,u_t|u_{1:t-1},\gamma_{1:t})} \tag{2.52b}$$

$$= \frac{P^\psi(x_t|u_{1:t-1},\gamma_{1:t-1})\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)}{\sum_{\hat{x}_t} P^\psi(\hat{x}_t|u_{1:t-1},\gamma_{1:t-1})\mathbf{1}_{\{\gamma_t(\hat{x}_t)\}}(u_t)} \tag{2.52c}$$

$$= \frac{\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)}{\sum_{\hat{x}_t} \xi^1_t(\hat{x}^1_t)\xi^2_t(\hat{x}^2_t)\mathbf{1}_{\{\gamma_t(\hat{x}_t)\}}(u_t)} \tag{2.52d}$$

where (2.52c) is true since given policy ψ, $\gamma_t = \psi_t(u_{1:t-1})$ and (2.52d) is true since X^1_t and X^2_t are conditionally independent given U_{t-1} (Lemma 2.2). Thus,

$$\xi^1_{t+1}(x^1_{t+1}) = \sum_{x_t} P(x^1_{t+1}|x^1_t,u_t)\frac{\xi^1_t(x^1_t)\xi^2_t(x^2_t)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)}{\sum_{\hat{x}_t} \xi^1_t(\hat{x}^1_t)\xi^2_t(\hat{x}^2_t)\mathbf{1}_{\{\gamma_t(\hat{x}_t)\}}(u_t)} \tag{2.53a}$$

$$= \sum_{x_t^1} P(x_{t+1}^1 | x_t^1, u_t)$$
$$\frac{\xi_t^1(x_t^1)\mathbf{1}_{\{\gamma_t^1(x_t^1)\}}(u_t^1)\sum_{x_t^2}\mathbf{1}_{\{\gamma_t^2(x_t^2)\}}(u_t^2)\xi_t^2(x_t^2)}{\sum_{\hat{x}_t^1}\xi_t^1(\hat{x}_t^1)\mathbf{1}_{\{\gamma_t^1(\hat{x}_t^1)\}}(u_t^1)\sum_{\hat{x}_t^2}\mathbf{1}_{\{\gamma_t^2(\hat{x}_t^2)\}}(u_t^2)\xi_t^2(\hat{x}_t^2)} \tag{2.53b}$$
$$= \sum_{x_t^1} P(x_{t+1}^1 | x_t^1, u_t) \frac{\xi_t^1(x_t^1)\mathbf{1}_{\{\gamma_t^1(x_t^1)\}}(u_t^1)}{\sum_{\hat{x}_t^1}\xi_t^1(\hat{x}_t^1)\mathbf{1}_{\{\gamma_t^1(\hat{x}_t^1)\}}(u_t^1)} \tag{2.53c}$$
$$= G^1(\xi_t^1, \gamma_t^1, u_t)(x_{t+1}^1) \tag{2.53d}$$

Similarly $\xi_{t+1}^2 = G^2(\xi_t^2, \gamma_t^2, u_t)$ where G^1 and G^2 are deterministic functions independent of policy ψ.

2.10 Appendix E (Proof of Theorem 2.2)

Proof. In the following we use the notation $G := (G^1, G^2)$

$$P^{\phi}(\xi_{t+1}^1, \xi_{t+1}^2 | \xi_{1:t}^1, \xi_{1:t}^2, \gamma_{1:t}) = \sum_{u_t} P^{\phi}(\xi_{t+1}^1, \xi_{t+1}^2, u_t | \xi_{1:t}^1, \xi_{1:t}^2, \gamma_{1:t}) \tag{2.54a}$$
$$= \sum_{u_t} \mathbf{1}_{\{G(\xi_t^1, \xi_t^2, \gamma_t, u_t)\}}(\xi_{t+1}^1, \xi_{t+1}^2) P^{\phi}(u_t | \xi_{1:t}^1, \xi_{1:t}^2, \gamma_{1:t}) \tag{2.54b}$$
$$= \sum_{u_t, x_t} \mathbf{1}_{\{G(\xi_t^1, \xi_t^2, \gamma_t, u_t)\}}(\xi_{t+1}^1, \xi_{t+1}^2)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)$$
$$P^{\phi}(x_t | \xi_{1:t}^1, \xi_{1:t}^2, \gamma_{1:t}) \tag{2.54c}$$
$$= \sum_{u_t, x_t} \xi_t^1(x_t^1)\xi_t^2(x_t^2)\mathbf{1}_{\{G(\xi_t^1, \xi_t^2, \gamma_t, u_t)\}}(\xi_{t+1}^1, \xi_{t+1}^2)\mathbf{1}_{\{\gamma_t(x_t)\}}(u_t)$$
$$\tag{2.54d}$$
$$= \sum_{x_t} \xi_t^1(x_t^1)\xi_t^2(x_t^2)\mathbf{1}_{\{G(\xi_t^1, \xi_t^2, \gamma_t, u_t)\}}(\xi_{t+1}^1, \xi_{t+1}^2) \tag{2.54e}$$
$$= P(\xi_{t+1}^1, \xi_{t+1}^2 | \xi_t^1, \xi_t^2, \gamma_t) \tag{2.54f}$$

$$\mathbb{E}(g(x_t, u_t)|\xi_{1:t}, \gamma_{1:t}) = \sum_{x_t, u_t} (g(x_t, u_t)) P(x_t, u_t|\xi_{1:t}, \gamma_{1:t}) \tag{2.55a}$$

$$= \sum_{x_t, u_t} (g(x_t, u_t)) P(x_t|\xi_{1:t}, \gamma_{1:t}) \mathbf{1}_{\{\gamma_t(x_t)\}}(u_t) \tag{2.55b}$$

$$= \sum_{x_t, u_t} (g(x_t, u_t)) \xi_t^1(x_t^1) \xi_t^2(x_t^2) \mathbf{1}_{\{\gamma_t(x_t)\}}(u_t) \tag{2.55c}$$

$$= \tilde{g}(\xi_t, \gamma_t) \tag{2.55d}$$

CHAPTER 3

Structured perfect Bayesian equilibria in dynamic games with asymmetric information

3.1 Introduction

There are many practical scenarios where strategic players with different sets of observations are involved in a time-evolving dynamical process such that their actions influence each others' payoffs. Such scenarios include repeated online advertisement auctions, wireless resource sharing, competing sellers and energy markets. In the case of repeated online advertisement auctions, advertisers place bids for locations on a website to sell a product. These bids are based on the value of that product, which is privately observed by an advertiser, and past actions of everybody else, which are observed publicaly. Each advertiser's goal is to maximize its reward, which depends on the value of the products and on the actions taken by everybody else. A similar scenario can be considered for wireless resource sharing where players are allocated channels that interfere with each other. Each player privately observes its channel gain and takes actions, which may be the choice of modulation and coding scheme and also the transmission power. The reward here is the rate each player gets at time t, which is a function of everyone's channel gain and actions. Consider another scenario where different sellers compete to sell different but related goods which are complementary, substitutable or in general, with externalities. The true value of the goods is private information of a seller who, at each stage, takes an action to stock some amount of goods for sale. Her profit is based on some market mechanism (say through Walrasian prices) based on the true value of all the goods and their availability in the market, which depends on the actions of the other sellers. Each seller wants to maximize her own profit. Finally, a similar scenario also exists for energy markets, where different suppliers (to their different end consumers) bid their estimated power outputs to an independent system operator (ISO) that forms the market mechanism to determine the prices assessed to the different suppliers. Each supplier wants to maximize its returns, which depend on its

cost of production of energy, which is their private information, and the market-determined prices, which depend on all the bids.

Such dynamical systems with strategic players are modeled as dynamic games. In dynamic games with perfect and symmetric information, subgame perfect equilibrium (SPE) is an appropriate equilibrium concept [45], [4], [13] and there is a backward recursive algorithm to find all subgame perfect equilibria of such games. Maskin and Tirole in [38] introduced the concept of Markov perfect equilibrium (MPE) for dynamic games with perfect and symmetric information, where equilibrium strategies are dependent on some payoff relevant state of the system rather than on the entire history. However, for games with asymmetric information, since players have different information sets in each period, they need to form a belief on the information sets of other players, based upon which they predict their strategies. As a result, SPE or MPE are not appropriate equilibrium concepts for such setting. There are several notions of equilibrium for such games, such as perfect Bayesian equilibrium (PBE), sequential equilibrium, trembling hand equilibrium [13, 45]. Each of these notions of equilibrium consists of a strategy and a belief profile of all players. The equilibrium strategies are optimal given the beliefs and the beliefs are derived from the equilibrium strategy profile and using Bayes' rule (whenever possible), with some equilibrium concepts requiring further refinements. Due to this circular argument of beliefs being consistent with strategies, which are in turn optimal given the beliefs, finding such equilibria is a difficult task. Moreover, strategies are function of histories, which belong to an ever-expanding space, and thus the space of optimization also becomes computationally intractable. There is no known methodology to find such equilibria for general dynamic games with asymmetric information.

In this chapter, we consider a model where players observe their types privately and publicly observe the actions taken by other players at the end of each period. Their instantaneous rewards depend on everyones' types and actions. We provide a two-step algorithm involving a backward recursion followed by a forward recursion to construct a class of PBE for the dynamic game in consideration, which we call *structured perfect Bayesian equilibria* (SPBE). In these equilibria, players' strategies are based on their current type and a set of beliefs on each type, which is common to all players and lie in a time-invariant space. These beliefs on players' types form independent controlled Markov processes that together summarize the common information history, and are updated individually and sequentially, based on corresponding agents' actions and (partial) strategies. The algorithm works as follows. In a backward recursive way, for each stage, the algorithm finds an equilibrium strategy function for all possible beliefs on types of the players, which involves solving a fixed point equation on the space of probability simplexes. Then, the equilibrium

strategies and beliefs are obtained through forward recursion by operating on the function obtained in the backward step. The SBPEs that are developed in this chapter are analogous to the MPEs for dynamic games with perfect information in the sense that players choose their actions based on beliefs that depend on common information and have Markovian dynamics, where actions of a players are now partial functions from their private information to their action sets.

Related literature on this topic include [16, 41] and [46]. Nayyar et al. in [16, 41] consider a model of dynamic games with asymmetric information. There is an underlying controlled Markov process, where players jointly observe part of the process and also make some observations privately. It is shown in [16, 41] that the considered game with asymmetric information, under certain assumptions, can be transformed to another game with symmetric information. Once this is established, a backward recursive algorithm is provided to find MPE of the transformed game, which are equivalently Nash equilibria of the transformed symmetric information game. For this strong equivalence to hold, authors in [16, 41] make a critical assumption in their model: based on the common information, a player's posterior beliefs about the system state and about other players' information are independent of the strategies used by the players in the past. Our model is different from the model considered in [16, 41]. We assume that the underlying state of the system has independent components, each constituting the type of a player. However, we do not make any assumption regarding update of beliefs and allow the common information based belief state to depend on players' strategies.

Ouyang et al. in [46] consider a dynamic oligopoly game with N strategic sellers of different goods and M strategic buyers. Each seller privately observes the valuation of their good, which is assumed to have independent Markovian dynamics, thus resulting in a dynamic game of asymmetric information. In each period, sellers post prices for their goods and buyers make decisions regarding buying the goods. Then a public signal indicating buyers experience is revealed, which depends on sellers' valuation of the goods. Authors in [46] consider a policy-dependent common information based belief state based on which they define the concept of common information based equilibria. They show that for any given update function of this belief state, which is consistent with strategies of the players, if all other players play actions based on this common belief and their private information, then player i faces a Markov decision process (MDP) with respect to its action with state as common belief and its type. For every prior distribution, this defines a fixed point equation on belief update functions and strategies of all players. They provide necessary and sufficient conditions for common information based strategy profile and belief update functions to constitute PBE of the game; however they do no provide a systematic way to

find such equilibria. In addition, because of the special structure of the reward function, the problem admits a degenerate solution where agents' strategies do not depend on their private information, and therefore no signaling takes place. This allows existence of myopic, type-independent equilibrium policies (although other equilibria may also exist).

The chapter is organized as follows. In Section 3.2, we present our model. In Section 3.3 we present structural results that serve as motivation for SPBE. In Section 3.4 we present the main result by providing a two-step backward-forward recursive algorithm to construct a strategy profile and a sequence of beliefs and show that it is a PBE of the dynamic game considered. As an illustration, we apply this algorithm on a discrete version of an example from [13] on repeated public good game in Section 3.5. We conclude in Section 3.6. All proofs are presented in Appendices.

3.1.1 Notation

In this chapter, for an independent probabilistic strategy profile of players, $(\beta_t^i)_{i\in\mathcal{N}}$, where probability of action a_t^i conditioned on $a_{1:t-1}, x_{1:t}^i$ is given by $\beta_t^i(a_t^i|a_{1:t-1}, x_{1:t}^i)$, we use the short hand notation $\beta_t^{-i}(a_t^{-i}|a_{1:t-1}, x_{1:t}^{-i})$ to represent $\prod_{j\neq i}\beta_t^j(a_t^j|a_{1:t-1}, x_{1:t}^j)$.

3.2 Model

We consider a discrete-time dynamical system with N strategic players in the set $\mathcal{N} \triangleq \{1, 2, \dots N\}$, over a time horizon $\mathcal{T} \triangleq \{1, 2, \dots T\}$, and with perfect recall. There is a dynamic state of the system $X_t \triangleq (X_t^1, X_t^2, \dots X_t^N)$, where $X_t^i \in \mathcal{X}^i$ is the type of player i at time t, which is perfectly observed and is its private information. Types of the players evolve as conditionally independent, controlled Markov processes such that

$$P(x_1) = \prod_{i=1}^{N} Q_1^i(x_1^i) \tag{3.1a}$$

$$P(x_t|a_{1:t-1}, x_{1:t-1}) = P(x_t|a_{t-1}, x_{t-1}) \tag{3.1b}$$

$$= \prod_{i=1}^{N} Q_t^i(x_t^i|a_{t-1}, x_{t-1}^i), \tag{3.1c}$$

where Q_t^i are known kernels. Player i at time t takes action $a_t^i \in \mathcal{A}^i$ on observing $a_{1:t-1}$, which is common information among players, and $x_{1:t}^i$, which it observes privately. The sets $\mathcal{A}^i, \mathcal{X}^i$ are assumed to be finite. Let $g^i = (g_t^i)_{t\in\mathcal{T}}$ be a probabilistic strategy of player i where $g_t^i : \mathcal{A}^{t-1} \times (\mathcal{X}^i)^t \to \mathcal{P}(\mathcal{A}^i)$ such that player i plays action A_t^i according to

$A_t^i \sim g_t^i(\cdot|a_{1:t-1}, x_{1:t}^i)$. Let $g \triangleq (g^i)_{i\in\mathcal{N}}$ be a strategy profile of all players. At the end of interval t, player i receives an instantaneous reward $R^i(x_t, a_t)$. The objective of player i is to maximize its total expected reward

$$J^{i,g} \triangleq \mathbb{E}^g \left\{ \sum_{t=1}^{T} R^i(X_t, A_t) \right\}. \tag{3.2}$$

With all players being strategic, this problem is modeled as a dynamic game $\mathfrak{D}$ with imperfect and asymmetric information, and with simultaneous moves.

3.3 Motivation for structured equilibria

In this section we present structural results for the considered dynamical process that serve as a motivation for finding SPBE of the underlying game $\mathfrak{D}$. Specifically, we define a belief state based on common information history, and show that any reward profile that can be obtained through a general strategy profile can also be obtained through strategies that depend on this belief state and player's current type, which is its private information. These structural results are inspired by the analysis of decentralized team problems, which serve as guiding principles to design our equilibrium strategies. While these structural results provide intuition and the required notation, they are not directly used in the proofs for finding SPBEs, later, in Section 3.4.

At any time t, player i has information $(a_{1:t-1}, x_{1:t}^i)$ where $a_{1:t-1}$ is the common information among players, and $x_{1:t}^i$ is the private information of player i. Since $(a_{1:t-1}, x_{1:t}^i)$ increases with time, any strategy of the form $A_t^i \sim g_t^i(\cdot|a_{1:t-1}, x_{1:t}^i)$ becomes unwieldy. Thus it is desirable to have an information state in a time-invariant space that succinctly summarizes $(a_{1:t-1}, x_{1:t}^i)$, and that can be sequentially updated. We first show in Lemma 3.1 that given common information $a_{1:t-1}$ and its current type x_t^i, player i can discard its type history $x_{1:t-1}^i$ and play a strategy of the form $s_t^i(a_t^i|a_{1:t-1}, x_t^i)$. Then in Lemma 3.2, we show that $a_{1:t-1}$ can be summarized through a belief π_t, defined as follows. For any strategy profile g, belief π_t on X_t, $\pi_t \in \mathcal{P}(\mathcal{X})$, is defined as $\pi_t(x_t) \triangleq P^g(X_t = x_t|a_{1:t-1})\ \forall x_t \in \mathcal{X}$. We also define the marginals $\pi_t^i(x_t^i) \triangleq P^g(x_t^i = x_t^i|a_{1:t-1})\ \forall x_t^i \in \mathcal{X}^i$.

For player i, we use notation g to denote a general policy of type $A_t^i \sim g_t^i(\cdot|a_{1:t-1}, x_{1:t}^i)$, notation s, where $s_t^i : (\mathcal{A})^{t-1} \times \mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)$, to denote a policy of the form $s_t^i(a_t^i|a_{1:t-1}, x_t^i)$, and notation m, where $m_t^i : \mathcal{P}(\times_{i\in\mathcal{N}}\mathcal{X}^i) \times \mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)$, to denote a policy of the form $m_t^i(a_t^i|\pi_t, x_t^i)$. It should be noted that since π_t is a function of random variables $a_{1:t-1}$, an

m policy is a special type of an s policy, which in turn, is a special type of a g policy.

Using the agent-by-agent approach [18], we show in Lemma 3.1 that any expected reward profile of the players that can be achieved by any general strategy profile g can also be achieved by a strategy profile s.

Lemma 3.1. Given a fixed strategy g^{-i} of all players other than player i and for any strategy g^i of player i, there exists a strategy s^i of player i such that

$$P^{s^i g^{-i}}(x_t, a_t) = P^{g^i g^{-i}}(x_t, a_t) \quad \forall t \in \mathcal{T}, x_t \in \mathcal{X}, a_t \in \mathcal{A}, \tag{3.3}$$

which implies $J^{i,s^i g^{-i}} = J^{i,g^i g^{-i}}$.

Proof. See Appendix A.

Since any s^i policy is also a g^i type policy, the above lemma can be iterated over all players, which implies that for any g policy profile there exists an s policy profile that achieves the same reward profile i.e. $(J^{i,s})_{i\in\mathcal{N}} = (J^{i,g})_{i\in\mathcal{N}}$.

Policies of types s still have increasing domain due to increasing common information, $a_{1:t-1}$. In order to summarize this information, we take an equivalent view of the system dynamics through a common agent, as taken by Nayyar et al. in [43]. The common agent approach is a general approach that has been used extensively for dynamic team problems [33, 35, 44, 56]. Using this approach, the problem can be equivalently described as follows: player i at time t observes $a_{1:t-1}$ and takes action γ_t^i, where $\gamma_t^i : \mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)$ is a partial (stochastic) function from its private information x_t^i to a_t^i of the form $\gamma_t^i(a_t^i|x_t^i)$. These actions are generated through some policy $\psi^i = (\psi_t^i)_{t\in\mathcal{T}}$, $\psi_t^i : \mathcal{A}^{t-1} \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)\}$, that operates on the common information $a_{1:t-1}$ so that $\gamma_t^i = \psi_t^i[a_{1:t-1}]$. Then any policy of the form $A_t^i \sim s_t^i(\cdot|a_{1:t-1}, x_t^i)$ is equivalent to $A_t^i \sim \psi_t^i[a_{1:t-1}](\cdot|x_t^i)$ [43].

We call a player i's policy through common agent to be of type ψ^i if its actions γ_t^i are taken as $\gamma_t^i = \psi_t^i[a_{1:t-1}]$. We call a player i's policy through common agent to be of type θ^i where $\theta_t^i : \mathcal{P}(\mathcal{X}) \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)\}$, if its actions γ_t^i are taken as $\gamma_t^i = \theta_t^i[\pi_t]$. A policy of type θ^i is also a policy of type ψ^i. There is a one-to-one correspondence between policies of type s^i and of type ψ^i and between policies of type m^i and of type θ^i.

In the following lemma, we show that the space of profiles of type s is outcome-equivalent to the space of profiles of type m.

Lemma 3.2. For any given strategy profile s of all players, there exists a strategy profile m such that

$$P^m(x_t, a_t) = P^s(x_t, a_t) \quad \forall t \in \mathcal{T}, x_t \in \mathcal{X}, a_t \in \mathcal{A}, \tag{3.4}$$

which implies $(J^{i,m})_{i\in\mathcal{N}} = (J^{i,s})_{i\in\mathcal{N}}$. Furthermore π_t can be factorized as $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(x_t^i)$ where each π_t^i can be updated through an update function

$$\pi_{t+1}^i = \bar{F}(\pi_t^i, \gamma_t^i, a_t), \tag{3.5}$$

where $\bar{F}$ is independent of s.

Proof. See Appendix B.

The above two lemmas show that any reward profile that can be generated through policy profile of type g can also be generated through policy profile of type m. It should be noted that the construction of s^i, as in (3.31), depends only on g^i, while the construction of m^i depends on the whole policy profile g and not just on g^i, since construction of θ^i depends on ψ in (3.43). Thus any unilateral deviation of player i in g policy profile does not necessarily translate to unilateral deviation of player i in the corresponding m policy profile. Therefore g being an equilibrium of the game (in some appropriate notion) does not necessitate the corresponding m also being an equilibrium.

As shown in the previous lemmas, due to the independence of types and their evolution as independent controlled Markov processes, for any strategy of the players, joint beliefs on types can be factorized as product of their marginals i.e. $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(x_t^i)$. Since in this chapter, we only deal with such joint beliefs, to accentuate this independence structure, we define $\underline{\pi}_t \in \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i)$ as vector of marginal beliefs where $\underline{\pi}_t := (\pi_t^i)_{i\in\mathcal{N}}$. In the rest of the chapter, we will use $\underline{\pi}_t$ instead of π_t whenever appropriate, where, of course, π_t can be constructed from $\underline{\pi}_t$. Similarly, we define a vector of belief updates as $F(\underline{\pi}, \gamma, a) := (\bar{F}(\pi^i, \gamma^i, a))_{i\in\mathcal{N}}$. We also change the notation of policies of type m as $m_t^i : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i) \times \mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)$ and common agent's policies of type θ as $\theta_t^i : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i) \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)\}$.

We end this section by noting that finding general PBEs of type g of the game $\mathfrak{D}$ would be a desirable goal, but due to the space of strategies growing exponentially with time, that would be computationally intractable. However lemma 3.1 suggests that strategies of type m form a class that is rich in the sense that they achieve every possible reward profile. Since these strategies are functions of beliefs π_t that lie in a time-invariant space and are easily updatable, equilibria of this type are potential candidates for computation through backward recursion. In this chapter our goal is to devise an algorithm to find structured equilibria of type m of the dynamic game $\mathfrak{D}$.

3.4 Algorithm for SPBE computation

3.4.1 Preliminaries

Any history of this game at which players take action is of the form $h_t = (a_{1:t-1}, x_{1:t})$. Let $\mathcal{H}_t$ be the set of such histories of the game at time t when players take action, $\mathcal{H} \triangleq \cup_{t=0}^{T} \mathcal{H}_t$ be the set of all possible such histories. At any time t player i observes $h_t^i = (a_{1:t-1}, x_{1:t}^i)$ and all players together have $h_t^c = a_{1:t-1}$ as common history. Let $\mathcal{H}_t^i$ be the set of observed histories of player i at time t and $\mathcal{H}_t^c$ be the set of common histories at time t. An appropriate concept of equilibrium for such games is PBE [13], which consists of a pair (β^*, μ^*) of strategy profile $\beta^* = (\beta_t^{*,i})_{t\in\mathcal{T}, i\in\mathcal{N}}$ where $\beta_t^{*,i} : \mathcal{H}_t^i \to \mathcal{P}(\mathcal{A}^i)$ and a belief profile $\mu^* = ({}^i\mu_t^*)_{t\in\mathcal{T}, i\in\mathcal{N}}$ where ${}^i\mu_t^* : \mathcal{H}_t^i \to \mathcal{P}(\mathcal{H}_t)$ that satisfy sequential rationality so that $\forall i \in \mathcal{N}, t \in \mathcal{T}, h_t^i \in \mathcal{H}_t^i, \beta^i$

$$\mathbb{E}^{\beta^{*,i}\beta^{*,-i},\,\mu^*[h_t^i]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\Big|h_t^i\right\} \geq \mathbb{E}^{\beta^i\beta^{*,-i},\,\mu^*[h_t^i]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\Big|h_t^i\right\}, \tag{3.6}$$

and the beliefs satisfy some consistency conditions as described in [13, p. 331]. In general, a belief for player i at time t, ${}^i\mu_t^*$ is defined on history $h_t = (a_{1:t-1}, x_{1:t})$ given its private history $h_t^i = (a_{1:t-1}, x_{1:t}^i)$. Here player i's private history $h_t^i = (a_{1:t-1}, x_{1:t}^i)$ consists of a public part $h_t^c = a_{1:t-1}$ and a private part $x_{1:t}^i$. At any time t, the relevant uncertainty player i has is about other players' type x_t^{-i}. In our setting, due to independence of types, player i's current type x_t^i does not provide any information about x_t^{-i}, as will be shown later. For this reason we consider beliefs that are functions of each agent's history h_t^i only through the common history h_t^c. Hence, for each agent i, its belief for each history $h_t^c = a_{1:t-1}$ is derived from a common belief $\mu_t^*[a_{1:t-1}]$, which itself factorizes into a product of marginals $\prod_{j\in\mathcal{N}} \mu_t^{*,j}[a_{1:t-1}]$, as will be shown later. Thus we can sufficiently use the system of beliefs, $\mu^* = (\mu_t^*)_{t\in\mathcal{T}}$ with $\mu_t^* : \mathcal{H}_t^c \to \mathcal{P}(\mathcal{X})$, with the understanding that agent i's belief on x_t^{-i} is $\mu_t^{*,-i}[a_{1:t-1}](x_t^{-i}) = \prod_{j\neq i} \mu_t^{*,j}[a_{1:t-1}](x_t^j)$. Under the above structure, all consistency conditions that are required for PBEs [13, p. 331] are automatically satisfied.

Structural results from Section 3.3 provide us motivation to study equilibria of the form $(m_t^i(a_t^i|\underline{\pi}_t, x_t^i))_{i\in\mathcal{N}}$, which are equivalent to policy profiles of the form $(\theta_t^i[\underline{\pi}_t](a_t^i|x_t^i))_{i\in\mathcal{N}}$ and have the advantage of being defined on a time-invariant space.

3.4.2 Backward recursion

In this section, we define an equilibrium generating function $\theta = (\theta_t^i)_{i\in\mathcal{N},t\in\mathcal{T}}$, where $\theta_t^i : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i) \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)\}$ and a sequence of functions $(V_t^i)_{i\in\mathcal{N},t\in\{1,2,\dots T+1\}}$, where $V_t^i : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i) \times \mathcal{X}^i \to \mathbb{R}$, in a backward recursive way, as follows.

1. Initialize $\forall \underline{\pi}_{T+1} \in \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i), x_{T+1}^i \in \mathcal{X}^i$,

$$V_{T+1}^i(\underline{\pi}_{T+1}, x_{T+1}^i) \stackrel{\triangle}{=} 0. \tag{3.7}$$

2. For $t = T, T-1, \dots 1,\ \forall \underline{\pi}_t \in \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i), \pi_t = \prod_{i\in\mathcal{N}} \pi_t^i$, let $\theta_t[\underline{\pi}_t]$ be generated as follows. Set $\tilde{\gamma}_t = \theta_t[\underline{\pi}_t]$, where $\tilde{\gamma}_t$ is the solution, if it exists[1], of the following equation, $\forall i \in \mathcal{N}, x_t^i \in \mathcal{X}^i$,

$$\tilde{\gamma}_t^i(\cdot|x_t^i) \in \arg\max_{\gamma_t^i(\cdot|x_t^i)} \mathbb{E}^{\gamma_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R^i(X_t, A_t) + V_{t+1}^i(F(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i)\big|x_t^i\right\}, \tag{3.8}$$

where expectation in (3.8) is with respect to random variables $(X_t^{-i}, A_t, X_{t+1}^i)$ through the measure $\pi_t^{-i}(x_t^{-i})\gamma_t^i(a_t^i|x_t^i)\tilde{\gamma}_t^{-i}(a_t^{-i}|x_t^{-i})Q_{t+1}^i(x_{t+1}^i|x_t^i, a_t)$, and F is defined in the proof of Lemma 3.2 and in particular Claim 3.5.

Furthermore, set

$$V_t^i(\underline{\pi}_t, x_t^i) \stackrel{\triangle}{=} \mathbb{E}^{\tilde{\gamma}_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R^i(X_t, A_t) + V_{t+1}^i(F(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i)\big|x_t^i\right\}. \tag{3.9}$$

It should be noted that in (3.8), $\tilde{\gamma}_t^i$ is not the outcome of the maximization operation as in a best response equation similar to that of a Bayesian Nash equilibrium. Rather (3.8) has characteristics of a fixed point equation. This is because the maximizer $\tilde{\gamma}_t^i$ appears in both, the left-hand-side and the right-hand-side of the equation. This distinct construction allows the maximization operation to be done with respect to the variable $\gamma_t^i(\cdot|x_t^i)$ for every x_t^i separately as opposed to be done with respect to the whole function $\gamma_t^i(\cdot|\cdot)$, and is pivotal in the construction.

To highlight the significance of structure of (3.8), we contrast it with two alternate incorrect constructions:

[1] Similar to the existence results shown in [46], in the special case of uncontrolled types and where agents' instantaneous rewards do not depend on their own private types, the fixed point equation always has a type-independent, myopic solution $\tilde{\gamma}_t^i(\cdot)$, since it degenerates to a best-response-like equation.

(a) Following the common information approach as in decentralized team problems [43], instead of (3.8), suppose γ_t^i were constructed as equilibrium on common agents' actions γ_t, i.e. for a fixed $\underline{\pi}_t, \pi_t = \prod_{i\in\mathcal{N}} \pi_t^i, \forall i \in \mathcal{N}$,

$$\tilde{\gamma}_t^i \in \arg\max_{\gamma_t^i} \mathbb{E}^{\gamma_t^i \tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \gamma_t^i \tilde{\gamma}_t^{-i}, A_t), X_{t+1}^i) \right\}. \tag{3.10}$$

It should be noted that in (3.10), the argument of the maximization operation, γ_t^i, appears both, in generation of action A_t^i and in the update of the belief π_t. Moreover, (3.10) is not conditioned on x_t^i, the private information of player i, similar to the case in the corresponding team problem. This is because the common agent who does not observe the private information of the player i, averages out that information. While this averaging of private information works for the team problem whose objective is to maximize the total expected reward, for the case with strategic players, it is incompatible with the sequential rationality condition in (3.6), which requires conditioning on the entire history $(a_{1:t-1}, x_t^i)$ and not just the common information $a_{1:t-1}$.

If the private information is also conditioned on, the construction still remains invalid, as discussed next.

(b) Instead of (3.8), suppose γ_t^i were constructed as best response of player i to other players actions $\tilde{\gamma}_t^{-i}$, similar to a standard Bayesian Nash equilibrium. For a fixed $\underline{\pi}_t, \pi_t = \prod_{i\in\mathcal{N}} \pi_t^i, \forall i \in \mathcal{N}, x_t^i \in \mathcal{X}^i$,

$$\tilde{\gamma}_t^i \in \arg\max_{\gamma_t^i} \mathbb{E}^{\gamma_t^i(\cdot|x_t^i) \tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \gamma_t^i \tilde{\gamma}_t^{-i}, A_t), X_{t+1}^i) \big| x_t^i \right\}. \tag{3.11}$$

Then $\tilde{\gamma}_t^i$ would be a function of $\tilde{\gamma}_t^{-i}$ and x_t^i through a best response relation $\tilde{\gamma}_t^i \in BR_{x_t^i}^i(\tilde{\gamma}_t^{-i})$, where $BR_{x_t^i}^i$ is appropriately defined from (3.62). Consequently, every component of the solution of the fixed point equation $(\tilde{\gamma}_t^i \in BR_{x_t^i}^i(\tilde{\gamma}_t^{-i}))_{x_t^i \in \mathcal{X}^i, i\in\mathcal{N}}$, if it existed, would be a function of the whole type profile x_t, resulting in a mapping $\tilde{\gamma}_t^i = \theta_t^i[\underline{\pi}_t, x_t]$. Since player i only observes its own type x_t^i, it would not be able to implement the corresponding $\tilde{\gamma}_t^i$, and therefore the construction would be invalid.

3.4.3 Forward recursion

As discussed above, a pair of strategy and belief profile (β^*, μ^*) is a PBE if it satisfies (3.6). Based on θ defined above in (3.7)–(3.9), we now construct a set of strategies β^* and beliefs

μ^* for the game $\mathfrak{D}$ in a forward recursive way, as follows[2]. As before, we will use the notation $\underline{\mu}_t^*[a_{1:t-1}] := (\mu_t^{*,i}[a_{1:t-1}])_{i\in\mathcal{N}}$ where $\mu_t^*[a_{1:t-1}]$ can be constructed from $\underline{\mu}_t^*[a_{1:t-1}]$ as $\mu_t^*[a_{1:t-1}](x_t) = \prod_{i=1}^N \mu_t^{*,i}[a_{1:t-1}](x_t^i)\ \forall a_{1:t-1} \in \mathcal{H}_t^c$ where $\mu_t^{*,i}[a_{1:t-1}]$ is a belief on x_t^i.

1. Initialize at time $t = 1$,

$$\mu_1^*[\phi](x_1) := \prod_{i=1}^N Q_1^i(x_1^i). \tag{3.12}$$

2. For $t = 1, 2 \dots T, \forall i \in \mathcal{N}, a_{1:t} \in \mathcal{H}_{t+1}^c, x_{1:t}^i \in (\mathcal{X}^i)^t$

$$\beta_t^{*,i}(a_t^i|a_{1:t-1}, x_{1:t}^i) = \beta_t^{*,i}(a_t^i|a_{1:t-1}, x_t^i) := \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]](a_t^i|x_t^i) \tag{3.13}$$

and

$$\mu_{t+1}^{*,i}[a_{1:t}] := \bar{F}(\mu_t^{*,i}[a_{1:t-1}], \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]], a_t) \tag{3.14}$$

where $\bar{F}$ is defined in the proof of Lemma 3.2 and in particular Claim 3.5.

We now state our main result.

Theorem 3.1. A strategy and belief profile (β^*, μ^*), constructed through backward/forward recursion algorithm described in Section 3.4 is a PBE of the game, i.e. $\forall i \in \mathcal{N}, t \in \mathcal{T}, a_{1:t-1} \in \mathcal{H}_t^c, x_{1:t}^i \in (\mathcal{X}^i)^t, \beta^i$,

$$\begin{aligned}
&\mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i},\,\mu_t^*[a_{1:t-1}]}\left\{\sum_{n=t}^T R^i(X_n, A_n)\Big|a_{1:t-1}, x_{1:t}^i\right\} \\
&\geq \mathbb{E}^{\beta_{t:T}^{i}\beta_{t:T}^{*,-i},\,\mu_t^*[a_{1:t-1}]}\left\{\sum_{n=t}^T R^i(X_n, A_n)\Big|a_{1:t-1}, x_{1:t}^i\right\}.
\end{aligned} \tag{3.15}$$

Proof. See Appendix C.

An intuitive explanation for why all players are able to use a common belief is the following. The sequence of beliefs defined above serve two purposes. First, for any player i, it puts a belief on x_t^{-i} to compute an expectation on the current and future rewards. Secondly, it predicts the actions of the other players since their strategies are functions of these beliefs. Since for any strategy profile, x_t^i is conditionally independent of x_t^{-i} given the common history $a_{1:t-1}$, and since other players do not observe x_t^i, knowledge of x_t^i does

[2] As discussed in starting of Section 3.4, beliefs at time t are functions of each agent's history h_t^i only through the common history h_t^c, and are the same for all agents.

not affect this belief and thus in our definition, all players can use the same belief μ^* which is independent of their private information.

Independence of types is a crucial assumption in proving the above result, which manifests itself in Lemma 3.5 in Appendix D, used in the proof of Theorem 3.1. This is because, at equilibrium, player i's reward-to-go at time t, conditioned on its type x_t^i, depends on its strategy at time t, β_t^i, only through its action a_t^i and is independent of the corresponding partial function $\beta_t^i(\cdot|a_{1:t-1}, \cdot)$. In other words, given x_t^i and a_t^i, player i's reward-to-go is independent of β_t^i. We discuss this in more detail below.

At equilibrium, all players observe past actions $a_{1:t-1}$ and update their belief π_t, which is the same as $\mu_t^*[a_{1:t-1}]$, through the equilibrium strategy profile β^*. Now suppose at time t, player i decides to unilaterally deviate to $\hat{\beta}_t^i$ at time t for some history $a_{1:t-1}$, keeping the rest of its strategy the same. Then other players still update their beliefs $(\pi_t)_{t\in\{t+1,\dots T\}}$ same as before and take their actions through equilibrium strategy $\beta_t^{*,-i}$ operated on π_t and x_t^{-i}, whereas player i forms a new belief $\hat{\pi}_{t+1}$ on x_t which depends on strategy profile $\beta_{1:t-1}^*\hat{\beta}_t, \beta_t^{*,-i}$. Thus, at time t, player i would need both the beliefs $\pi_{t+1}, \hat{\pi}_{t+1}$ to compute its expected future reward (as also discussed in [41]); π_{t+1} to predict other players' actions and $\hat{\pi}_{t+1}$ to form a true belief on x_t based on its information. As it turns out, due to independence of types, $\hat{\pi}_{t+1}$ does not provide additional information to player i to compute its future expected reward, and thus it can be discarded. Intuitively, this is so because the belief on type j, π_{t+1}^j is a function of strategy and action of player j till time t (as shown in Claim 3.5 in Appendix B); thus $\pi_{t+1}^{-i} = \hat{\pi}_{t+1}^{-i}$. Now since player i already observes its type x_t^i, its belief $\hat{\pi}_t^i$ on x_t^i does not provide any additional information to player i, and thus π_t (which is the same as $\mu_t^*[a_{1:t-1}]$) sufficiently computes future expected reward for player i. Also π_{t+1} is updated from π_t, $\beta_t^*(\cdot|a_{1:t-1}, \cdot)$ and a_t, and is independent of $\hat{\beta}_t^i$ given a_t^i. This implies player i can use the equilibrium strategy β_t^* to update its future belief, as used in (3.8). Then by construction of θ and specifically due to (3.8), player i does not gain by unilaterally deviating at time t keeping the remainder of its strategy the same.

Finally, we note that in the two-step backward-forward algorithm described above, once the equilibrium generating function θ is defined through backward recursion, the SPBEs can be generated through forward recursion for any prior distribution Q on types X. Since, in comparison to the backward recursion, the forward recursive part of the algorithm is computationally insignificant, the algorithm computes SPBEs for different prior distributions at the same time.

In the following lemma we show that all SPBE can be found through the backward-forward methodology described before. In general, an SPBE can be defined as a PBE (β^*, μ^*) of the game that is generated through forward recursion in (3.12)–(3.14), us-

ing an equilibrium generating function ϕ, where $\phi = (\phi_t^i)_{i\in\mathcal{N},t\in\mathcal{T}}$, $\phi_t^i : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{X}^i) \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{A}^i)\}$, common belief update function F^i and prior distributions Q^i. As a consequence, $\beta_t^{*,i}$ only depends on current type x_t^i of player i, and on the common information $a_{1:t-1}$ through the set of marginals $\underline{\mu}_t[a_{1:t-1}]$, and $\mu^{*,i}$ depends only on common information history $a_{1:t-1}$.

Lemma 3.3. Let (β^*, μ^*) be an SPBE. Then there exists an equilibrium generating function ϕ that satisfies (3.8) in backward recursion $\forall\ \pi_t = \mu_t^*[a_{1:t-1}],\ \forall\ a_{1:t-1}$, such that (β^*, μ^*) is defined through forward recursion using ϕ [3] .

Proof. See Appendix E.

3.4.4 Existence

While it is known that for any finite dynamic game with asymmetric information and perfect recall, there always exists a PBE [45, Prop. 249.1], existence of the fixed point equation in (3.8) is an unresolved question. Generally, existence of a fixed point equation is shown through Kakutani's fixed point theorem, as is done in proving existence of a mixed strategy Nash equilibrium for any finite game [40, 45] by showing existence of fixed point of the best-response correspondences of the game. Among other conditions, it requires closed graph property of the correspondences, which is implied by the continuity property of the utility functions involved. For (3.8), continuity of the term to be optimized, with respect to actions γ_t^i, is not guaranteed. This is due to two reasons: (a) potential discontinuity of the π_t update function F when the denominator in the Bayesian update is 0, (b) as it is observed in the numerical example in the next section, the value functions, V_t^i, need not be continuous. Thus the standard arguments for existence of the fixed point equation can not be directly applied and existence of solution of (3.8) remains an open question.

In the next section, we discuss an example to illustrate the methodology described above for the construction of SPBEs.

3.5 Illustrative example: A two stage public goods game

We consider a discrete version of Example 8.3 from [13, ch.8], which is an instance of a repeated public good game. There are two players who play a two period game. In each period t, they simultaneously decide whether to contribute to the period t public good,

[3] Note that for $\underline{\pi}_t \neq \underline{\mu}_t[a_{1:t-1}]$ for any $a_{1:t-1}$, ϕ can be arbitrarily defined without affecting the definition of (β^*, μ^*)

which is a binary decision $a_t^i \in \{0,1\}$ for player $i = 1,2$. Before the start of period 2, both players know the actions taken by them in period 1. For both periods, each player gets reward 1 if at least one of them contributed and 0 if none does. Player i's cost of contributing is x^i which is its private information. Both players believe that x^is are drawn independently and identically with probability distribution Q with support $\{x^L, x^H\}$; $0 < x^L < 1 < x^H$, such that $P^Q(X^i = x^H) = q$ where $0 < q < 1$.

This example is similar to our model where $N = 2, T = 2$ and reward for player i in period t is

$$R^i(x, a_t) = \begin{cases} a_t^{-i} & \text{if } a_t^i = 0 \\ 1 - x^i & \text{if } a_t^i = 1. \end{cases} \tag{3.16}$$

We will use the backward recursive algorithm, defined in Section 3.4, to find an SPBE of this game. For period $t = 1, 2$ and for $i = 1, 2$, the partial functions γ_t^i can equivalently be defined through scalars p_t^{iL} and p_t^{iH} such that $\gamma_t^i(1|x^L) = p_t^{iL}$, $\gamma_t^i(0|x^L) = 1 - p_t^{iL}$ and $\gamma_t^i(1|x^H) = p_t^{iH}$, $\gamma_t^i(0|x^H) = 1 - p_t^{iH}$, where $p_t^{iL}, p_t^{iH} \in [0,1]$. Henceforth, we will use p_t^{iL} and p_t^{iH} interchangeably with the corresponding γ_t^i.

For $t = 2$ and for any fixed $\underline{\pi}_2 = (\pi_2^1, \pi_2^2)$, where $\pi_2^i = \pi_2^i(x^H) \in [0,1]$ represents a probability measure on the event $\{X^i = x^H\}$, player i's reward is

$$\mathbb{E}^{\gamma_2}\{R_2^i(X, A_2)|\pi_2, X^i = x^L\} = (1 - p_2^{iL})\left((1 - \pi_2^{-i})p_2^{-iL} + \pi_2^{-i}p_2^{-iH}\right) + p_2^{iL}(1 - x^L), \tag{3.17a}$$

$$\mathbb{E}^{\gamma_2}\{R_2^i(X, A_2)|\pi_2, X^i = x^H\} = (1 - p_2^{iH})\left((1 - \pi_2^{-i})p_2^{-iL} + \pi_2^{-i}p_2^{-iH}\right) + p_2^{iH}(1 - x^H). \tag{3.17b}$$

Let $\tilde{\gamma}_2 = \theta_2[\underline{\pi}_2]$ and equivalently $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = \theta_2[\pi_2]$ be defined through the following fixed point equation, which is equivalent to (3.8). For $i = 1, 2$

$$\tilde{p}_2^{iL} \in \arg\max_{p_2^{iL}} \quad (1 - p_2^{iL})\left((1 - \pi_2^{-i})\tilde{p}_2^{-iL} + \pi_2^{-i}\tilde{p}_2^{-iH}\right) + p_2^{iL}(1 - x^L), \tag{3.18a}$$

$$\tilde{p}_2^{iH} \in \arg\max_{p_2^{iH}} \quad (1 - p_2^{iH})\left((1 - \pi_2^{-i})\tilde{p}_2^{-iL} + \pi_2^{-i}\tilde{p}_2^{-iH}\right) + p_2^{iH}(1 - x^H). \tag{3.18b}$$

Since $1 - x^H < 0$, $\tilde{p}_2^{iH} = 0$ achieves the maximum in (3.18b). Thus (3.18a)–(3.18b) can be reduced to, $\forall i \in \{1, 2\}$

$$\tilde{p}_2^{iL} \in \arg\max_{p_2^{iL}} \quad (1 - p_2^{iL})(1 - \pi_2^{-i})\tilde{p}_2^{-iL} + p_2^{iL}(1 - x^L). \tag{3.19}$$

This implies,

$$\tilde{p}_2^{iL} = \begin{cases} 0 & \text{if} \quad x^L > 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}, \\ 1 & \text{if} \quad x^L < 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}, \\ \text{arbitrary} & \text{if} \quad x^L = 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}. \end{cases} \tag{3.20}$$

The fixed point equation (3.20) has the following solutions,

1. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (0, 1, 0, 0)$ for $\pi_2^1 \in [0, 1], \pi_2^2 \leq x^L$
   - $V_2^1(\pi_2, x^L) = 1 - \pi_2^2$
   - $V_2^1(\pi_2, x^H) = 1 - \pi_2^2$
   - $V_2^2(\pi_2, x^L) = 1 - x^L$
   - $V_2^2(\pi_2, x^H) = 0.$
2. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (1, 0, 0, 0)$ for $\pi_2^1 \leq x^L, \pi_2^1 \in [0, 1]$
   - $V_2^1(\pi_2, x^L) = 1 - x^L$
   - $V_2^1(\pi_2, x^H) = 0$
   - $V_2^2(\pi_2, x^L) = 1 - \pi_2^1$
   - $V_2^2(\pi_2, x^H) = 1 - \pi_2^1.$
3. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (1, 1, 0, 0)$ for $\pi_2^1 \geq x^L, \pi_2^2 \geq x^L$
   - $V_2^1(\pi_2, x^L) = 1 - x^L$
   - $V_2^1(\pi_2, x^H) = 1 - \pi_2^2$
   - $V_2^2(\pi_2, x^L) = 1 - x^L$
   - $V_2^2(\pi_2, x^H) = 1 - \pi_2^1.$
4. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (1, \tilde{p}_2^{2L}, 0, 0)$ for $\pi_2^1 = x^L, \pi_2^2 \in [0, 1]$ where $\tilde{p}_2^{2L} \in \left[0, \max\left\{1, \frac{1-x^L}{1-\pi_2^2}\right\}\right]$
   - $V_2^1(\pi_2, x^L) = 1 - x^L$
   - $V_2^1(\pi_2, x^H) = 1 - \pi_2^2.\tilde{p}_2^{2L}$
   - $V_2^2(\pi_2, x^L) = 1 - x^L$
   - $V_2^2(\pi_2, x^H) = 1 - \pi_2^1.$

5. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (\tilde{p}_2^{1L}, 1, 0, 0)$ for $\pi_2^1 \in [0,1], \pi_2^2 = x^L$ where $\tilde{p}_2^{1L} \in \left[0, \max\left\{1, \frac{1-x^L}{1-\pi_2^1}\right\}\right]$
   - $V_2^1(\pi_2, x^L) = 1 - x^L$
   - $V_2^1(\pi_2, x^H) = 1 - \pi_2^2$
   - $V_2^2(\pi_2, x^L) = 1 - x^L$
   - $V_2^2(\pi_2, x^H) = 1 - \pi_2^1.\tilde{p}_2^{1L}$.
6. $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = (\frac{1-x^L}{1-\pi_2^1}, \frac{1-x^L}{1-\pi_2^2}, 0, 0)$ for $\pi_2^1 \leq x^L, \pi_2^2 \leq x^L$
   - $V_2^1(\pi_2, x^L) = 1 - x^L$
   - $V_2^1(\pi_2, x^H) = 1 - x^L$
   - $V_2^2(\pi_2, x^L) = 1 - x^L$
   - $V_2^2(\pi_2, x^H) = 1 - x^L$.

Figure 3.1 shows these solutions in the space of (π_2^1, π_2^2).

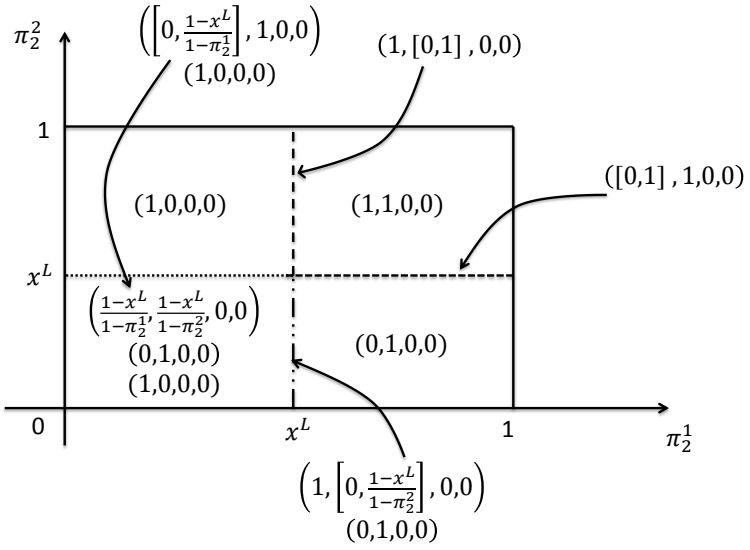


Figure 3.1: Solutions of fixed point equation in (3.20)

Thus for any $\underline{\pi}_2$, there can exist multiple equilibria and correspondingly multiple $\theta_2[\underline{\pi}_2]$ can be defined. For any particular θ_2, at $t = 1$, the fixed point equation that needs to be

solved is of the form, $\forall i \in \{1,2\}$

$$\tilde{p}_1^{iL} \in \arg\max_{p_1^{iL}} \quad (1-p_1^{iL})\left((1-q)\tilde{p}_1^{-iL} + q\tilde{p}_1^{-iH} + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(F(Q^2,\tilde{\gamma}_1,(0,A_1^{-i})),x^L)\}\right)$$
$$+ p_1^{iL}\left(1 - x^L + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(F(Q^2,\tilde{\gamma}_1,(1,A_1^{-i})),x^L)\}\right). \tag{3.21a}$$
$$\tilde{p}_1^{iH} \in \arg\max_{p_1^{iH}} \quad (1-p_1^{iH})\left((1-q)\tilde{p}_1^{-iL} + q\tilde{p}_1^{-iH} + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(F(Q^2,\tilde{\gamma}_1,(0,A_1^{-i})),x^H)\}\right)$$
$$+ p_1^{iH}\left(1 - x^H + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(F(Q^2,\tilde{\gamma}_1,(1,A_1^{-i})),x^H)\}\right). \tag{3.21b}$$

where $F(Q^2,\tilde{\gamma},(A^1,A^2)) = \bar{F}(Q,\tilde{\gamma}^1,A^1)\bar{F}(Q,\tilde{\gamma}^2,A^2)$ and

$$\bar{F}(Q,\tilde{\gamma}_1^i,0) = \frac{q(1-\tilde{p}_1^{iH})}{q(1-\tilde{p}_1^{iH}) + (1-q)(1-\tilde{p}_1^{iL})}, \tag{3.22a}$$
$$\bar{F}(Q,\tilde{\gamma}_1^i,1) = \frac{q\tilde{p}_1^{iH}}{q\tilde{p}_1^{iH} + (1-q)\tilde{p}_1^{iL}}, \tag{3.22b}$$

if the denominators in (3.22a)–(3.22b) are strictly positive, else $\bar{F}(Q,\tilde{\gamma}_1^i,A^i) = Q$ as in the proof of Lemma 3.2, and in particular Claim 3.5. A solution of the fixed point equation in (3.21a)-(3.21b) defines $\theta_1[Q^2]$.

Using one such θ defined as follows, we find an SPBE of the game for $q = 0.1, x^L = 0.2, x^H = 1.2$. We use $\theta_2[\underline{\pi}_2]$ as one possible set of solutions of (3.20), shown in Figure 3.2 and described below,

$$\theta_2[\pi_2] = (\tilde{p}_2^{1L},\tilde{p}_2^{2L},\tilde{p}_2^{1H},\tilde{p}_2^{2H}) = \begin{cases} (\frac{1-x^L}{1-\pi_2^1},\frac{1-x^L}{1-\pi_2^2},0,0) & \pi_2^1 \in [0,x^L), \pi_2^2 \in [0,x^L) \\ (1,0,0,0) & \pi_2^1 \in [0,x^L], \pi_2^2 \in [x^L,1] \\ (0,1,0,0) & \pi_2^1 \in [x^L,1], \pi_2^2 \in [0,x^L] \\ (1,1,0,0) & \pi_2^1 \in (x^L,1], \pi_2^2 \in (x^L,1]. \end{cases} \tag{3.23}$$

Then, through iteration on the fixed point equation (3.21a)-(3.21b) and using the aforementioned $\theta_2[\underline{\pi}_2]$, we numerically find (and analytically verify) that $\theta_1[Q^2] = (\tilde{p}_1^{1L},\tilde{p}_1^{2L},\tilde{p}_1^{1H},\tilde{p}_1^{2H}) = (0,1,0,0)$ is a fixed point. Thus

$$\beta_1^1(A_1^1 = 1|X^1 = x^L) = 0 \qquad \beta_1^2(A_1^2 = 1|X^2 = x^L) = 1$$
$$\beta_1^1(A_1^1 = 1|X^1 = x^H) = 0 \qquad \beta_1^2(A_1^2 = 1|X^2 = x^H) = 0$$

with beliefs $\mu_2^*[00] = (q,1), \mu_2^*[01] = (q,0), \mu_2^*[10] = (q,1), \mu_2^*[11] = (q,0)$ and

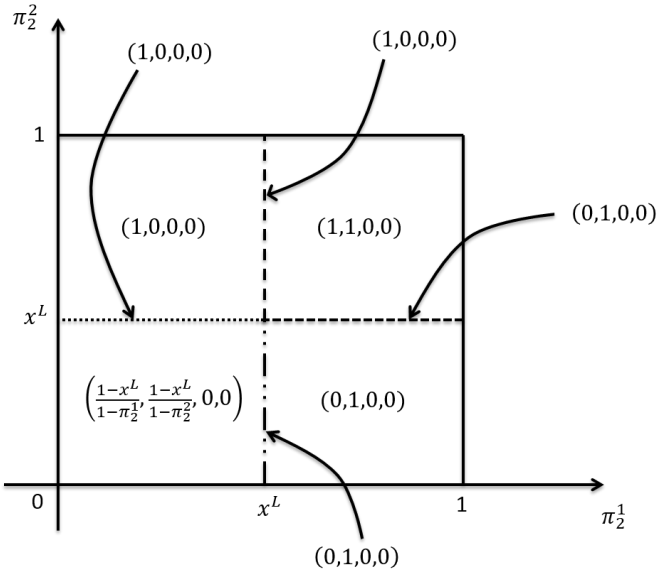


Figure 3.2: $\theta_2[\pi_2]$ described in (3.23)

$(\beta_2^i(\cdot|a_1,\cdot))_{i\in\{1,2\}} = \theta_2[\mu_2^*[a_1]]$ is an SPBE of the game. In this equilibrium, player 2 at time $t = 1$, contributes according to her type whereas player 1 never contributes, thus player 2 reveals her private information through her action whereas player 1 does not. Since θ_2 is symmetric, there also exists an (antisymmetric) equilibrium where at time $t = 1$, players' strategies reverse i.e. player 2 never contributes and player 1 contributes according to her type. We also obtain a symmetric equilibrium where $\theta_1[Q^2] = (\frac{1-x^L}{(1-q)(1+x^L)}, \frac{1-x^L}{(1-q)(1+x^L)}, 0, 0)$ as a fixed point when $x^L > \frac{q}{2-q}$, resulting in beliefs $\mu_2^*[00] = (p,p), \mu_2^*[01] = (p,0), \mu_2^*[10] = (0,p), \mu_2^*[11] = (0,0)$ where $p = \frac{q(1+x^L)}{q(1+x^L)+(1-x^L)}$.

3.6 Conclusion

In this chapter, we study a class of dynamic games with asymmetric information where player i observes its true private type x_t^i and together with other players, observe past actions of everybody else. The types of the players evolve as conditionally independent, controlled Markov processes, conditioned on players current actions. We present a two-step backward-forward recursive algorithm to find SPBE of this game, where equilibrium strategies are function of a Markov belief state π_t, which depends on the common information, and current private types of the players. The backward recursive part of this algorithm defines an equilibrium generating function. Each period in backward recursion involves solving a fixed point equation on the space of probability simplexes for every possible belief on types. Then using this function, equilibrium strategies and beliefs are defined

through a forward recursion.

In this chapter we consider perfectly observable, independent dynamic types of the players. In chapter 5, we consider the case where players do not perfectly observe their types, rather they make independent, noisy observations. In general, this methodology opens the door for finding PBEs for many applications, analytically or numerically, which was not feasible before. One such case would be dynamic LQG games where types evolve linearly with Gaussian noise and players incur quadratic cost, which we discuss in next chapter.

3.7 Appendix A (Proof of Lemma 3.1)

We prove this lemma in the following steps.

(a) In Claim 3.1, we prove that for any policy profile g and $\forall t \in \mathcal{T}$, $x^i_{1:t}$ for $i \in \mathcal{N}$ are conditionally independent given the common information $a_{1:t}$.

(b) In Claim 3.2, using Claim 3.1, we prove that for every fixed strategy g^{-i} of the players $-i$, $((a_{1:t-1}, x^i_t), a^i_t)_{t\in\mathcal{T}}$ is a controlled Markov process for player i.

(c) For a given policy g, we define a policy s^i of player i from g as $s^i_t(a^i_t|a_{1:t-1}, x^i_t) \stackrel{\triangle}{=} P^g(a^i_t|a_{1:t-1}, x^i_t)$.

(d) In Claim 3.3, we prove that the dynamics of this controlled Markov process $((x^i_t, a_{1:t-1}), a^i_t)_{t\in\mathcal{T}}$ under $(s^i g^{-i})$ are same as under g i.e. $P^{s^i g^{-i}}(x^i_t, x^i_{t+1}, a_{1:t}) = P^g(x^i_t, x^i_{t+1}, a_{1:t})$.

(e) In Claim 3.4, we prove that w.r.t. random variables (x_t, a_t), x^i_t is sufficient for player i's private information history $x^i_{1:t}$ i.e. $P^g(x_t, a_t|a_{1:t-1}, x^i_{1:t}, a^i_t) = P^{g^{-i}}(x_t, a_t|a_{1:t-1}, x^i_t, a^i_t)$.

(f) From (c), (d) and (e) we then prove the result of the lemma that $P^{s^i g^{-i}}(x_t, a_t) = P^g(x_t, a_t)$.

Claim 3.1. For any policy profile g and $\forall t$,

$$P^g(x_{1:t}|a_{1:t-1}) = \prod_{i=1}^{N} P^{g^i}(x^i_{1:t}|a_{1:t-1}) \tag{3.25}$$

Proof.

$$
\begin{aligned}
&P^g(x_{1:t}|a_{1:t-1}) \\
&= \frac{P^g(x_{1:t}, a_{1:t-1})}{\sum_{\bar{x}_{1:t}} P^g(\bar{x}_{1:t}, a_{1:t-1})} && \text{(3.26a)} \\
&= \frac{\prod_{i=1}^N \left(Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)\right)}{\sum_{\bar{x}_{1:t}} \prod_{i=1}^N \left(Q^i(\bar{x}_1^i) g_1^i(a_1^i|\bar{x}_1^i) \prod_{n=2}^t Q_n^i(\bar{x}_n^i|\bar{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \bar{x}_{1:n}^i)\right)} && \text{(3.26b)} \\
&= \frac{\prod_{i=1}^N \left(Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)\right)}{\prod_{i=1}^N \left(\sum_{\bar{x}_{1:t}^i} Q^i(\bar{x}_1^i) g_1^i(a_1^i|\bar{x}_1^i) \prod_{n=2}^t Q_n^i(\bar{x}_n^i|\bar{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \bar{x}_{1:n}^i)\right)} && \text{(3.26c)} \\
&= \prod_{i=1}^N \frac{Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)}{\sum_{\bar{x}_{1:t}^i} Q^i(\bar{x}_1^i) g_1^i(a_1^i|\bar{x}_1^i) \prod_{n=2}^t Q_n^i(\bar{x}_n^i|\bar{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \bar{x}_{1:n}^i)} && \text{(3.26d)} \\
&= \prod_{i=1}^N P^{g^i}(x_{1:t}^i|a_{1:t-1}) && \text{(3.26e)}
\end{aligned}
$$

Claim 3.2. For a fixed g^{-i}, $\{(a_{1:t-1}, x_t^i), a_t^i\}_t$ is a controlled Markov process with state $(a_{1:t-1}, x_t^i)$ and control action a_t^i.

Proof.

$$
\begin{aligned}
&P^g(\tilde{a}_{1:t}, x_{t+1}^i|a_{1:t-1}, x_{1:t}^i, a_{1:t}^i) \\
&= \sum_{x_{1:t}^{-i}} P^g(\tilde{a}_{1:t}, x_{t+1}^i, x_{1:t}^{-i}|a_{1:t-1}, x_{1:t}^i, a_t^i) && \text{(3.27a)} \\
&= \sum_{x_{1:t}^{-i}} P^g(\tilde{a}_t^{-i}, x_{t+1}^i, x_{1:t}^{-i}|a_{1:t-1}, x_{1:t}^i, a_t^i) I_{(a_{1:t-1}, a_t^i)}(\tilde{a}_{1:t-1}, \tilde{a}_t^i) && \text{(3.27b)} \\
&= \sum_{x_{1:t}^{-i}} P^{g^{-i}}(x_{1:t}^{-i}|a_{1:t-1}) \left(\prod_{j \neq i} g_t^j(\tilde{a}_t^j|a_{1:t-1}, x_{1:t}^j)\right) Q_t^i(x_{t+1}^i|x_t^i, a_t^i, \tilde{a}_t^{-i}) I_{(a_{1:t-1}, a_t^i)}(\tilde{a}_{1:t-1}, \tilde{a}_t^i) \\
& && \text{(3.27c)} \\
&= P^{g^{-i}}(\tilde{a}_{1:t}, x_{t+1}^i|a_{1:t-1}, x_t^i, a_t^i), && \text{(3.27d)}
\end{aligned}
$$

where (3.27c) follows from Claim 3.1 since $x_{1:t}^{-i}$ is conditionally independent of $x_{1:t}^i$ given $a_{1:t-1}$ and the corresponding probability is only a function of g^{-i}.

For any given policy profile g, we construct a policy s^i in the following way,

$$s_t^i(a_t^i|a_{1:t-1}, x_t^i) \triangleq P^g(a_t^i|a_{1:t-1}, x_t^i) \tag{3.28}$$

$$= \frac{\sum_{x_{1:t-1}^i} P^g(a_t^i, x_{1:t}^i|a_{1:t-1})}{\sum_{\tilde{a}_t^i}\sum_{\tilde{x}_{1:t-1}^i} P^g(\tilde{a}_t^i, \tilde{x}_{1:t-1}^i x_t^i|a_{1:t-1})} \tag{3.29}$$

$$= \frac{\sum_{x_{1:t-1}^i} P^{g^i}(x_{1:t}^i|a_{1:t-1}) g_t^i(a_t^i|a_{1:t-1}, x_{1:t}^i)}{\sum_{\tilde{a}_t^i}\sum_{\tilde{x}_{1:t-1}^i} P^{g^i}(\tilde{x}_{1:t-1}^i x_t^i|a_{1:t-1}) g_t^i(\tilde{a}_t^i|a_{1:t-1}, \tilde{x}_{1:t-1}^i x_t^i)} \tag{3.30}$$

$$= P^{g^i}(a_t^i|a_{1:t-1}, x_t^i), \tag{3.31}$$

where dependence of (3.30) on only g^i is due to Claim 3.1.

Claim 3.3. The dynamics of the Markov process $\{(x_t^i, a_{1:t-1}), a_t^i\}_t$ under $(s^i g^{-i})$ are the same as under g i.e.

$$P^{s^i g^{-i}}(x_t^i, x_{t+1}^i, a_{1:t}) = P^g(x_t^i, x_{t+1}^i, a_{1:t}) \quad \forall t \tag{3.32}$$

Proof. We prove this by induction. Clearly,

$$P^g(x_1^i) = P^{s^i g^{-i}}(x_1^i) = Q_1^i(x_1^i) \tag{3.33}$$

Now suppose (3.32) is true for $t-1$ which also implies that the marginals $P^g(x_t^i, a_{1:t-1}) = P^{s^i g^{-i}}(x_t^i, a_{1:t-1})$. Then

$$P^g(x_t^i, a_{1:t-1}, x_{t+1}^i, a_t) = P^g(x_t^i, a_{1:t-1}) P^g(a_t^i|a_{1:t-1}, x_t^i) P^g(x_{t+1}^i, a_{1:t}|x_t^i, a_{1:t-1}, a_t^i) \tag{3.34a}$$

$$= P^{s^i g^{-i}}(x_t^i, a_{1:t-1}) s_t^i(a_t^i|a_{1:t-1}, x_t^i) P^{g^{-i}}(x_{t+1}^i, a_{1:t}|x_t^i, a_{1:t-1}, a_t^i) \tag{3.34b}$$

$$= P^{s^i g^{-i}}(x_t^i, a_{1:t-1}, x_{t+1}^i, a_t) \tag{3.34c}$$

where (3.34b) is true from induction hypothesis, definition of s^i in (3.31) and since $\{(a_{1:t-1}, x_t^i), a_t^i\}_t$ is a controlled Markov process as proved in Claim 3.2 and its update kernel does not depend on policy g^i.This completes the induction step.

Claim 3.4. For any policy g,

$$P^g(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_{1:t}^i, a_t^i) = P^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i) \tag{3.35}$$

Proof.

$$P^g(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_{1:t}, a^i_t) = I_{x^i_t, a^i_t}(\tilde{x}^i_t, \tilde{a}^i_t) P^g(\tilde{x}^{-i}_t, \tilde{a}^{-i}_t | a_{1:t-1}, x^i_{1:t}) \tag{3.36}$$

Now

$$\begin{aligned}
P^g(\tilde{x}^{-i}_t, \tilde{a}^{-i}_t | a_{1:t-1}, x^i_{1:t}) &= \sum_{\tilde{x}^{-i}_{1:t-1}} P^g(\tilde{x}^{-i}_{1:t}, \tilde{a}^{-i}_t | a_{1:t-1}, x^i_{1:t}) && \text{(3.37a)} \\
&= \sum_{\tilde{x}^{-i}_{1:t-1}} P^g(\tilde{x}^{-i}_{1:t} | a_{1:t-1}, x^i_{1:t}) \left(\prod_{j \neq i} g^j_t(\tilde{a}^j_t | a_{1:t-1}, \tilde{x}^j_{1:t}) \right) && \text{(3.37b)} \\
&= \sum_{\tilde{x}^{-i}_{1:t}} P^{g^{-i}}(\tilde{x}^{-i}_{1:t} | a_{1:t-1}) \left(\prod_{j \neq i} g^j_t(\tilde{a}^j_t | a_{1:t-1}, \tilde{x}^j_{1:t}) \right) && \text{(3.37c)} \\
&= P^{g^{-i}}(\tilde{x}^{-i}_t, \tilde{a}^{-i}_t | a_{1:t-1}) && \text{(3.37d)}
\end{aligned}$$

where (3.37c) follows from Claim 3.1.

Hence

$$\begin{aligned}
P^g(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_{1:t}, a^i_t) &= I_{x^i_t, a^i_t}(\tilde{x}^i_t, \tilde{a}^i_t) P^{g^{-i}}(\tilde{x}^{-i}_t, \tilde{a}^{-i}_t | a_{1:t-1}) && \text{(3.38a)} \\
&= P^{g^{-i}}(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_t, a^i_t) && \text{(3.38b)}
\end{aligned}$$

Finally,

$$\begin{aligned}
P^g(\tilde{x}_t, \tilde{a}_t) &= \sum_{a_{1:t-1} x^i_{1:t} a^i_t} P^g(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_{1:t}, a^i_t) P^g(a_{1:t-1}, x^i_{1:t}, a^i_t) && \text{(3.39a)} \\
&= \sum_{a_{1:t-1} x^i_{1:t}, a^i_t} P^{g^{-i}}(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_t, a^i_t) P^g(a_{1:t-1}, x^i_{1:t}, a^i_t) && \text{(3.39b)} \\
&= \sum_{a_{1:t-1} x^i_t, a^i_t} P^{g^{-i}}(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_t, a^i_t) P^g(a_{1:t-1}, x^i_t, a^i_t) && \text{(3.39c)} \\
&= \sum_{a_{1:t-1} x^i_t, a^i_t} P^{g^{-i}}(\tilde{x}_t, \tilde{a}_t | a_{1:t-1}, x^i_t, a^i_t) P^{s^i g^{-i}}(a_{1:t-1}, x^i_t, a^i_t) && \text{(3.39d)} \\
&= P^{s^i g^{-i}}(\tilde{x}_t, \tilde{a}_t). && \text{(3.39e)}
\end{aligned}$$

where (3.39b) follows from (3.35) in Claim 3.4 and (3.39d) from (3.32) in Claim 3.3.

3.8 Appendix B (Proof of Lemma 3.2)

For this proof we will assume the common agents strategies to be probabilistic as opposed to being deterministic, as was the case in Section 3.3. This means actions of the common agent, γ_t^i's are generated probabilistically from ψ^i as $\Gamma_t^i \cdot \psi_t^i(\cdot|a_{1:t-1})$, as opposed to being deterministically generated as $\gamma_t^i = \psi_t^i[a_{1:t-1}]$, as before. These two are equivalent ways of generating actions a_t^i from $a_{1:t-1}$ and x_t^i. We avoid using the probabilistic strategies of common agent throughout the main text for ease of exposition, and because it conceptually does not affect the results.

Proof. We prove this lemma in the following steps. We view this problem from the perspective of a common agent. Let ψ be the coordinator's policy corresponding to policy profile g. Let $\pi_t^i(x_t^i) = P^{\psi^i}(x_t^i|a_{1:t-1})$.

(a) In Claim 3.5, we show that π_t can be factorized as $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(x_t^i)$ where each π_t^i can be updated through an update function $\pi_{t+1}^i = \bar{F}(\pi_t^i, \gamma_t^i, a_t)$ and $\bar{F}$ is independent of common agent's policy ψ.

(b) In Claim 3.6, we prove that $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ is a controlled Markov process.

(c) We construct a policy profile θ from g such that $\theta_t(d\gamma_t|\pi_t) \stackrel{\triangle}{=} P^{\psi}(d\gamma_t|\pi_t)$.

(d) In Claim 3.7, we prove that dynamics of this Markov process $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ under θ is same as under ψ i.e. $P^{\theta}(d\pi_t, d\gamma_t, d\pi_{t+1}) = P^{\psi}(d\pi_t, d\gamma_t, d\pi_{t+1})$.

(e) In Claim 3.8, we prove that with respect to random variables (X_t, A_t), π_t can summarize common information $a_{1:t-1}$ i.e. $P^{\psi}(x_t, a_t|a_{1:t-1}, \gamma_t) = P(x_t, a_t|\pi_t, \gamma_t)$.

(f) From (c), (d) and (e) we that prove the result of the lemma that $P^{\psi}(x_t, a_t) = P^{\theta}(x_t, a_t)$ which is equivalent to $P^g(x_t, a_t) = P^m(x_t, a_t)$, where m is the policy profile of players corresponding to θ .

Claim 3.5. π_t can be factorized as $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(x_t^i)$ where each π_t^i can be updated through an update function $\pi_{t+1}^i = \bar{F}(\pi_t^i, \gamma_t^i, a_t)$ and $\bar{F}$ is independent of common agent's policy ψ. We also say $\underline{\pi}_{t+1} = F(\underline{\pi}_t, \gamma_t, a_t)$.

Proof. We prove this by induction. Since $\pi_1(x_1) = \prod_{i=1}^N Q_t^i(x_1^i)$, the base case is verified.

Now suppose $\pi_t = \prod_{i=1}^N \pi_t^i$. Then,

$$\pi_{t+1}(x_{t+1}) = P^\psi(x_{t+1}|a_{1:t}, \gamma_{1:t+1}) \tag{3.40a}$$
$$= P^\psi(x_{t+1}|a_{1:t}, \gamma_{1:t}) \tag{3.40b}$$
$$= \frac{\sum_{x_t} P^\psi(x_t, a_t, x_{t+1}|a_{1:t-1}, \gamma_{1:t})}{\sum_{\tilde{x}_{t+1}\tilde{x}_t} P^\psi(\tilde{x}_t, \tilde{x}_{t+1}, a_t|a_{1:t-1}, \gamma_{1:t})} \tag{3.40c}$$
$$= \frac{\sum_{x_t} \pi_t(x_t) \prod_{i=1}^N \gamma_t^i(a_t^i|x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t)}{\sum_{\tilde{x}_t\tilde{x}_{t+1}} \pi_t(\tilde{x}_t) \prod_{i=1}^N \gamma_t^i(a_t^i|\tilde{x}_t^i) Q_t^i(\tilde{x}_{t+1}^i|\tilde{x}_t^i, a_t)} \tag{3.40d}$$
$$= \prod_{i=1}^N \frac{\sum_{x_t^i} \pi_t^i(x_t^i)\gamma_t^i(a_t^i|x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t)}{\sum_{\tilde{x}_t^i} \pi_t^i(\tilde{x}_t^i)\gamma_t^i(a_t^i|\tilde{x}_t^i)}, \tag{3.40e}$$
$$= \prod_{i=1}^N \pi_{t+1}^i(x_{t+1}^i) \tag{3.40f}$$

where (3.40e) follows from induction hypothesis. It is assumed in (3.40c)-(3.40e) that the denominator is not 0. If denominator corresponding to any γ_t^i is zero, we define

$$\pi_{t+1}^i(x_{t+1}^i) = \sum_{x_t^i} \pi_t^i(x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t), \tag{3.41}$$

where π_{t+1} still satisfies (3.40f). Thus $\pi_{t+1}^i = \bar{F}(\pi_t^i, \gamma_t^i, a_t)$ and $\underline{\pi}_{t+1} = F(\underline{\pi}_t, \gamma_t, a_1)$ where $\bar{F}$ and F are appropriately defined from above.

Claim 3.6. $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ is a controlled Markov process with state Π_t and control action Γ_t

Proof.

$$P^\psi(d\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}) = \sum_{a_t, x_t} P^\psi(d\pi_{t+1}, a_t, x_t|\pi_{1:t}, \gamma_{1:t}) \tag{3.42a}$$
$$= \sum_{a_t, x_t} P^\psi(x_t|\pi_{1:t}, \gamma_{1:t}) \left\{ \prod_{i=1}^N \gamma_t^i(a_t^i|x_t^i) \right\} I_{F(\pi_t, \gamma_t, a_t)}(\pi_{t+1}) \tag{3.42b}$$
$$= \sum_{a_t, x_t} \pi_t(x_t) \left\{ \prod_{i=1}^N \gamma_t^i(a_t^i|x_t^i) \right\} I_{F(\pi_t, \gamma_t, a_t)}(\pi_{t+1}) \tag{3.42c}$$
$$= P(d\pi_{t+1}|\pi_t, \gamma_t). \tag{3.42d}$$

For any given policy profile ψ, we construct policy profile θ in the following way.

$$\theta_t(d\gamma_t|\pi_t) \stackrel{\triangle}{=} P^{\psi}(d\gamma_t|\pi_t). \tag{3.43}$$

Claim 3.7.

$$P^{\psi}(d\pi_t, d\gamma_t, d\pi_{t+1}) = P^{\theta}(d\pi_t, d\gamma_t, d\pi_{t+1}) \quad \forall t \in \mathcal{T}. \tag{3.44}$$

Proof. We prove this by induction. For $t = 1$,

$$P^{\psi}(d\pi_1) = P^{\theta}(d\pi_1) = I_Q(\pi_1). \tag{3.45}$$

Now suppose $P^{\psi}(d\pi_t) = P^{\theta}(d\pi_t)$ is true for t, then

$$\begin{aligned} P^{\psi}(d\pi_t, d\gamma_t, d\pi_{t+1}) &= P^{\psi}(d\pi_t)P^{\psi}(d\gamma_t|\pi_t)P^{\psi}(d\pi_{t+1}|\pi_t\gamma_t) && \text{(3.46a)} \\ &= P^{\theta}(d\pi_t)\theta_t(d\gamma_t|\pi_t)P(d\pi_{t+1}|\pi_t, \gamma_t) && \text{(3.46b)} \\ &= P^{\theta}(d\pi_t, d\gamma_t, d\pi_{t+1}). && \text{(3.46c)} \end{aligned}$$

where (3.46b) is true from induction hypothesis, definition of θ in (3.43) and since $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ is a controlled Markov process as proved in Claim 3.6 and thus its update kernel does not depend on policy ψ. This completes the induction step.

Claim 3.8. For any policy ψ,

$$P^{\psi}(x_t, a_t|a_{1:t-1}, \gamma_t) = P(x_t, a_t|\pi_t, \gamma_t). \tag{3.47}$$

Proof.

$$\begin{aligned} P^{\psi}(x_t, a_t|a_{1:t-1}, \gamma_t) &= P^{\psi}(x_t|a_{1:t-1}, \gamma_t)\prod_{i\in\mathcal{N}}\gamma_t^i(a_t^i|x_t^i) && \text{(3.48a)} \\ &= \pi_t(x_t)\prod_{i\in\mathcal{N}}\gamma_t^i(a_t^i|x_t^i) && \text{(3.48b)} \\ &= P(x_t, a_t|\pi_t, \gamma_t). && \text{(3.48c)} \end{aligned}$$

Finally,

$$P^{\psi}(x_t, a_t) = \sum_{a_{1:t-1}, \gamma_t} P^{\psi}(x_t, a_t | a_{1:t-1}, \gamma_t) P^{\psi}(a_{1:t-1}, \gamma_t) \tag{3.49a}$$

$$= \sum_{a_{1:t-1} \gamma_t} P(x_t, a_t | \pi_t, \gamma_t) P^{\psi}(a_{1:t-1}, \gamma_t) \tag{3.49b}$$

$$= \sum_{\pi_t, \gamma_t} P(x_t, a_t | \pi_t, \gamma_t) P^{\psi}(\pi_t, \gamma_t) \tag{3.49c}$$

$$= \sum_{\pi_t, \gamma_t} P(x_t, a_t | \pi_t, \gamma_t) P^{\theta}(\pi_t, \gamma_t) \tag{3.49d}$$

$$= P^{\theta}(x_t, a_t). \tag{3.49e}$$

where (3.49b) follows from (3.47), (3.49c) is change of variable and (3.49d) from (3.44).

3.9 Appendix C (Proof of Theorem 3.1)

Proof. We prove (3.15) using induction and from results in Lemma 3.4, 3.5 and 3.6 proved in Appendix D. For base case at $t = T$, $\forall i \in \mathcal{N}$, $(a_{1:T-1}, x^i_{1:T}) \in \mathcal{H}^i_T$, β^i

$$\mathbb{E}^{\beta^{*,i}_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i(X_T, A_T) \middle| a_{1:T-1}, x^i_{1:T} \right\} = V^i_T(\underline{\mu}^*_T[a_{1:T-1}], x^i_T) \tag{3.50a}$$

$$\geq \mathbb{E}^{\beta^i_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i(X_T, A_T) \middle| a_{1:T-1}, x^i_{1:T} \right\}. \tag{3.50b}$$

where (3.50a) follows from Lemma 3.6 and (3.50b) follows from Lemma 3.4 in Appendix D.

Let the induction hypothesis be that for $t+1$, $\forall i \in \mathcal{N}$, $a_{1:t} \in \mathcal{H}^c_{t+1}$, $x^i_{1:t+1} \in (\mathcal{X}^i)^{t+1}$, β^i,

$$\mathbb{E}^{\beta^{*,i}_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, x^i_{1:t+1} \right\} \tag{3.51a}$$

$$\geq \mathbb{E}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, x^i_{1:t+1} \right\}. \tag{3.51b}$$

Then $\forall i \in \mathcal{N}, (a_{1:t-1}, x^i_{1:t}) \in \mathcal{H}^i_t, \beta^i$, we have

$$\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, x^i_{1:t}\right\}$$

$$= V^i_t(\underline{\mu}^*_t[a_{1:t-1}], x^i_t) \tag{3.52a}$$

$$\geq \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\left\{R^i(X_t, A_t) + V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], X^i_{t+1})\big|a_{1:t-1}, x^i_{1:t}\right\} \tag{3.52b}$$

$$= \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\Big\{R^i(X_t, A_t)+ \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, A_t, x^i_{1:t+1}\right\}\Big|a_{1:t-1}, x^i_{1:t}\Big\} \tag{3.52c}$$

$$\geq \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\Big\{R^i(X_t, A_t)+ \mathbb{E}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T}\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, A_t, x^i_{1:t}, X^i_{t+1}\right\}\Big|a_{1:t-1}, x^i_{1:t}\Big\} \tag{3.52d}$$

$$= \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\Big\{R^i(X_t, A_t) + \mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, A_t, x^i_{1:t}, X^i_{t+1}\right\}\Big|a_{1:t-1}, x^i_{1:t}\Big\} \tag{3.52e}$$

$$= \mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}\,\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, x^i_{1:t}\right\}, \tag{3.52f}$$

where (3.52a) follows from Lemma 3.6, (3.52b) follows from Lemma 3.4, (3.52c) follows from Lemma 3.6, (3.52d) follows from induction hypothesis in (3.51b) and (3.52e) follows from Lemma 3.5. Moreover, construction of θ in (3.8), and consequently definition of β^* in (3.13) are pivotal for (3.52e) to follow from (3.52d).

We note that μ^* satisfies the consistency condition of [13, p. 331] from the fact that (a) for all t and for every common history $a_{1:t-1}$, all players use the same belief $\mu^*_t[a_{1:t-1}]$ on x_t and (b) the belief μ^*_t can be factorized as $\mu^*_t[a_{1:t-1}] = \prod_{i=1}^N \mu^{*,i}_t[a_{1:t-1}]\ \forall a_{1:t-1} \in \mathcal{H}^c_t$ where $\mu^{*,i}_t$ is updated through Bayes' rule ($\bar{F}$) as in Claim 3.5 in Appendix B.

3.10 Appendix D

Lemma 3.4. $\forall t \in \mathcal{T}, i \in \mathcal{N}, (a_{1:t-1}, x^i_{1:t}) \in \mathcal{H}^i_t, \beta^i_t$

$$V^i_t(\underline{\mu}^*_t[a_{1:t-1}], x^i_t) \geq$$
$$\mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[a_{1:t-1}], \beta^*_t(\cdot|a_{1:t-1}, \cdot), A_t), X^i_{t+1}) \middle| a_{1:t-1}, x^i_{1:t} \right\}. \tag{3.53}$$

Proof. We prove this Lemma by contradiction.

Suppose the claim is not true for t. This implies $\exists i, \hat{\beta}^i_t, \hat{a}_{1:t-1}, \hat{x}^i_{1:t}$ such that

$$\mathbb{E}^{\hat{\beta}^i_t \beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \middle| \hat{a}_{1:t-1}, \hat{x}^i_{1:t} \right\}$$
$$> V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{x}^i_t). \tag{3.54}$$

We will show that this leads to a contradiction.

Construct $\hat{\gamma}^i_t(a^i_t|x^i_t) = \begin{cases} \hat{\beta}^i_t(a^i_t|\hat{a}_{1:t-1}, \hat{x}^i_{1:t}) & x^i_t = \hat{x}^i_t \\ \text{arbitrary} & \text{otherwise.} \end{cases}$

Then for $\hat{a}_{1:t-1}, \hat{x}^i_{1:t}$, we have

$$V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{x}^i_t) \tag{3.55a}$$
$$= \max_{\gamma^i_t(\cdot|\hat{x}^i_t)} \mathbb{E}^{\gamma^i_t(\cdot|\hat{x}^i_t)\beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(\hat{x}^i_t x^{-i}_t, a_t) \right.$$
$$\left. + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \middle| \hat{x}^i_t \right\}, \tag{3.55b}$$
$$\geq \mathbb{E}^{\hat{\gamma}^i_t(\cdot|\hat{x}^i_t)\beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \middle| \hat{x}^i_t \right\}$$
$$= \sum_{x^{-i}_t, a_t, x_{t+1}} \left\{ R^i(\hat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), a_t), x^i_{t+1}) \right\} \times$$
$$\mu^{*,-i}_t[\hat{a}_{1:t-1}](x^{-i}_t) \hat{\gamma}^i_t(a^i_t|\hat{x}^i_t) \beta^{*,-i}_t(a^{-i}_t|\hat{a}_{1:t-1}, x^{-i}_t) Q^i_t(x^i_{t+1}|\hat{x}^i_t, a_t) \tag{3.55c}$$
$$= \sum_{x^{-i}_t, a_t, x_{t+1}} \left\{ R^i(\hat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), a_t), x^i_{t+1}) \right\} \times$$
$$\mu^{*,-i}_t[\hat{a}_{1:t-1}](x^{-i}_t) \hat{\beta}^i_t(a^i_t|\hat{a}_{1:t-1}, \hat{x}^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t|\hat{a}_{1:t-1}, x^{-i}_t) Q^i_t(x^i_{t+1}|\hat{x}^i_t, a_t) \tag{3.55d}$$

$$= \mathbb{E}^{\hat{\beta}^i_t \beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \big\{ R^i(\hat{x}^i_t x^{-i}_t, a_t) \\ + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \big| \hat{a}_{1:t-1}, \hat{x}^i_{1:t} \big\} \quad (3.55e)$$

$$> V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{x}^i_t) \quad (3.55f)$$

where (3.55b) follows from definition of V^i_t in (3.9), (3.55d) follows from definition of $\hat{\gamma}^i_t$ and (3.55f) follows from (3.54). However this leads to a contradiction.

Lemma 3.5. $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, x^i_{1:t+1}) \in \mathcal{H}^i_{t+1}$ and β^i_t

$$\mathbb{E}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t}, x^i_{1:t+1} \right\} \\ = \mathbb{E}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t}, x^i_{1:t+1} \right\}. \quad (3.56)$$

Thus the above quantities do not depend on β^i_t.

Proof. Essentially this claim stands on the fact that $\mu^{*,-i}_{t+1}[a_{1:t}]$ can be updated from $\mu^{*,-i}_t[a_{1:t-1}], \beta^{*,-i}_t$ and a_t, as $\mu^{*,-i}_{t+1}[a_{1:t}] = \prod_{j \neq i} \bar{F}(\mu^{*,-i}_t[a_{1:t-1}], \beta^{*,-i}_t, a_t)$ as in Claim 3.5. Since the above expectations involve random variables $X^{-i}_{t+1}, A_{t+1:T}, X_{t+2:T}$, we consider the probability
$P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t}, x^i_{1:t+1})$.

$$P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t}, x^i_{1:t+1}) \\ = \frac{\sum_{x^{-i}_t} P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_t, a_t, x_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t-1}, x^i_{1:t})}{\sum_{\tilde{x}^{-i}_t} P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(\tilde{x}^{-i}_t, a_t, x^i_{t+1} \big| a_{1:t-1}, x^i_{1:t})} \quad (3.57a)$$

We consider the numerator and the denominator separately. The numerator in (3.57a) is given by

$$Nr = \sum_{x^{-i}_t} P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_t \big| a_{1:t-1}, x^i_{1:t}) \beta^i_t(a^i_t | a_{1:t-1}, x^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t | a_{1:t-1}, x^{-i}_t) \\ Q(x_{t+1}|x_t, a_t) P^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, x^i_{1:t-1}, x_{t:t+1}) \quad (3.57b)$$

$$= \sum_{x^{-i}_t} \mu^{*,-i}_t[a_{1:t-1}](x^{-i}_t) \beta^i_t(a^i_t | a_{1:t-1}, x^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t | a_{1:t-1}, x^{-i}_t) Q^i(x^i_{t+1} | x^i_t, a_t) \\ Q^{-i}(x^{-i}_{t+1} | x^{-i}_t, a_t) P^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, x^i_{1:t}, x_{t+1}) \quad (3.57c)$$

where (3.57c) follows from the conditional independence of types given common informa-

tion, as shown in Claim 3.1, and the fact that probability on $(a_{t+1:T}, x_{2+t:T})$ given $a_{1:t}, x^i_{1:t-1}, x_{t:t+1}, \mu^*_t[a_{1:t-1}]$ depends on $a_{1:t}, x^i_{1:t}, x_{t+1}, \mu^*_{t+1}[a_{1:t}]$ through $\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T}$. Similarly, the denominator in (3.57a) is given by

$$
\begin{aligned}
&Dr = \\
&\sum_{\tilde{x}^{-i}_t} P^{\beta^i_{t:T}\beta^{*,-i}_{t:T},\mu^*_t}(\tilde{x}^{-i}_t|a_{1:t-1}, x^i_{1:t})\beta^i_t(a^i_t|a_{1:t-1}, x^i_{1:t})\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1}, \tilde{x}^{-i}_t)Q^i(x^i_{t+1}|x^i_t, a_t) && \text{(3.57d)}\\
=&\sum_{\tilde{x}^{-i}_t} \mu^{*,-i}_t[a_{1:t-1}](\tilde{x}^{-i}_t)\beta^i_t(a^i_t|a_{1:t-1}, x^i_{1:t})\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1}, \tilde{x}^{-i}_t)Q^i(x^i_{t+1}|x^i_t, a_t) && \text{(3.57e)}
\end{aligned}
$$

By canceling the terms $\beta^i_t(\cdot)$ and $Q^i(\cdot)$ in the numerator and the denominator, (3.57a) is given by

$$
\begin{aligned}
&\frac{\sum_{x^{-i}_t} \mu^{*,-i}_t[a_{1:t-1}](x^{-i}_t)\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1}, x^{-i}_t)Q^{-i}_{t+1}(x^{-i}_{t+1}|x^{-i}_t, a_t)}{\sum_{\tilde{x}^{-i}_t} \mu^{*,-i}_t[a_{1:t-1}](\tilde{x}^{-i}_t)\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1}, \tilde{x}^{-i}_t)}\times \\
&P^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(a_{t+1:T}, x_{t+2:T}|a_{1:t}, x^i_{1:t}, x_{t+1}) && \text{(3.57f)}\\
=&\mu^{*,-i}_{t+1}[a_{1:t}](x^{-i}_{t+1})P^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(a_{t+1:T}, x_{t+2:T}|a_{1:t}, x^i_{1:t}, x_{t+1}) && \text{(3.57g)}\\
=&P^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(x^{-i}_{t+1}, a_{t+1:T}, x_{t+2:T}|a_{1:t}, x^i_{1:t+1}), && \text{(3.57h)}
\end{aligned}
$$

where (3.57g) follows from using the definition of $\mu^{*,-i}_{t+1}[a_{1:t}](x^{-i}_t)$ in the forward recursive step in (3.14) and the definition of the belief update in (3.40).

Lemma 3.6. $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, x^i_{1:t}) \in \mathcal{H}^i_t$,

$$
V^i_t(\underline{\mu}^*_t[a_{1:t-1}], x^i_t) = \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\Big|a_{1:t-1}, x^i_{1:t}\right\}. \tag{3.58}
$$

Proof. We prove the Lemma by induction. For $t = T$,

$$
\begin{aligned}
&\mathbb{E}^{\beta^{*,i}_T\beta^{*,-i}_T,\mu^*_T[a_{1:T-1}]}\left\{R^i(X_T, A_T)\big|a_{1:T-1}, x^i_{1:T}\right\} \\
&= \sum_{x^{-i}_T a_T} R^i(x_T, a_T)\mu^*_T[a_{1:T-1}](x^{-i}_T)\beta^{*,i}_T(a^i_T|a_{1:T-1}, x^i_T)\beta^{*,-i}_T(a^{-i}_T|a_{1:T-1}, x^{-i}_T) && \text{(3.59a)}\\
&= V^i_T(\underline{\mu}^*_T[a_{1:T-1}], x^i_T), && \text{(3.59b)}
\end{aligned}
$$

where (3.59b) follows from the definition of V_t^i in (3.9) and the definition of β_T^* in the forward recursion in (3.13).

Suppose the claim is true for $t+1$, i.e., $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, x_{1:t+1}^i) \in \mathcal{H}_{t+1}^i$

$$V_{t+1}^i(\underline{\mu}_{t+1}^*[a_{1:t}], x_{t+1}^i) = \mathbb{E}^{\beta_{t+1:T}^{*,i}\beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t}]}\left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t}, x_{1:t+1}^i \right\}. \quad (3.60)$$

Then $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, x_{1:t}^i) \in \mathcal{H}_t^i$, we have

$$\begin{aligned}
&\mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}\left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, x_{1:t}^i \right\} \\
&= \mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}\Big\{ R^i(X_t, A_t) + \\
&\qquad \mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}\left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, x_{1:t}^i, X_{t+1}^i \right\} \Big| a_{1:t-1}, x_{1:t}^i \Big\} \qquad (3.61a) \\
&= \mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}\Big\{ R^i(X_t, A_t) + \\
&\qquad \mathbb{E}^{\beta_{t+1:T}^{*,i}\beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t-1}, A_t]}\left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, x_{1:t}^i, X_{t+1}^i \right\} \Big| a_{1:t-1}, x_{1:t}^i \Big\} \qquad (3.61b) \\
&= \mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}\left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{\mu}_{t+1}^*[a_{1:t-1}A_t], X_{t+1}^i) \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.61c) \\
&= \mathbb{E}^{\beta_t^{*,i}\beta_t^{*,-i}, \mu_t^*[a_{1:t-1}]}\left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{\mu}_{t+1}^*[a_{1:t-1}A_t], X_{t+1}^i) \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.61d) \\
&= V_t^i(\underline{\mu}_t^*[a_{1:t-1}], x_t^i), \qquad (3.61e)
\end{aligned}$$

where (3.61b) follows from Lemma 3.5 in Appendix D, (3.61c) follows from the induction hypothesis in (3.60), (3.61d) follows because the random variables involved in expectation, X_t^{-i}, A_t, X_{t+1}^i do not depend on $\beta_{t+1:T}^{*,i}\beta_{t+1:T}^{*,-i}$ and (3.61e) follows from the definition of β_t^* in the forward recursion in (3.13), the definition of μ_{t+1}^* in (3.14) and the definition of V_t^i in (3.9).

3.11 Appendix E (Proof of Lemma 3.3)

Proof. We prove this by contradiction. Suppose for any equilibrium generating function ϕ that generates (β^*, μ^*) through forward recursion, there exists $t \in \mathcal{T}, i \in \mathcal{N}, a_{1:t-1} \in \mathcal{H}_t^c$, such that for $\underline{\pi}_t = \underline{\mu}_t[a_{1:t-1}]$, (3.8) is not satisfied for ϕ i.e. for $\tilde{\gamma}_t^i = \phi^i[\underline{\pi}_t] =$

$\beta_t^{*,i}(\cdot|\underline{\mu}_t[a_{1:t-1}], x_t^i)$,

$$\tilde{\gamma}_t^i \notin \arg\max_{\gamma_t^i} \mathbb{E}^{\gamma_t^i(\cdot|x^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i) \big| x_t^i \right\}. \tag{3.62}$$

Let t be the first instance in the backward recursion when this happens. This implies $\exists\ \hat{\gamma}_t^i$ such that

$$\begin{aligned} &\mathbb{E}^{\hat{\gamma}_t^i(\cdot|x^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i) \big| x_t^i \right\} \\ > &\mathbb{E}^{\tilde{\gamma}_t^i(\cdot|x^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i) \big| x_t^i \right\} \end{aligned} \tag{3.63}$$

This implies for $\hat{\beta}_t(\cdot|\underline{\mu}_t[a_{1:t-1}], \cdot) = \hat{\gamma}_t^i$,

$$\begin{aligned} &\mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, x_{1:t}^i \right\} \\ &= \mathbb{E}^{\beta_t^{*,i}\beta_t^{*,-i}, \mu_t^*[a_{1:t-1}]} \Big\{ R^i(X_t, A_t) + \\ &\qquad \mathbb{E}^{\beta_{t:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, x_{1:t+1}^i \right\} \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.64) \\ &= \mathbb{E}^{\beta_t^{*,i}\beta_t^{*,-i}, \mu_t^*[a_{1:t-1}]} \Big\{ R^i(X_t, A_t) + \\ &\qquad \mathbb{E}^{\beta_{t+1:T}^{*,i}\beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t-1}, A_t]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, x_{1:t+1}^i \right\} \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.65) \\ &= \mathbb{E}^{\tilde{\gamma}_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i) \big| x_t^i \right\} \qquad (3.66) \\ &< \mathbb{E}^{\hat{\beta}_t^i(\cdot|\underline{\mu}_t[a_{1:t-1}], x_t^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X_{t+1}^i) \big| x_t^i \right\} \qquad (3.67) \\ &= \mathbb{E}^{\hat{\beta}_t^i\beta_t^{*,-i}, \mu_t^*[a_{1:t-1}]} \Big\{ R^i(X_t, A_t) \\ &\qquad + \mathbb{E}^{\beta_{t+1:T}^{*,i}\beta_{t+1:T}^{*,-i}\mu_{t+1}^*[a_{1:t-1}, A_t]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, x_{1:t}^i, X_{t+1}^i \right\} \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.68) \\ &= \mathbb{E}^{\hat{\beta}_t^i, \beta_{t+1:T}^{*,i}\beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, x_{1:t}^i \right\} \qquad (3.69) \end{aligned}$$

where (3.65) follows from Lemma 3.5, (3.66) follows from the definitions of $\tilde{\gamma}_t^i$ and $\mu_{t+1}[a_{1:t-1}, A_t]$ and Lemma 3.6, (3.67) follows from (3.63) and the definition of $\hat{\beta}_t^i$, (3.68) follows from Lemma 3.4, (3.69) follows from Lemma 3.5. However, this leads to a contradiction since (β^*, μ^*) is a PBE of the game.

CHAPTER 4

Signaling equilibria for dynamic LQG games with asymmetric information

4.1 Introduction

Linear quadratic Gaussian (LQG) team problems have been studied extensively under the framework of classical stochastic control with single controller and perfect recall [31, Ch.7]. In such a system, the state evolves linearly and the controller makes a noisy observation of the state which is also linear in the state and noise. The controller incurs a quadratic instantaneous cost. With all basic random variables being independent and Gaussian, the problem is modeled as a partially observed Markov decision process (POMDP). The belief state process under any control law happens to be Gaussian and thus can be sufficiently described by the corresponding mean and covariance processes, which can be updated by the Kalman filter equations. Moreover, the covariance can be computed offline and thus the mean (state estimate) is a sufficient statistic for control. Finally, due to the quadratic nature of the costs, the optimal control strategy is linear in the state. Thus, unlike most POMDP problems, the LQG stochastic control problem can be solved analytically and admits an easy-to-implement optimal strategy.

LQG team problems have also been studied under non-classical information structure such as in multi-agent decentralized team problems where 2 controllers with different information sets minimize the same objective. Such systems with asymmetric information structure are of special interest today because of the emergence of large scale networks such as social or power networks, where there are multiple decision makers with local or partial information about the system. It is well known that for decentralized LQG team problems, linear control policies are not optimal in general [59]. However there exist special information structures, such as partially nested [19] and stochastically nested [62], where linear control is shown to be optimal. Furthermore, due to their strong appeal for ease of imple-

mentation, linear strategies have been studied on their own for decentralized teams even at the possibility of being suboptimal (see [34] and references therein).

Authors in [3] studied a discrete-time dynamic LQG game with one step delayed sharing of observations. Authors in [41] studied a class of dynamic games with asymmetric information under the assumption that player's posterior beliefs about the system state conditioned on their common information are independent of the strategies used by the players in the past. Due to this independence of beliefs and past strategies, the authors were able to provide a backward recursive algorithm similar to dynamic programming to find Markov perfect equilibria [38] of a transformed game which are equivalently a class of Nash equilibria of the original game. The same authors specialized their results in [16] to find non-signaling equilibria of dynamic LQG games with asymmetric information.

We considered a general class of dynamic games with asymmetric information and independent private types in chapter 3 and provided a sequential decomposition methodology to find a class of PBE of the game considered. In our model, beliefs depend on the players' strategies, so it allows the possibility of signaling equilibria. In this chapter, we build on this methodology to find signaling equilibria for two-player dynamic LQG games with asymmetric information. We show that players' strategies that are linear in their private types in conjunction with consistent Gaussian beliefs form a PBE of the game. Our contributions are:

(a) Under strategies that are linear in players' private types, we show that the belief updates are Gaussian and the corresponding mean and covariance are updated through Kalman filtering equations which depend on the players' strategies, unlike the case in classical stochastic control and the model considered in [16]. Thus there is signaling [18, 30].

(b) We sequentially decompose the problem by specializing the forward-backward algorithm presented in chapter 3 for the dynamic LQG model. The backward algorithm requires, at each step, solving a fixed point equation in 'partial' strategies of the players for all possible beliefs. We show that in this setting, solving this fixed point equation reduces to solving a matrix algebraic equation for each realization of the state estimate covariance matrices.

(c) The cost-to-go value functions are shown to be quadratic in the private type and state estimates, which together with quadratic instantaneous costs and mean updates being linear in the control action, implies that at every time t player i faces an optimization problem which is quadratic in her control. Thus linear control strategies are shown to satisfy the optimality conditions in chapter 3.

(d) For the special case of scalar actions, we provide sufficient algorithmic conditions for existence of a solution of the algebraic matrix equation. Finally, we present numerical results on the steady state solution for specific parameters of the problem.

The chapter is structured as follows. In Section 4.2, we define the model. In Section 4.3, we summarize the general methodology in chapter 3. In Section 4.4, we present our main results where we construct equilibrium strategies and belief through a forward-backward recursion. In Section 4.5 we discuss existence issues and present numerical steady state solutions. We conclude in Section 4.6.

4.1.1 Notation

We use $\delta(\cdot)$ for the Dirac delta function. We use the notation $X \sim F$ to denote that the random variable X has distribution F. For any Euclidean set $\mathcal{S}$, $\mathcal{P}(\mathcal{S})$ represents the space of probability measures on $\mathcal{S}$ with respect to the Borel sigma algebra. We denote by P^g (or $\mathbb{E}^g$) the probability measure generated by (or expectation with respect to) strategy profile g. For any random vector X and event A, we use the notation $sm(\cdot|\cdot)$ to denote the conditional second moment, $sm(X|A) := \mathbb{E}[XX^\dagger|A]$. For any matrices $\mathbf{A}$ and $\mathbf{B}$, we will also use the notation $quad(\cdot;\cdot)$ to denote the quadratic function, $quad(\mathbf{A};\mathbf{B}) := \mathbf{B}^\dagger\mathbf{A}\mathbf{B}$. We denote trace of a matrix $\mathbf{A}$ by $tr(\mathbf{A})$. $N(\hat{x}, \mathbf{\Sigma})$ represents the vector Gaussian distribution with mean vector $\hat{x}$ and covariance matrix $\mathbf{\Sigma}$. All inequalities in matrices are to be interpreted in the sense of positive definitedness. All matrix inverses are interpreted as pseudo-inverses.

4.2 Model

We consider a discrete-time dynamical system with 2 strategic players over a finite time horizon $\mathcal{T} := \{1, 2, \dots T\}$ and with perfect recall. There is a dynamic state of the system $x_t := (x_t^1, x_t^2)$, where $x_t^i \in \mathcal{X}^i := \mathbb{R}^{n_i}$ is private type of player i at time t which is perfectly observed by her. Player i at time t takes action $u_t^i \in \mathcal{U}^i := \mathbb{R}^{m_i}$ after observing $u_{1:t-1}$, which is common information between the players, and $x_{1:t}^i$, which it observes privately. Thus at any time $t \in \mathcal{T}$, player i's information is $u_{1:t-1}, x_{1:t}^i$. Players' types evolve linearly as

$$x_{t+1}^i = \mathbf{A}_t^i x_t^i + \mathbf{B}_t^i u_t + w_t^i, \tag{4.1}$$

where $\mathbf{A}_t^i, \mathbf{B}_t^i$ are known matrices. $(X_1^1, X_1^2, (W_t^i)_{t\in\mathcal{T}})$ are basic random variables of the system which are assumed to be independent and Gaussian such that $X_1^i \sim N(0, \mathbf{\Sigma}_1^i)$ and

$W_t^i \sim N(0, \mathbf{Q}^i)$. As a consequence, types evolve as conditionally independent, controlled Markov processes,

$$P(x_{t+1}|u_{1:t}, x_{1:t}) = P(x_{t+1}|u_t, x_t) = \prod_{i=1}^{2} Q^i(x_{t+1}^i|u_t, x_t^i). \tag{4.2}$$

where $Q^i(x_{t+1}^i|u_t, x_t^i) := P(w_t^i = x_{t+1}^i - \mathbf{A}_t^i x_t^i - \mathbf{B}_t^i u_t)$. At the end of interval t, player i incurs an instantaneous cost $R^i(x_t, u_t)$,

$$\begin{aligned} R^i(x_t, u_t) &= u_t^\dagger \mathbf{T}^i u_t + x_t^\dagger \mathbf{P}^i x_t + 2u_t^\dagger \mathbf{S}^i x_t \\ &= \begin{bmatrix} u_t^\dagger & x_t^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{T}^i & \mathbf{S}^i \\ \mathbf{S}^{i\dagger} & \mathbf{P}^i \end{bmatrix} \begin{bmatrix} u_t \\ x_t \end{bmatrix}, \end{aligned} \tag{4.3}$$

where $\mathbf{T}^i, \mathbf{P}^i, \mathbf{S}^i$ are real matrices of appropriate dimensions and $\mathbf{T}^i, \mathbf{P}^i$ are symmetric. We define the instantaneous cost matrix $\mathbf{R}^i$ as $\mathbf{R}^i := \begin{bmatrix} \mathbf{T}^i & \mathbf{S}^i \\ \mathbf{S}^{i\dagger} & \mathbf{P}^i \end{bmatrix}$. Let $g^i = (g_t^i)_{t\in\mathcal{T}}$ be a probabilistic strategy of player i, where $g_t^i : (\mathcal{U}^i)^{t-1} \times (\mathcal{X}^i)^t \to \mathcal{P}(\mathcal{U}^i)$ such that player i plays action u_t^i according to distribution $g_t^i(\cdot|u_{1:t-1}, x_{1:t}^i)$. Let $g := (g^i)_{i=1,2}$ be a strategy profile of both players. The distribution of the basic random variables and their independence structure together with the system evolution in (4.1) and players strategy profile g define a joint distribution on all random variables involved in the dynamical process. The objective of player i is to maximize her total expected cost

$$J^{i,g} := \mathbb{E}^g \left\{ \sum_{t=1}^{T} R^i(X_t, U_t) \right\}. \tag{4.4}$$

With both players being strategic, this problem is modeled as a dynamic LQG game with asymmetric information and with simultaneous moves.

4.3 Structured perfect Bayesian equilibria

In Chapter 3, we considered a general class of dynamic games with asymmetric information, where players' types evolve as conditionally independent controlled Markov processes. We introduced the notion of equilibria for such games in Section 3.4.1 and subsequently, a backward-forward algorithm was provided to find a class of PBE of the game called structured perfect Bayesian equilibria (SPBE). In these equilibria, player i's strategy is of the form $U_t^i \sim m_t^i(\cdot|\pi_t^1, \pi_t^2, x_t^i)$ where $m_t^i : \mathcal{P}(\mathcal{X}^1) \times \mathcal{P}(\mathcal{X}^2) \times \mathcal{X}^i \to \mathcal{P}(\mathcal{U}^i)$.

Specifically, player i's action at time t depends on her private history $x^i_{1:t}$ only through x^i_t. Furthermore, it depends on the common information $u_{1:t-1}$ through a common belief vector $\underline{\pi}_t := (\pi^1_t, \pi^2_t)$ where $\pi^i_t \in \mathcal{P}(\mathcal{X}^i)$ is belief on player i's current type x^i_t conditioned on common information $u_{1:t-1}$, i.e. $\pi^i_t(x^i_t) := P^g(X^i_t = x^i_t | u_{1:t-1})$.

The common information $u_{1:t-1}$ was summarized into the belief vector (π^1_t, π^2_t) following the common agent approach used for dynamic decentralized team problems [43]. Using this approach, player i's strategy can be equivalently described as follows: player i at time t observes $u_{1:t-1}$ and takes action γ^i_t, where $\gamma^i_t : \mathcal{X}^i \to \mathcal{P}(\mathcal{U}^i)$ is a partial (stochastic) function from her private information x^i_t to u^i_t of the form $U^i_t \sim \gamma^i_t(\cdot|x^i_t)$. These actions are generated through some policy $\psi^i = (\psi^i_t)_{t\in\mathcal{T}}$, $\psi^i_t : (\mathcal{U}^i)^{t-1} \to \{\mathcal{X}^i \to \mathcal{P}(\mathcal{U}^i)\}$, that operates on the common information $u_{1:t-1}$ so that $\gamma^i_t = \psi^i_t[u_{1:t-1}]$. Then any policy of the player i of the form $U^i_t \sim g^i_t(\cdot|u_{1:t-1}, x^i_t)$ is equivalent to $U^i_t \sim \psi^i_t[u_{1:t-1}](\cdot|x^i_t)$ [43].

The common belief π^i_t is shown in Claim 3.5 of chapter 3 to be updated as

$$\pi^i_{t+1}(x^i_{t+1}) = \frac{\int_{x^i_t} \pi^i_t(x^i_t)\gamma^i_t(u^i_t|x^i_t)Q^i_t(x^i_{t+1}|x^i_t, u_t)dx^i_t}{\int_{\tilde{x}^i_t} \pi^i_t(\tilde{x}^i_t)\gamma^i_t(u^i_t|\tilde{x}^i_t)d\tilde{x}^i_t}, \tag{4.5a}$$

if the denominator is not 0, and as

$$\pi^i_{t+1}(x^i_{t+1}) = \int_{x^i_t} \pi^i_t(x^i_t)Q^i_t(x^i_{t+1}|x^i_t, u_t)dx^i_t, \tag{4.5b}$$

if the denominator is 0. The belief update can be summarized as,

$$\pi^i_{t+1} = \bar{F}(\pi^i_t, \gamma^i_t, u_t), \tag{4.6}$$

where $\bar{F}$ is independent of players' strategy profile g. The SPBE of the game can be found through a two-step backward-forward algorithm. In the backward recursive part, an equilibrium generating function θ is defined based on which a strategy and belief profile (β^*, μ^*) are defined through a forward recursion.

4.4 SPBE of the dynamic LQG game

In this section, we apply the general methodology for finding SPBE described in chapter 3, on the specific dynamic LQG game model described in Section 4.2. We show that players' strategies that are linear in their private types in conjunction with Gaussian beliefs, form an SPBE of the game. We prove this result by constructing an equilibrium generating function θ using backward recursion such that for all Gaussian belief vectors $\underline{\pi}_t$, $\tilde{\gamma}_t = \theta_t[\underline{\pi}_t]$, $\tilde{\gamma}^i_t$ is

of the form $\tilde{\gamma}_t^i(u_t^i|x_t^i) = \delta(u_t^i - \tilde{\mathbf{L}}_t^i x^i - \tilde{m}_t^i)$ and satisfies (3.8). Based on θ, we construct an equilibrium belief and strategy profile.

The following lemma shows that common beliefs remain Gaussian under linear deterministic γ_t of the form $\gamma_t^i(u_t^i|x_t^i) = \delta(u_t^i - \mathbf{L}_t^i x_t^i - m_t^i)$.

Lemma 4.1. If π_t^i is a Gaussian distribution with mean $\hat{x}_t^i$ and covariance $\mathbf{\Sigma}_t^i$, and $\gamma_t^i(u_t^i|x_t^i) = \delta(u_t^i - \mathbf{L}_t^i x_t^i - m_t^i)$ then π_{t+1}^i, given by (4.5), is also Gaussian distribution with mean $\hat{x}_{t+1}^i$ and covariance $\mathbf{\Sigma}_{t+1}^i$, where

$$\hat{x}_{t+1}^i = \mathbf{A}_t^i \hat{x}_t^i + \mathbf{B}_t^i u_t + \mathbf{A}_t^i \mathbf{G}_t^i (u_t^i - \mathbf{L}_t^i \hat{x}_t^i - m_t^i) \tag{4.7a}$$

$$\mathbf{\Sigma}_{t+1}^i = \mathbf{A}_t^i (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i)^\dagger \mathbf{\Sigma}_t^i (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i) \mathbf{A}_t^{i\dagger} + \mathbf{Q}^i. \tag{4.7b}$$

where

$$\mathbf{G}_t^i = \mathbf{\Sigma}_t^i \mathbf{L}_t^{i\dagger} (\mathbf{L}_t^i \mathbf{\Sigma}_t^i \mathbf{L}_t^{i\dagger})^{-1}. \tag{4.8}$$

Proof. See Appendix A.

Based on previous lemma, we define ϕ_x^i, ϕ_s^i as update functions of mean and covariance matrix, respectively, as defined in (4.7), such that

$$\hat{x}_{t+1}^i = \phi_x^i(\hat{x}_t^i, \mathbf{\Sigma}_t^i, \mathbf{L}_t^i, m_t^i, u_t) \tag{4.9a}$$

$$\mathbf{\Sigma}_{t+1}^i = \phi_s^i(\mathbf{\Sigma}_t^i, \mathbf{L}_t^i). \tag{4.9b}$$

We also say,

$$\hat{x}_{t+1} = \phi_x(\hat{x}_t, \mathbf{\Sigma}_t, \mathbf{L}_t, m_t, u_t) \tag{4.10}$$

$$\mathbf{\Sigma}_{t+1} = \phi_s(\mathbf{\Sigma}_t, \mathbf{L}_t). \tag{4.11}$$

The previous lemma shows that with linear deterministic γ_t^i, the next update of the mean of the common belief, $\hat{x}_{t+1}^i$ is linear in $\hat{x}_t^i$ and the control action u_t^i. Furthermore, these updates are given by appropriate Kalman filter equations. It should be noted however that the covariance update in (4.7b) depends on the strategy through γ_t^i and specifically through the matrix $\mathbf{L}_t^i$. This specifically shows how belief updates depend on strategies on the players which leads to signaling, unlike the case in classical stochastic control and the model considered in [16].

Now we will construct an equilibrium generating function θ using the backward recursion in (3.7)–(3.9). The θ function generates linear deterministic partial functions γ_t,

which, from Lemma 4.1 and the fact that initial beliefs (or priors) are Gaussian, generates only Gaussian belief vectors $(\pi_t^1, \pi_t^2)_{t\in\mathcal{T}}$ for the whole time horizon. These beliefs can be sufficiently described by their mean and covariance processes $(\hat{x}_t^1, \mathbf{\Sigma}_t^1)_{t\in\mathcal{T}}$ and $(\hat{x}_t^2, \mathbf{\Sigma}_t^2)_{t\in\mathcal{T}}$ which are updated using (4.7).

For $t = T+1, T, \ldots, 1$, we define the vectors

$$e_t^i := \begin{bmatrix} x_t^i \\ \hat{x}_t^1 \\ \hat{x}_t^2 \end{bmatrix} \quad z_t^i := \begin{bmatrix} u_t^i \\ x_t^i \\ \hat{x}_t^1 \\ \hat{x}_t^2 \end{bmatrix} \quad y_t^i := \begin{bmatrix} u_t^1 \\ u_t^2 \\ x_t^1 \\ x_t^2 \\ x_{t+1}^i \\ \hat{x}_{t+1}^1 \\ \hat{x}_{t+1}^2 \end{bmatrix}. \tag{4.12}$$

Theorem 4.1. The backward recursion (3.7)–(3.9) admits[1] a solution of the form $\theta_t[\pi_t] = \theta_t[\hat{x}_t, \mathbf{\Sigma}_t] = \tilde{\gamma}_t$ where $\tilde{\gamma}_t^i(u_t^i|x_t^i) = \delta(u_t^i - \tilde{\mathbf{L}}_t^i x_t^i - \tilde{m}_t^i)$ and $\tilde{\mathbf{L}}_t^i, \tilde{m}_t^i$ are appropriately defined matrices and vectors, respectively. Furthermore, the value function reduces to

$$V_t^i(\underline{\pi}_t, x_t^i) = V_t^i(\hat{x}_t, \mathbf{\Sigma}_t, x_t^i) \tag{4.13a}$$
$$= quad(\mathbf{V}_t^i(\mathbf{\Sigma}_t); e_t^i) + \rho_t^i(\mathbf{\Sigma}_t). \tag{4.13b}$$

with $\mathbf{V}_t^i(\mathbf{\Sigma}_t)$ and $\rho_t^i(\mathbf{\Sigma}_t)$ as appropriately defined matrix and scalar quantities, respectively.

Proof. We construct such a θ function through the backward recursive construction and prove the properties of the corresponding value functions inductively.

(a) For $i = 1, 2, \forall\ \mathbf{\Sigma}_{T+1}$, let $\mathbf{V}_{T+1}^i(\mathbf{\Sigma}_{T+1}) := \mathbf{0}, \rho_{T+1}^i(\mathbf{\Sigma}_{T+1}) := 0$. Then $\forall\ \hat{x}_{T+1}^1, \hat{x}_{T+1}^2$, $\mathbf{\Sigma}_{T+1}^1, \mathbf{\Sigma}_{T+1}^2, x_{T+1}^i$ and for $\underline{\pi}_t = (\pi_t^1, \pi_t^2)$, where π_t^i is $N(\hat{x}_t^i, \mathbf{\Sigma}_t^i)$,

$$V_{T+1}^i(\underline{\pi}_{T+1}, x_{T+1}^i) := 0 \tag{4.14a}$$
$$= V_{T+1}^i(\hat{x}_{T+1}, \mathbf{\Sigma}_{T+1}, x_{T+1}^i) \tag{4.14b}$$
$$= quad(\mathbf{V}_{T+1}^i(\mathbf{\Sigma}_{T+1}), e_{T+1}^i) + \rho_{T+1}^i(\mathbf{\Sigma}_{T+1}). \tag{4.14c}$$

(b) For all $t \in \{T, T-1, \ldots, 1\}, i = 1, 2$,

Suppose $V_{t+1}^i(\underline{\pi}_{t+1}, x_{t+1}^i) = quad(\mathbf{V}_{t+1}^i(\mathbf{\Sigma}_{t+1}), e_{t+1}^i) + \rho_{t+1}^i(\mathbf{\Sigma}_{t+1})$ (from induction

[1] Under certain conditions, stated in the proof.

hypothesis) where $\mathbf{V}^i_{t+1}$ is a symmetric matrix defined recursively. Define $\bar{\mathbf{V}}^i_t$ as

$$\bar{\mathbf{V}}^i_t(\mathbf{\Sigma}_t, \mathbf{L}_t) := \begin{bmatrix} \mathbf{T}^i & \mathbf{S}^i & \mathbf{0} \\ \mathbf{S}^{i\dagger} & \mathbf{P}^i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}^i_{t+1}(\phi_s(\mathbf{\Sigma}_t, \mathbf{L}_t)) \end{bmatrix}. \tag{4.15}$$

Since $\mathbf{T}^i, \mathbf{P}^i$ are symmetric by assumption, $\bar{\mathbf{V}}^i_t$ is also symmetric.

For ease of exposition, we will assume $i = 1$ and for player 2, a similar argument holds. At time t, the quantity that is minimized for player $i = 1$ in (3.8) can be written as

$$\mathbb{E}^{\gamma^1_t(\cdot|x^1_t)} \left[\mathbb{E}^{\tilde{\gamma}^2_t} \left[R^1(X_t, U_t) + V^1_{t+1}(F(\underline{\pi}_t, \tilde{\gamma}_t, U_t), X^1_{t+1}) \big| \underline{\pi}_t, x^1_t, u^1_t \right] \big| \underline{\pi}_t, x^1_t \right]. \tag{4.16}$$

The inner expectation can be written as follows, where $\tilde{\gamma}^2_t(u^2_t|x^2_t) = \delta(u^2_t - \tilde{\mathbf{L}}^2_t x^2_t - \tilde{m}^2_t)$,

$$\begin{aligned} &\mathbb{E}^{\tilde{\gamma}^2_t} \Bigg[quad \left(\begin{bmatrix} \mathbf{T}^1 & \mathbf{S}^1 \\ \mathbf{S}^{1\dagger} & \mathbf{P}^1 \end{bmatrix}; z^i_t \right) \\ &\qquad + quad \left(\mathbf{V}^1_{t+1}(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)); e^i_{t+1} \right) + \rho^1_{t+1}(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)) \big| \pi_t, x^1_t, u^i_t \Bigg] \end{aligned} \tag{4.17a}$$

$$= \mathbb{E}^{\tilde{\gamma}^2_t} \left[quad \left(\bar{\mathbf{V}}^1_t(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); y^1_t \right) + \rho^1_{t+1}(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)) \big| \pi_t, x^1_t, u^1_t \right] \tag{4.17b}$$

$$= quad \left(\bar{\mathbf{V}}^1_t(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \mathbf{D}^1_t z^1_t + \mathbf{C}^1_t \begin{bmatrix} m^1_t \\ \tilde{m}^2_t \end{bmatrix} \right) + \rho^1_t(\mathbf{\Sigma}_t), \tag{4.17c}$$

where $\bar{\mathbf{V}}^i_t$ is defined in (4.15) and function ϕ_s is defined in (4.11); y^i_t, z^i_t are defined in (4.12); ρ^i_t is given by

$$\begin{aligned} \rho^i_t(\mathbf{\Sigma}_t) = tr &\left(\mathbf{\Sigma}_t^{-i} quad \left(\bar{\mathbf{V}}^i_t(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \mathbf{J}^i_t \right) \right) \\ &+ tr(\mathbf{Q}^i V^i_{11,t+1}(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t))) + \rho^i_{t+1}(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)), \end{aligned} \tag{4.18}$$

where $V^i_{11,t+1}$ is the matrix corresponding to x^i_{t+1} in V^i_{t+1} i.e. in the first row and first

column of the matrix V_{t+1}^i; and matrices $\mathbf{D}_t^i, \mathbf{C}_t^i, \mathbf{J}_t^i$ are as follows,

$$\mathbf{D}_t^1 := \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{L}}_t^2 \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \\ \mathbf{B}_{1,t}^1 & \mathbf{A}_t^1 & \mathbf{0} & \mathbf{B}_{2,t}^1 \tilde{\mathbf{L}}_t^2 \\ \mathbf{A}_t^1 \mathbf{G}_t^1 + \mathbf{B}_{1,t}^1 & \mathbf{0} & \mathbf{A}_t^1 (\mathbf{I} - \mathbf{G}_t^1 \mathbf{L}_t^1) & \mathbf{B}_{2,t}^1 \tilde{\mathbf{L}}_t^2 \\ \mathbf{B}_{1,t}^2 & \mathbf{0} & \mathbf{0} & \mathbf{A}_t^2 + \mathbf{B}_{2,t}^2 \tilde{\mathbf{L}}_t^2 \end{bmatrix} \tag{4.19a}$$

$$\mathbf{D}_t^2 := \begin{bmatrix} \mathbf{0} & \mathbf{0} & \tilde{\mathbf{L}}_t^1 & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{B}_{2,t}^2 & \mathbf{A}_t^2 & \mathbf{B}_{1,t}^2 \tilde{\mathbf{L}}_t^1 & \mathbf{0} \\ \mathbf{B}_{2,t}^1 & \mathbf{0} & \mathbf{A}_t^1 + \mathbf{B}_{1,t}^1 \tilde{\mathbf{L}}_t^1 & \mathbf{0} \\ \mathbf{A}_t^2 \mathbf{G}_t^2 + \mathbf{B}_{2,t}^2 & \mathbf{0} & \mathbf{B}_{1,t}^2 \tilde{\mathbf{L}}_t^1 & \mathbf{A}_t^2 (\mathbf{I} - \mathbf{G}_t^2 \mathbf{L}_t^2) \end{bmatrix} \tag{4.19b}$$

$$\mathbf{C}_t^1 := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{2,t}^1 \\ -\mathbf{A}_t^1 \mathbf{G}_t^1 & \mathbf{B}_{2,t}^1 \\ \mathbf{0} & \mathbf{B}_{2,t}^2 \end{bmatrix} \quad \mathbf{C}_t^2 := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{B}_{1,t}^2 & \mathbf{0} \\ \mathbf{B}_{1,t}^1 & \mathbf{0} \\ \mathbf{B}_{1,t}^2 & -\mathbf{A}_t^2 \mathbf{G}_t^2 \end{bmatrix} \tag{4.20}$$

$$\begin{aligned} \mathbf{J}_t^{1\dagger} &:= \begin{bmatrix} \mathbf{0} & \mathbf{L}_t^2 & \mathbf{0} & \mathbf{I} & \mathbf{B}_{2,t}^1 \mathbf{L}_t^2 & \mathbf{B}_{2,t}^1 \mathbf{L}_t^2 & (\mathbf{B}_{2,t}^2 + \mathbf{A}_t^2 \mathbf{G}_t^2) \mathbf{L}_t^2 \end{bmatrix}^\dagger \\ \mathbf{J}_t^{2\dagger} &:= \begin{bmatrix} \mathbf{L}_t^1 & \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{B}_{1,t}^2 \mathbf{L}_t^1 & (\mathbf{B}_{1,t}^1 + \mathbf{A}_t^1 \mathbf{G}_t^1) \mathbf{L}_t^1 & \mathbf{B}_{1,t}^2 \mathbf{L}_t^1 \end{bmatrix}^\dagger \end{aligned} \tag{4.21}$$

where $\mathbf{B}_t^i =: \begin{bmatrix} \mathbf{B}_{1,t}^i & \mathbf{B}_{2,t}^i \end{bmatrix}$, $\mathbf{B}_{1,t}^i, \mathbf{B}_{2,t}^i$ are the parts of the matrix $\mathbf{B}_t^i$ that corresponds to u_t^1, u_t^2 respectively. Let $\mathbf{D}_t^1 =: \begin{bmatrix} \mathbf{D}_t^{u1} & \mathbf{D}_t^{e1} \end{bmatrix}$ where $\mathbf{D}_t^{u1}$ is the first column matrix of $\mathbf{D}_t^1$ corresponding to u_t^1 and $\mathbf{D}_t^{e1}$ is the matrix composed of remaining three column matrices of $\mathbf{D}_t^1$ corresponding to e_t^1. The expression in (4.17c) is averaged with

respect to u_t^1 using the measure $\gamma_t^1(\cdot|x_t^1)$ and minimized in (3.8) over $\gamma_t^1(\cdot|x_t^1)$. This minimization can be performed component wise leading to a deterministic policy $\tilde{\gamma}_t^1(u_t^1|x_t^1) = \delta(u_t^1 - \tilde{\mathbf{L}}_t^1 x_t^1 - \tilde{m}_t^1) = \delta(u_t^1 - u_t^{1*})$, assuming that the matrix $\tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1\tilde{\mathbf{D}}_t^{u1}$ is positive definite[2]. In that case, the unique minimizer $u_t^{1*} = \tilde{\mathbf{L}}_t^1 x_t^1 + \tilde{m}_t^1$ can be found by differentiating (4.17c) w.r.t. $u_t^{1\dagger}$ and equating it to $\mathbf{0}$, resulting in the equation,

$$\mathbf{0} = 2\begin{bmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}\tilde{\mathbf{D}}_t^{1\dagger}\bar{\mathbf{V}}_t^1(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\left(\tilde{\mathbf{D}}_t^1 z_t^1 + \tilde{\mathbf{C}}_t^1\tilde{m}_t\right) \tag{4.22a}$$

$$\mathbf{0} = \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\left(\tilde{\mathbf{D}}_t^{u1}u_t^{1*} + \tilde{\mathbf{D}}_t^{e1}e_t^1 + \tilde{\mathbf{C}}_t^1\tilde{m}_t\right) \tag{4.22b}$$

$$\mathbf{0} = \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\left(\tilde{\mathbf{D}}_t^{u1}(\tilde{\mathbf{L}}_t^1 x_t^1 + \tilde{m}_t^1) + [\tilde{\mathbf{D}}_t^{e1}]_1 x_t^1 + [\tilde{\mathbf{D}}_t^{e1}]_{23}\hat{x}_t + \tilde{\mathbf{C}}_t^1\tilde{m}_t\right), \tag{4.22c}$$

where $[\mathbf{D}^{ei}]_1$ is the first matrix column of $\mathbf{D}^{ei}$, $[\mathbf{D}^{ei}]_{23}$ is the matrix composed of the second and third column matrices of $\mathbf{D}^{ei}$. Matrices $\tilde{\mathbf{D}}_t^i, \tilde{\mathbf{C}}_t^i$ are obtained by substituting $\tilde{\mathbf{L}}_t^i, \tilde{\mathbf{G}}_t^i$ in place of $\mathbf{L}_t^i, \mathbf{G}_t^i$ in the definition of $\tilde{\mathbf{D}}_t^i, \tilde{\mathbf{C}}_t^i$ in (4.20), respectively, and $\tilde{\mathbf{G}}_t^i$ is the matrix obtained by substituting $\tilde{\mathbf{L}}_t^i$ in place of $\mathbf{L}_t^i$ in (4.8).

Thus (4.22c) is equivalent to (3.8) and with a similar analysis for player 2, it implies that $\tilde{\mathbf{L}}_t^i$ is solution of the following algebraic fixed point equation,

$$\left(\tilde{\mathbf{D}}_t^{ui\dagger}\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\tilde{\mathbf{D}}_t^{ui}\right)\tilde{\mathbf{L}}_t^i = -\tilde{\mathbf{D}}_t^{ui\dagger}\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)[\tilde{\mathbf{D}}_t^{ei}]_1. \tag{4.23a}$$

For player 1, it reduces to,

$$\begin{aligned}&\left[\mathbf{T}_{11}^1 + \begin{bmatrix}\mathbf{B}_{1,t}^1 \\ \mathbf{A}_t^1\mathbf{G}_t^1 + \mathbf{B}_{1,t}^1 \\ \mathbf{B}_{1,t}^2\end{bmatrix}^\dagger \mathbf{V}_{t+1}^1(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t))\begin{bmatrix}\mathbf{B}_{1,t}^1 \\ \mathbf{A}_t^1\mathbf{G}_t^1 + \mathbf{B}_{1,t}^1 \\ \mathbf{B}_{1,t}^2\end{bmatrix}\right]\tilde{\mathbf{L}}_t^1 \\ &\quad = -\left[\mathbf{S}_{11}^{1\dagger} + \begin{bmatrix}\mathbf{B}_{1,t}^1 \\ \mathbf{A}_t^1\mathbf{G}_t^1 + \mathbf{B}_{1,t}^1 \\ \mathbf{B}_{1,t}^2\end{bmatrix}^\dagger \mathbf{V}_{t+1}^1(\phi_s(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t))\begin{bmatrix}\mathbf{A}_t^1 \\ 0 \\ 0\end{bmatrix}\right],\end{aligned} \tag{4.23b}$$

and a similar expression holds for player 2.

[2]This condition is true if the instantaneous cost matrix $\mathbf{R}^i = \begin{bmatrix}\mathbf{T}^i & \mathbf{S}^i \\ \mathbf{S}^{i\dagger} & \mathbf{P}^i\end{bmatrix}$ is positive definite and can be proved inductively in the proof by showing that $\mathbf{V}_t^i$ and $\bar{\mathbf{V}}_t^i$ are positive definite.

In addition, $\tilde{m}_t$ can be found from (4.22c) as

$$\begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1\tilde{\mathbf{D}}_t^{u1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2\tilde{\mathbf{D}}_t^{u2} \end{bmatrix} \tilde{m}_t = -\begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1[\tilde{\mathbf{D}}_t^{e1}]_{23} \\ \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2[\tilde{\mathbf{D}}_t^{e2}]_{23} \end{bmatrix} \hat{x}_t - \begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1\tilde{\mathbf{C}}_t^1 \\ \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2\tilde{\mathbf{C}}_t^2 \end{bmatrix} \tilde{m}_t \tag{4.24a}$$

$$\tilde{m}_t = -\left[\begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1\tilde{\mathbf{D}}_t^{u1} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2\tilde{\mathbf{D}}_t^{u2} \end{bmatrix} + \begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1\tilde{\mathbf{C}}_t^1 \\ \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2\tilde{\mathbf{C}}_t^2 \end{bmatrix}\right]^{-1} \begin{bmatrix} \tilde{\mathbf{D}}_t^{u1\dagger}\bar{\mathbf{V}}_t^1[\tilde{\mathbf{D}}_t^{e1}]_{23} \\ \tilde{\mathbf{D}}_t^{u2\dagger}\bar{\mathbf{V}}_t^2[\tilde{\mathbf{D}}_t^{e2}]_{23} \end{bmatrix} \hat{x}_t \tag{4.24b}$$

$$=: \tilde{\mathbf{M}}_t\hat{x}_t =: \begin{bmatrix} \tilde{\mathbf{M}}_t^1 \\ \tilde{\mathbf{M}}_t^2 \end{bmatrix} \hat{x}_t, \tag{4.24c}$$

Finally, the resulting cost for player i is,

$$V_t^i(\underline{\pi}_t, x_t^i) = V_t^i(\hat{x}_t, \mathbf{\Sigma}_t, x_t^i) \tag{4.25a}$$

$$:= quad\left(\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \begin{bmatrix} \tilde{\mathbf{D}}_t^{ui} & \tilde{\mathbf{D}}_t^{ei} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{L}}_t^i x_t^i + \tilde{\mathbf{M}}_t^i\hat{x}_t \\ e_t^i \end{bmatrix} + \tilde{\mathbf{C}}_t^i\tilde{\mathbf{M}}_t\hat{x}_t\right) + \rho_t^i(\mathbf{\Sigma}_t) \tag{4.25b}$$

$$= quad\left(\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \tilde{\mathbf{D}}_t^{ui}(\tilde{\mathbf{L}}_t^i x_t^i + \tilde{\mathbf{M}}_t^i\hat{x}_t) + \tilde{\mathbf{D}}_t^{e1}e_t^i + \tilde{\mathbf{C}}_t^i\tilde{\mathbf{M}}_t\hat{x}_t\right) + \rho_t^i(\mathbf{\Sigma}_t) \tag{4.25c}$$

$$= quad\left(\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \left(\begin{bmatrix} \tilde{\mathbf{D}}_t^{ui}\tilde{\mathbf{L}}_t^i & \tilde{\mathbf{D}}_t^{ui}\tilde{\mathbf{M}}_t^i + \tilde{\mathbf{C}}_t^i\tilde{\mathbf{M}}_t \end{bmatrix} + \tilde{\mathbf{D}}_t^{ei}\right) e_t^i\right) + \rho_t^i(\mathbf{\Sigma}_t) \tag{4.25d}$$

$$= quad\left(\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t); \tilde{\mathbf{F}}_t^i e_t^i\right) + \rho_t^i(\mathbf{\Sigma}_t) \tag{4.25e}$$

$$= quad\left(\tilde{\mathbf{F}}_t^{i\dagger}\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\tilde{\mathbf{F}}_t^i; e_t^i\right) + \rho_t^i(\mathbf{\Sigma}_t) \tag{4.25f}$$

$$= quad\left(\mathbf{V}_t^i(\mathbf{\Sigma}_t); e_t^i\right) + \rho_t^i(\mathbf{\Sigma}_t), \tag{4.25g}$$

where,

$$\tilde{\mathbf{F}}_t^i := \begin{bmatrix} \tilde{\mathbf{D}}_t^{ui}\tilde{\mathbf{L}}_t^i & \tilde{\mathbf{D}}_t^{ui}\tilde{\mathbf{M}}_t^i + \tilde{\mathbf{C}}_t^i\tilde{\mathbf{M}}_t \end{bmatrix} + \tilde{\mathbf{D}}_t^{ei} \tag{4.26a}$$

$$\mathbf{V}_t^i(\mathbf{\Sigma}_t) := \tilde{\mathbf{F}}_t^{i\dagger}\bar{\mathbf{V}}_t^i(\mathbf{\Sigma}_t, \tilde{\mathbf{L}}_t)\tilde{\mathbf{F}}_t^i. \tag{4.26b}$$

Since $\bar{\mathbf{V}}_t^i$ is symmetric, so is $\mathbf{V}_t^i$. Thus the induction step is completed.

Taking motivation from the previous theorem and with slight abuse of notation, we

define

$$\tilde{\gamma}_t = \theta_t[\pi_t] = \theta_t[\hat{x}_t, \mathbf{\Sigma}_t] \tag{4.27}$$

and since $\tilde{\gamma}_t^i(u_t^i|x_t^i) = \delta(u_t^i - \tilde{\mathbf{L}}_t^i x_t^i - \tilde{m}_t^i)$, we define a reduced mapping (θ^L, θ^m) as

$$\theta_t^{Li}[\hat{x}_t, \mathbf{\Sigma}_t] = \theta_t^{Li}[\mathbf{\Sigma}_t] := \tilde{\mathbf{L}}_t^i \quad \text{and} \quad \theta_t^{mi}[\hat{x}_t, \mathbf{\Sigma}_t] := \tilde{m}_t^i, \tag{4.28}$$

where $\tilde{\mathbf{L}}_t^i$ does not depend on $\hat{x}_t$ and $\tilde{m}_t^i$ is linear in $\hat{x}_t$ and is of the form $\tilde{m}_t^i = \tilde{\mathbf{M}}_t^i \hat{x}_t$.

Now we construct the equilibrium strategy and belief profile (β^*, μ^*) through the forward recursion in (3.12)–(3.14), using the equilibrium generating function $\theta \equiv (\theta^L, \theta^m)$.

(a) Let $\mu_1^{*,i}[\phi](x_1^i) = N(0, \mathbf{\Sigma}_1^i)$.

For $t = 1, 2 \dots T-1, \forall u_{1:t} \in \mathcal{H}_{t+1}^c$, if $\mu_t^{*,i}[u_{1:t-1}] = N(\hat{x}_t^i, \mathbf{\Sigma}_t^i)$, let $\tilde{\mathbf{L}}_t^i = \theta_t^{Li}[\mathbf{\Sigma}_t], \tilde{m}_t^i = \theta_t^{mi}[\hat{x}_t, \mathbf{\Sigma}_t] = \tilde{\mathbf{M}}_t^i \hat{x}$, then

(b) For $\forall x_{1:t}^i \in (\mathcal{X}^i)^t$

$$\beta_t^{*,i}(u_t^i|u_{1:t-1}x_{1:t}^i) := \delta(u_t^i - \tilde{\mathbf{L}}_t^i x_t^i - \tilde{\mathbf{M}}_t^i \hat{x}_t) \tag{4.29a}$$

$$\mu_{t+1}^{*,i}[u_{1:t}] := N(\hat{x}_{t+1}^i, \mathbf{\Sigma}_{t+1}^i) \tag{4.29b}$$

$$\mu_{t+1}^{*}[u_{1:t}](x_t^1, x_t^2) := \prod_{i=1}^{2} \mu_{t+1}^{*,i}[u_{1:t}](x_t^i), \tag{4.29c}$$

where $\hat{x}_{t+1}^i = \phi_x^i(\hat{x}_t^i, \tilde{\mathbf{L}}_t^i, \tilde{m}_t^i, u_t)$ and $\mathbf{\Sigma}_{t+1}^i = \phi_s^i(\mathbf{\Sigma}_t^i, \tilde{\mathbf{L}}_t^i)$.

Theorem 4.2. (β^*, μ^*) constructed above is a PBE of the dynamic LQG game.

Proof. The strategy and belief profile (β^*, μ^*) is constructed using the forward recursion steps (3.12)–(3.14) on equilibrium generating function θ, which is defined through backward recursion steps (3.7)–(3.9) implemented in the proof of Theorem 4.1. Thus the result is directly implied by Theorem 3.1 in Chapter 3.

4.5 Discussion

4.5.1 Existence

In the proof of Theorem 4.1, $\tilde{\mathbf{D}}_t^{u1\dagger} \bar{\mathbf{V}}_t^1 \tilde{\mathbf{D}}_t^{u1}$ was assumed to be positive definite. This can be achieved if $\mathbf{R}^i$ is positive definite, through which it can be easily shown inductively in the proof of Theorem 4.1 that the matrices $\mathbf{V}_t^1, \bar{\mathbf{V}}_t^1$ are also positive definite.

Constructing the equilibrium generating function θ involves solving the algebraic fixed point equation in (4.23) for $\tilde{\mathbf{L}}_t$ for all $\mathbf{\Sigma}_t$. In general, the existence is not guaranteed, as is the case for existence of $\tilde{\gamma}_t$ in (3.8) for general dynamic games with asymmetric information. At this point, we don't have a general proof for existence. However, in the following lemma, we provide sufficient conditions on the matrices $\mathbf{A}_t^i, \mathbf{B}_t^i, \mathbf{T}^i, \mathbf{S}^i, \mathbf{P}^i, \mathbf{V}_{t+1}^i$ and for the case $m^i = 1$, for a solution to exist.

Lemma 4.2. For $m^1 = m^2 = 1$, there exists a solution to (4.23) if and only if for $i = 1, 2$, $\exists\, l^i \ \in \mathbb{R}^{n^i}$ such that $l^{i\dagger}\mathbf{\Delta}^i(l^1, l^2)l^i \geq 0$, or sufficiently $\mathbf{\Delta}^i(l^1, l^2) + \mathbf{\Delta}^{i,\dagger}(l^1, l^2)$ is positive definite, where $\mathbf{\Delta}^i, i = 1, 2$ are defined in Appendix B.

Proof. See Appendix B.

4.5.2 Steady state

In Chapter 3, we presented the backward/forward methodology to find SPBE for finite time-horizon dynamic games, and specialized that methodology in this chapter, in Section 4.4, to find SPBE for dynamic LQG games with asymmetric information, where equilibrium strategies are linear in players' types. It requires further investigation to find the conditions for which the backward-forward methodology could be extended to infinite time-horizon dynamic games, with either expected discounted or time-average, cost or reward criteria. Such a methodology for infinite time-horizon could be useful to characterize steady state behavior of the games. Specifically, for time homogenous dynamic LQG games with asymmetric information (where matrices $\mathbf{A}^i, \mathbf{B}^i$ are time independent), under the required technical conditions for which such a methodology is applicable, the steady state behavior can be characterized by the fixed point equation in matrices $(\mathbf{L}^i, \mathbf{\Sigma}^i, \mathbf{V}^i)_{i=1,2}$ through (4.11), (4.23b) and (4.26), where the time index is dropped in these equations, i.e.

1. $\mathbf{\Sigma} = \phi_s(\mathbf{\Sigma}, \tilde{\mathbf{L}})$
2. $\left(\tilde{\mathbf{D}}^{ui\dagger}\bar{\mathbf{V}}^i(\mathbf{\Sigma}, \tilde{\mathbf{L}})\tilde{\mathbf{D}}^{ui}\right)\tilde{\mathbf{L}}^i = -\tilde{\mathbf{D}}^{ui\dagger}\bar{\mathbf{V}}^i(\mathbf{\Sigma}, \tilde{\mathbf{L}})[\tilde{\mathbf{D}}^{ei}]_1$
3. $\mathbf{V}^i(\mathbf{\Sigma}) := \tilde{\mathbf{F}}^{i\dagger}\bar{\mathbf{V}}^i(\mathbf{\Sigma}, \tilde{\mathbf{L}})\tilde{\mathbf{F}}^i$,

where $\bar{\mathbf{V}}^i(\mathbf{\Sigma}, \tilde{\mathbf{L}}) := \begin{bmatrix} \mathbf{T}^i & \mathbf{S}^i & \mathbf{0} \\ \mathbf{S}^{i\dagger} & \mathbf{P}^i & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}^i(\phi_s(\mathbf{\Sigma}, \tilde{\mathbf{L}})) \end{bmatrix}$.

It is important to note that the steady state behavior for a general dynamic game with asymmetric information and independent types, if it exists, would involve functional fixed

point equations in value functions $(V^i(\cdot))_i$, on domain as space of $\underline{\pi}_t$. However, for the LQG case, it reduces to a fixed point equation in $(V^i(\mathbf{\Sigma}))_i$, i.e. value functions evaluated at specific $\mathbf{\Sigma}$, as shown in the above mentioned algebraic fixed point equation in matrices, which represents a significant reduction in complexity.

4.5.2.1 Numerical examples

We take a leap by assuming that methodology extends for infinite horizon problem for the model considered in this section, and present numerically found solutions for steady state as follows. We assume $\mathbf{B}^i = 0$ which implies that the state process $(X_t^i)_{t\in\mathcal{T}}$ is uncontrolled.

1. For $i = 1, 2$, $m^i = 1, n^i = 2, \mathbf{A}^i = 0.9\mathbf{I}, \mathbf{B}^i = \mathbf{0}, \mathbf{Q}^i = \mathbf{I}$,

$$\mathbf{T}^1 = \begin{bmatrix} \mathbf{I} & \frac{1}{4}\mathbf{I} \\ \frac{1}{4}\mathbf{I} & \mathbf{0} \end{bmatrix}, \quad \mathbf{T}^2 = \begin{bmatrix} \mathbf{0} & \frac{1}{4}\mathbf{I} \\ \frac{1}{4}\mathbf{I} & \mathbf{I} \end{bmatrix}, \quad \mathbf{P}^1 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$
$$\mathbf{P}^2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{S}^1 = \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{S}^2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}. \tag{4.30}$$

   This gives a symmetric solution, for $i = 1, 2$,

$$\tilde{\mathbf{L}}^i = -\begin{bmatrix} 1.062 & 1.062 \end{bmatrix}, \mathbf{\Sigma}^i = \begin{bmatrix} 3.132 & -2.132 \\ -2.132 & 3.132 \end{bmatrix}. \tag{4.31}$$

2. For $i = 1, 2$, $m^i = 2, n^i = 2, \mathbf{A}^1 = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.8 \end{bmatrix}$, $\mathbf{A}^2 = 0.9\mathbf{I}$, and $\mathbf{B}^i, \mathbf{T}^i, \mathbf{P}^i, \mathbf{S}^i$ used as before with appropriate dimensions, then,

$$\tilde{\mathbf{L}}^1 = -\begin{bmatrix} 1.680 & 1.600 \\ 0.191 & 0.286 \end{bmatrix} \quad , \tilde{\mathbf{L}}^2 = -\begin{bmatrix} 1.363 & 1.363 \\ 1.363 & 1.363 \end{bmatrix}$$
$$\mathbf{\Sigma}^1 = \mathbf{I}, \quad \mathbf{\Sigma}^2 = \begin{bmatrix} 3.132 & -2.132 \\ -2.132 & 3.132 \end{bmatrix}. \tag{4.32}$$

   It is interesting to note that for player 1, where $\mathbf{A}^1$ does not weigh the two components equally, the corresponding $\tilde{\mathbf{L}}^1$ is full rank, and thus reveals her complete private information. Whereas for player 2, where $\mathbf{A}^2$ has equal weight components, the corresponding $\tilde{\mathbf{L}}^2$ is rank deficient, which implies, at equilibrium player 2 does not completely reveal her private information. Also it is easy to check from (4.7b)

that with full rank $\tilde{\mathbf{L}}^i$ matrices, steady state $\mathbf{\Sigma}^i = \mathbf{Q}^i$.

4.6 Conclusion

In this chapter, we study a two-player dynamic LQG game with asymmetric information and perfect recall where players' private types evolve as independent controlled Markov processes. We show that under certain conditions, there exist strategies that are linear in players' private types which, together with Gaussian beliefs, form a PBE of the game. We show this by specializing the general methodology developed in chapter 3 to our model. Specifically, we prove that (a) the common beliefs remain Gaussian under the strategies that are linear in players' types where we find update equations for the corresponding mean and covariance processes; (b) using the backward recursive approach of chapter 3, we compute an equilibrium generating function θ by solving a fixed point equation in linear deterministic partial strategies γ_t for all possible common beliefs and all time epochs. Solving this fixed point equation reduces to solving a matrix algebraic equation for each realization of the state estimate covariance matrices. Also, the cost-to-go value functions are shown to be quadratic in private type and state estimates. This result is one of the very few results available on finding signaling perfect Bayesian equilibria of a truly dynamic game with asymmetric information .

4.7 Appendix A (Proof of Lemma 4.1)

This lemma could be interpreted as Theorem 2.30 in [31, Ch. 7] with appropriate matrix substitution where specifically, their observation matrix C_k should be substituted by our L_k. We provide an alternate proof here for convenience.

π^i_{t+1} is updated from π^i_t through (4.5). Since π^i_t is Gaussian, $\gamma^i_t(u^i_t|x^i_t) = \delta(u^i_t - L^i_t x^i_t - m^i_t)$ is a linear deterministic constraint and kernel Q^i is Gaussian, thus π^i_{t+1} is also Gaussian. We find its mean and covariance as follows.

We know that $X^i_{t+1} = \mathbf{A}^i_t X^i_t + \mathbf{B}^i_t U_t + W^i_t$. Then,

$$\mathbb{E}[X^i_{t+1}|\pi^i_t, \gamma^i_t, u_t] = \mathbb{E}[\mathbf{A}^i_t X^i_t + \mathbf{B}^i_t U_t + W^i_t|\pi^i_t, \gamma^i_t, u_t] \tag{4.33a}$$

$$= \mathbf{A}^i_t \mathbb{E}[X^i_t|\pi^i_t, \gamma^i_t, u_t] + \mathbf{B}^i_t u_t \tag{4.33b}$$

$$= \mathbf{A}^i_t \mathbb{E}[X^i_t|\mathbf{L}^i_t X^i_t = u^i_t - m^i_t] + \mathbf{B}^i_t u_t \tag{4.33c}$$

where (4.33b) follows because W^i_t is zero mean. Suppose there exists a matrix $\mathbf{G}^i_t$ such

that $X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i$ and $\mathbf{L}_t^i X_t^i$ are independent. Then

$$\mathbb{E}[X_t^i \big| \mathbf{L}_t^i X_t^i = u_t^i - m_t^i] = \mathbb{E}[X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i + \mathbf{G}_t^i \mathbf{L}_t^i X_t^i \big| \mathbf{L}_t^i X_t^i = u_t^i - m_t^i] \tag{4.34a}$$

$$= \mathbb{E}[X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i] + \mathbf{G}_t^i (u_t^i - m_t^i) \tag{4.34b}$$

$$= \hat{x}_t^i + \mathbf{G}_t^i (u_t^i - \mathbf{L}_t^i \hat{x}_t^i - m_t^i), \tag{4.34c}$$

where $\mathbf{G}_t^i$ satisfies

$$\mathbb{E}[(X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i)(\mathbf{L}_t^i X_t^i)^\dagger] = \mathbb{E}[(X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i)] \mathbb{E}[(\mathbf{L}_t^i X_t^i)^\dagger] \tag{4.35a}$$

$$(\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i) \mathbb{E}[X_t^i X_t^{i\dagger}] \mathbf{L}_t^{i\dagger} = (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i) \mathbb{E}[X_t^i] \mathbb{E}[X_t^{i\dagger}] \mathbf{L}_t^{i\dagger} \tag{4.35b}$$

$$(\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i)(\mathbf{\Sigma}_t^i + \hat{x}_t^i \hat{x}_t^{i\dagger}) \mathbf{L}_t^{i\dagger} = (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i) \hat{x}_t^i \hat{x}_t^{i\dagger} \mathbf{L}_t^{i\dagger} \tag{4.35c}$$

$$\mathbf{G}_t^i = \mathbf{\Sigma}_t^i \mathbf{L}_t^{i\dagger} (\mathbf{L}_t^i \mathbf{\Sigma}_t^i \mathbf{L}_t^{i\dagger})^{-1}. \tag{4.35d}$$

$$\mathbf{\Sigma}_{t+1}^i = sm \left(\mathbf{A}_t^i X_t^i - \mathbb{E}[\mathbf{A}_t^i X_t^i | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i] | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i \right) + \mathbf{Q}^i \tag{4.36a}$$

Now

$$sm \left(X_t^i - \mathbb{E}[X_t^i | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i] | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i \right) \tag{4.37a}$$

$$= sm \big((X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i)$$

$$- (\mathbb{E}[X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i]) | \mathbf{L}_t^i X_t^i = u_t^i - m_t^i \big) \tag{4.37b}$$

$$= sm \big((X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i) - (\mathbb{E}[X_t^i - \mathbf{G}_t^i \mathbf{L}_t^i X_t^i]) \big) \tag{4.37c}$$

$$= sm \big((\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i)(X_t^i - \mathbb{E}[X_t^i]) \big) \tag{4.37d}$$

$$= (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i) \mathbf{\Sigma}_t^i (\mathbf{I} - \mathbf{G}_t^i \mathbf{L}_t^i)^\dagger \tag{4.37e}$$

4.8 Appendix B (Proof of Lemma 4.2)

We prove the lemma for player 1 and the result follows for player 2 by similar arguments. For the scope of this appendix, we define $\bar{\mathbf{B}}_t^1 = \begin{bmatrix} \mathbf{B}_{1,t}^1 \\ \mathbf{B}_{1,t}^1 \\ \mathbf{B}_{1,t}^2 \end{bmatrix}$ and for any matrix $\mathbf{V}$, we define $\mathbf{V}_{*i}, \mathbf{V}_{i*}$ as the i^{th} column and the i^{th} row of $\mathbf{V}$, respectively. Then the fixed point equation

(4.23) can be written as,

$$
\begin{aligned}
0 = &\left[\mathbf{T}^1_{11} + (\mathbf{A}^1_t\mathbf{G}^1_t)^\dagger \mathbf{V}^1_{22,t+1}(\mathbf{A}^1_t\mathbf{G}^1_t) + \right. \\
&\left. \bar{\mathbf{B}}^{1\dagger}_t \mathbf{V}^1_{*2,t+1}\mathbf{A}^1_t\mathbf{G}^1_t + (\mathbf{A}^1_t\mathbf{G}^1_t)^\dagger \mathbf{V}^1_{2*,t+1}\bar{\mathbf{B}}^1_t + \bar{\mathbf{B}}^{1\dagger}_t \mathbf{V}^1_{t+1}\bar{\mathbf{B}}^1_t\right] \mathbf{L}^1_t \\
&+ \left[\mathbf{S}^{1\dagger}_{11} + (\mathbf{A}^1_t\mathbf{G}^1_t)^\dagger \mathbf{V}^1_{21,t+1}\mathbf{A}^1_t + \bar{\mathbf{B}}^{1\dagger}_t \mathbf{V}^1_{*1,t+1}\mathbf{A}^1_t\right]. \qquad (4.38)
\end{aligned}
$$

It should be noted that $\mathbf{V}^i_{t+1}$ is a function of $\mathbf{\Sigma}_{t+1}$ which is update through $\mathbf{\Sigma}_t$ and $\mathbf{L}_t$. Substituting $\mathbf{G}^1_t = \mathbf{\Sigma}^1_t\mathbf{L}^{1\dagger}_t(\mathbf{L}^1_t\mathbf{\Sigma}^1_t\mathbf{L}^{1\dagger}_t)^{-1}$ and multiplying (4.38) by $(\mathbf{L}^1_t\mathbf{\Sigma}^1_t\mathbf{L}^{1\dagger}_t)$ from left and $(\mathbf{\Sigma}^1_t\mathbf{L}^{1\dagger}_t)$ from right, we get

$$
\begin{aligned}
0 = \mathbf{L}^1_t\mathbf{\Sigma}^1_t &\left[\mathbf{L}^{1\dagger}_t(\mathbf{T}^1_{11} + \bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{t+1}\bar{\mathbf{B}}^1_t)\mathbf{L}^1_t + \mathbf{A}^{1\dagger}_t\mathbf{V}^1_{22,t+1}\mathbf{A}^1_t \right. \\
&+ \mathbf{L}^{1\dagger}_t(\bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{*2,t+1}\mathbf{A}^1_t + \mathbf{S}^{1\dagger}_{11} + \bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{*1,t+1}\mathbf{A}^1_t) \\
&\left.+(\mathbf{A}^{1\dagger}_t\mathbf{V}^1_{2*,t+1}\bar{\mathbf{B}}^1_t + \mathbf{A}^{1\dagger}_t\mathbf{V}^1_{21,t+1}\mathbf{A}^1_t)\mathbf{L}^1_t\right] \mathbf{\Sigma}^1_t\mathbf{L}^{1\dagger}_t \qquad (4.39)
\end{aligned}
$$

Let $\bar{\mathbf{L}}^i_t = \mathbf{L}^i_t(\mathbf{\Sigma}^i_t)^{1/2}$, $\bar{\mathbf{A}}^i_t = \mathbf{A}^i_t(\mathbf{\Sigma}^i_t)^{1/2}$,

$$
\begin{aligned}
\mathbf{\Lambda}^1_a(\mathbf{L}_t) &:= \mathbf{T}^1_{11} + \bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{t+1}\bar{\mathbf{B}}^1_t && (4.40a)\\
\mathbf{\Lambda}^1_b(\mathbf{L}_t) &:= \bar{\mathbf{A}}^{1\dagger}_t\mathbf{V}^1_{22,t+1}\bar{\mathbf{A}}^1_t && (4.40b)\\
\mathbf{\Lambda}^1_c(\mathbf{L}_t) &:= \bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{*2,t+1}\bar{\mathbf{A}}^1_t + \mathbf{S}^{1\dagger}_{11}(\mathbf{\Sigma}^1_t)^{1/2} + \bar{\mathbf{B}}^{1\dagger}_t\mathbf{V}^1_{*1,t+1}\bar{\mathbf{A}}^1_t && (4.40c)\\
\mathbf{\Lambda}^1_d(\mathbf{L}_t) &:= \bar{\mathbf{A}}^{1\dagger}_t\mathbf{V}^1_{2*,t+1}\bar{\mathbf{B}}^1_t + \bar{\mathbf{A}}^{1\dagger}_t\mathbf{V}^1_{21,t+1}\bar{\mathbf{A}}^1_t. && (4.40d)
\end{aligned}
$$

Then,

$$
0 = \bar{\mathbf{L}}^1_t\bar{\mathbf{L}}^{1\dagger}_t\mathbf{\Lambda}^1_a(\mathbf{L}_t)\bar{\mathbf{L}}^1_t\bar{\mathbf{L}}^{1\dagger}_t + \bar{\mathbf{L}}^1_t\mathbf{\Lambda}^1_b(\mathbf{L}_t)\bar{\mathbf{L}}^{1\dagger}_t + \bar{\mathbf{L}}^1_t\bar{\mathbf{L}}^{1\dagger}_t\mathbf{\Lambda}^1_c(\mathbf{L}_t)\bar{\mathbf{L}}^{1\dagger}_t + \bar{\mathbf{L}}^1_t\mathbf{\Lambda}^1_d(\mathbf{L}_t)\bar{\mathbf{L}}^1_t\bar{\mathbf{L}}^{1\dagger}_t \quad (4.41)
$$

Since m=1, Λ^1_a is a scalar. Let $\bar{\mathbf{L}}^i_t = \lambda^i l^{i\dagger}$, where $\lambda^i = ||\bar{\mathbf{L}}^i_t||_2$ and l^i is a normalized vector and $t^1 = T_{11}$. Moreover, since the update of $\mathbf{\Sigma}_t$ in (4.7b), is scaling invariant, $\mathbf{V}^1_{t+1}$ only depends on the directions $l = (l^1, l^2)$. Then, (4.41) reduces to the following quadratic equation in λ^1

$$
(\lambda^1)^2\Lambda^1_a(l) + \lambda^1(\mathbf{\Lambda}^1_c(l)l^1 + l^{1\dagger}\mathbf{\Lambda}^1_d(l)) + l^{1\dagger}\mathbf{\Lambda}^1_b(l)l^1 = 0. \qquad (4.42)
$$

There exists a real-valued solution[3] of this quadratic equation in λ^1 if and only if

$$(\mathbf{\Lambda}_c(l)l^1 + l^{1\dagger}\mathbf{\Lambda}_d^1(l))^2 \geq 4\Lambda_a^1(l)l^{1\dagger}\mathbf{\Lambda}_b^1(l)l^1 \tag{4.43a}$$

$$l^{1\dagger}(\mathbf{\Lambda}_c^{1\dagger}(l)\mathbf{\Lambda}_c^1(l) + \mathbf{\Lambda}_d^1(l)\mathbf{\Lambda}_d^{1\dagger}(l) + 2\mathbf{\Lambda}_d^1(l)\mathbf{\Lambda}_c^1(l) - 4\Lambda_a^1(l)\mathbf{\Lambda}_b^1(l))l^1 \geq 0. \tag{4.43b}$$

$$\text{Let } \mathbf{\Delta}^1(l) := (\mathbf{\Lambda}_c^{1\dagger}(l)\mathbf{\Lambda}_c^1(l) + \mathbf{\Lambda}_d^1(l)\mathbf{\Lambda}_d^{1\dagger}(l) + 2\mathbf{\Lambda}_d^1(l)\mathbf{\Lambda}_c^1(l) - 4\Lambda_a^1(l)\mathbf{\Lambda}_b^1(l)). \tag{4.44}$$

There exists a solution to the fixed point equation (4.23) if and only if $\exists l^1, l^2 \in \mathbb{R}^n$ such that $l^{1\dagger}\mathbf{\Delta}^1(l)l^1 \geq 0$, or sufficiently $\mathbf{\Delta}^1(l) + \mathbf{\Delta}^{1\dagger}(l)$ is positive definite.

[3]Note that a negative sign of λ^1 can be absorbed in l^1.

CHAPTER 5

Decentralized Bayesian learning in dynamic games

5.1 Introduction

In a classical Bayesian learning problem, there is a *single decision maker* who makes noisy observations of the state of nature and based on these observations eventually learns the true state. It is well known that through the likelihood ratio test, the probability of error converges exponentially fast to zero as the number of observations increases, and the true state is learnt asymptotically. With the advent of the Internet, in today's world, there are many scenarios, where strategic agents with different observations (i.e. information sets) interact with each other to learn the state of the system that in turn affects the spread of information in the system. One such scenario was studied in the seminal paper [8], where authors studied the occurrence of fads in a social network, which was later generalized in [52]. The authors in [8] and [52] study the problem of learning over a social network, where observations are made sequentially by *different decision makers* (users) who act *strategically* based on their own private information and actions of previous users. It is shown that herding (information cascade) can occur in such a case where a user discards its own private information and follows the majority action of its predecessors (fads in social networks). As a result, all future users repeat this behavior and a cascade occurs. While a good cascade is desirable, there's a positive probability of a bad cascade that hurts all the users in the community. Thus from a social (i.e. team) perspective, it is highly desirable to avoid such situations. Avoiding such bad cascades is an active area of research, for example [1] and [32] propose alternative learning models that aim at avoiding such bad cascades. There are however more general scenarios, such as cases where players participate (take actions and receive rewards) in the game more than once, deterministically or randomly through an exogenous or even an endogenous process. Furthermore there are practical

scenarios where players may be adversarial to each others' learning (e.g. dynamic zero-sum games). Studying such scenarios may reveal more interesting and richer equilibrium behaviors such as cascading phenomena, not manifested in the models considered in the current literature.

In this chapter, we study this problem from two different perspectives. Our first goal is to study this problem to design incentives to align social or team objective with strategic players' objectives, which implicitly promotes learning to continue in the game. In the second part, we seek to study learning dynamics of the system in a more general set up where players participate in the game throughout the duration of the game and not just once, as is the case for the models considered in the current literature. Since this requires studying PBE of the game, we also generalize the methodology described in chapter 3 to find perfect Bayesian equilibria (PBE) for the case where players' do not observe their types perfectly, but instead make noisy observations. This methodology then serves as a framework for studying information cascades in a more general setting.

The chapter is structured as follows. In Section 5.2, we study the problem of incentive design. Specifically, in Section 5.2.1, we present the model. In Section 5.2.2, we formulate the team problem as an instance of decentralized stochastic control and characterize its optimal policies. In Section 5.2.3, we consider the case with strategic users and design incentives for the users to align their objective with team objective. In Section 5.3, we consider a more general dynamic model. In Section 5.3.1, we provide a methodology to find a class of PBEs for such games. In Section 5.3.4, we specialize that methodology to study a specific Bayesian learning game with partially controlled observations. We characterize information cascades for this problem. While this example, limited as it is, provide analysis and intuition on the learning dynamics in decentralized games, it serves as motivation for exploring a vast landscape of the scenarios that can be studied through the proposed methodology. We conclude in Section 5.4.

5.1.1 Notation

For a probabilistic strategy profile of players $(\beta_t^i)_{i\in\mathcal{N}}$, where probability of action a_t^i conditioned on $a_{1:t-1}, x_{1:t}^i$ is given by $\beta_t^i(a_t^i|a_{1:t-1}, x_{1:t}^i)$, we use the short hand notation $\beta_t^{-i}(a_t^{-i}|a_{1:t-1}, x_{1:t}^{-i})$ to represent $\prod_{j\neq i} \beta_t^j(a_t^j|a_{1:t-1}, x_{1:t}^j)$. We use the terms users and buyers interchangeably.

5.2 Incentive design

In this section we first consider the problem of designing incentives so that the players' objectives can be aligned to the team objective. Most of the models of this problem considered in the literature assume time-invariant state of the nature. However, there are situations where the state of the nature, for e.g. popularity of a product, could change over time, as a consequence of endogenous or exogenous factors (for e.g., owing to the entering of a new competitor product or improvement/drop in quality of the product). In this section, we consider a simple scenario where users want to buy a product online. The product is either good or bad (popular or unpopular) and the value of the product (state of the system) is represented by X_t, which is changing exogenously via a Markov chain. The state is not directly observed by the users but each user receives a private noisy observation of the current state. Each user makes a decision to either buy or not buy the product, based on its private observation and action profile of all the users before its.

The strategic user wants to maximize its expected value of the product. However, its optimal action could be misaligned with the team objective of maximizing the expected average reward of the users. Thus the question we seek to address is whether it is possible to incentivize the users to align them with the team objective. To incentivize users to contribute in the learning, we assume that users can also send reports (at some cost) about their private observations after deciding to buy or to not buy the product. The idea is similar to leaving a review of the product. Thus users could be paid to report their observations to enrich the information of the future participants. Our objective is to use principles of mechanism design to construct the appropriate payment transfers (taxes/subsidies). Although, our approach deviates from general principles of mechanism design for solution of the game problem to *exactly* coincide with the team problem. However, this analysis could provide the bounds on the gap and an acceptable practical design.

5.2.1 Model

We consider a discrete-time dynamical system over infinite horizon. There is a product whose value varies over time as (a slowly varying) discrete time Markov process $(X_t)_t$, where X_t takes value in the set $\{0, 1\}$; 0 represents that product was bad (has low intrinsic value) and 1 represents and product is good (has high intrinsic value).

$$P(x_1) = \hat{Q}(x_1) \tag{5.1a}$$

$$P(x_t|x_{1:t-1}) = Q_x(x_t|x_{t-1}), \tag{5.1b}$$

such that $Q_x(x_t|x_{t-1}) = \epsilon$ if $x_t \neq x_{t-1}$, for $0 < \epsilon < 1$.

There are countably infinite number of exogenously selected, selfish buyers that act sequentially and exactly once in the process. Buyer t makes a noisy observation of the value of the product at time t, $v_t \in \mathcal{V} \stackrel{\triangle}{=} \{0, 1\}$, through a binary symmetric channel with crossover probability p such that these observations are conditionally independent across users given the system state (i.e. noise is i.i.d.) i.e. $P(v_t|x_{1:t}, v_{1:t-1}) = Q_v(v_t|x_t) = p$ if $v_t \neq x_t$. Based on actions of previous buyers and its private observation buyer t takes two actions: $a_t \in \mathcal{A} \stackrel{\triangle}{=} \{0, 1\}$, which correspond to either buying or not buying the good, and $b_t \in \mathcal{B} \stackrel{\triangle}{=} \{*, 1\}$ where * represents not reporting its observation and 1 represent reporting truthfully. Based on these actions and the state of the system, the buyer gets reward $R(x_t, a_t, b_t)$ where

$$R(x_t, a_t, b_t) = -c \cdot I(b_t = 1) + \begin{cases} 1/2, & x_t = 1, a_t = 1 \\ -1/2, & x_t = 0, a_t = 1 \\ 0, & a_t = 0 \end{cases}, \tag{5.2}$$

where c is cost of reporting its observation truthfully. The actions are publicly observed by future buyers whereas the observations $(v_t)_t$ are private information of the buyers.

5.2.2 Team problem

In this section, we study the team problem where the buyers are cooperative and want to maximize the expected average reward per unit time for the team. At time t, buyer t's information consists of its private information v_t and publicly available information $a_{1:t-1}, b_{1:t-1}$. It takes action a_t, b_t though a (deterministic) policy $g_t : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{A} \times \mathcal{B}$ as

$$(a_t, b_t) = g_t(a_{1:t-1}, b_{1:t-1}, v_t). \tag{5.3}$$

The objective as a team (or for a social planner) is to maximize the expected average reward per unit time for all the users i.e.

$$J \stackrel{\triangle}{=} \sup_g \limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^g\{R(X_t, A_t, B_t)\}. \tag{5.4}$$

Since the decision makers (i.e. the buyers) have different information sets, this is an instance of a decentralized stochastic control problem. We use techniques developed in [42] to find structural properties of the optimal policies. Specifically, we equivalently view the

system through the perspective of a common agent that observes at time t, the common information $a_{1:t-1}, b_{1:t-1}$ and takes action $\gamma_t : \mathcal{V} \to \mathcal{A} \times \mathcal{B}$, which is a partial function that, when acted upon buyer's private information v_t, generates its action (a_t, b_t). The common agent's actions $(\gamma_t)_t$ are taken through common agent's strategy $\psi = (\psi)_t$ as $\gamma_t = \psi_t[a_{1:t-1}, b_{1:t-1}]$ where $\psi_t : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \to (\mathcal{V} \to \mathcal{A} \times \mathcal{B})$. The corresponding common agent's problem is

$$J^c \triangleq \sup_{\psi} \limsup_{\tau \to \infty} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}^{\psi}\{R(X_t, A_t, B_t)\}. \tag{5.5}$$

This procedure transforms the original decentralized stochastic control problem of buyers to a centralized stochastic control problem of the common agent, which is a POMDP. Then, an optimal policy of the common agent can be translated to optimal policies for the buyers. In order to characterize common agent's optimal policy, we find an information state for the common agent's problem. We define a belief state π_t at time t as a probability measure on current state of the system given the common information i.e. $\pi_t(x_t) \triangleq P^{\psi}(x_t | a_{1:t-1}, b_{1:t-1}, \gamma_{1:t})$. The following lemma shows that the common agent faces a Markov decision problem (MDP).

Lemma 5.1. $(\Pi_t, \Gamma_t)_t$ is a controlled Markov process with state Π_t and action Γ_t such that

$$P^{\psi}(\pi_{t+1} | \pi_{1:t}, \gamma_{1:t}) = P(\pi_{t+1} | \pi_t, \gamma_t) \tag{5.6a}$$

$$\mathbb{E}^{\psi}\{R(X_t, A_t, B_t) | a_{1:t-1}, b_{1:t-1}, \gamma_{1:t}\} = \mathbb{E}\{R(X_t, A_t, B_t) | \pi_t, \gamma_t\} \tag{5.6b}$$

$$=: \hat{R}(\pi_t, \gamma_t) \tag{5.6c}$$

and there exists an update function F, independent of ψ such that $\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t)$.

Proof. See Appendix A

Lemma 5.1 implies that for common agent's problem, it can summarize the common information $a_{1:t-1}, b_{1:t-1}$ in the belief state π_t. Furthermore there exists an optimal policy for the common agent of the form $\theta_t : \mathcal{P}(\mathcal{X}) \to (\mathcal{V} \to \mathcal{A} \times \mathcal{B})$ that can be found as solution of the following dynamic programming equation in the space of public beliefs π_t as, $\forall \pi, \gamma^* = \theta[\pi]$ is the maximizer in the following equation

$$\rho + V(\pi) = \max_{\gamma} \ \hat{R}(\pi, \gamma) + \mathbb{E}\{V(\Pi') | \pi, \gamma\}, \tag{5.7}$$

where the distribution of π' is given through the kernel $P(\cdot | \pi, \gamma)$ in (5.6a) and $\rho \in \mathbb{R}, V : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ are solutions of the above fixed point equation. Based on this public belief π_t

and its private information x_t, each user t takes actions as

$$(a_t, b_t) = m_t(\pi_t, v_t) = \theta_t[\pi_t](v_t). \tag{5.8}$$

We note that since states, actions and observations belong to a binary set, there are sixteen partial functions γ possible that are shown in Table 5.1 below where $\gamma = \begin{bmatrix}\gamma(v_t=0)\\ \gamma(v_t=1)\end{bmatrix} = \begin{bmatrix}a_t, b_t(v_t=0)\\ a_t, b_t(v_t=1)\end{bmatrix}$. Since the common belief is updated as $\pi_{t+1} = F(\pi_t, \gamma, \gamma(v_t))$ and v_t is binary valued, there exist two types of γ functions: learning (γ^L) and non-learning (γ^{NL}). γ^L leads to update of belief through $F(\cdot)$ in (5.6a) that is informative of the private observation v_t, whereas γ^{NL} leads to uninformative update of belief. Eight of them are dominated in reward, for example v_t need not be reported if it is revealed through a_t, or if it can be revealed indirectly by absence of reporting.

Table 5.1: Learning vs. Non-learning γ

γ^L	$\begin{bmatrix}0,*\\1,*\end{bmatrix}$	$\begin{bmatrix}1,*\\0,*\end{bmatrix}$	$\begin{bmatrix}1,1\\1,*\end{bmatrix}$	$\begin{bmatrix}1,*\\1,1\end{bmatrix}$	$\begin{bmatrix}0,1\\0,*\end{bmatrix}$	$\begin{bmatrix}0,*\\0,1\end{bmatrix}$
	~~$\begin{bmatrix}0,1\\1,1\end{bmatrix}$~~	~~$\begin{bmatrix}1,1\\0,1\end{bmatrix}$~~	~~$\begin{bmatrix}0,1\\1,*\end{bmatrix}$~~	~~$\begin{bmatrix}1,1\\0,*\end{bmatrix}$~~	~~$\begin{bmatrix}0,*\\1,1\end{bmatrix}$~~	~~$\begin{bmatrix}1,*\\0,1\end{bmatrix}$~~
	~~$\begin{bmatrix}0,1\\0,1\end{bmatrix}$~~	~~$\begin{bmatrix}1,1\\1,1\end{bmatrix}$~~				
γ^{NL}	$\begin{bmatrix}0,*\\0,*\end{bmatrix}$	$\begin{bmatrix}1,*\\1,*\end{bmatrix}$				

5.2.3 Game problem

We now consider the case when the buyers are strategic. As before, buyer t observes public history $a_{1:t-1}, b_{1:t-1}$ and its private observation v_t and thus takes its actions as $(a_t, b_t) = g_t(a_{1:t-1}, b_{1:t-1}, v_t)$. Its objective is to maximize its expected reward

$$J_t = \max_{g_t} \mathbb{E}^g\{R(X_t, A_t, B_t)\}. \tag{5.9}$$

Since all buyers have different information, this defines a dynamic game with asymmetric information. An appropriate solution concept is Perfect Bayesian Equilibrium (PBE) [45] that requires specification of an assessment $(g_t^*, \mu_t^*)_t$ of strategy and belief profile where g_t^* is the strategy of buyer t, $g_t^* : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{P}(\mathcal{A} \times \mathcal{B})$, and μ_t^* is a belief

as a function of buyer t's history on the random variables not observed by it till time t i.e. $\mu_t^* : \mathcal{A}^{t-1} \times \mathcal{B}^{t-1} \times \mathcal{V} \to \mathcal{P}(\mathcal{X}^t \times \mathcal{V}^t)$. In general, finding a PBE is hard [45] since it involves solving a fixed point equation in strategies and beliefs that are function of histories, although there are few cases where there exists an algorithm to find them [41,57]. For this problem, since users act exactly once in the game and are thus myopic, it can be found easily in a forward inductive way, as in [8, 52]. Moreover, a belief on X_t, $\mu_t^*(x) \stackrel{\triangle}{=} P^{g^*}(X_t = x | a_{1:t-1}, b_{1:t-1}, v_t), x \in \{0, 1\}$ is sufficient and any joint belief consistent with $\mu_t^*(x)$ along with equilibrium strategy profile g^* constitute a PBE. For any history, users compute a belief equilibrium strategy depending on v_t and π_t as

$$\gamma_t^* = \phi[\pi_t] = \arg\max_{\gamma_t} \hat{R}(\pi_t, \gamma_t). \tag{5.10}$$

With $\phi[\cdot]$ defined through (5.10), for every history $(a_{1:t-1}, b_{1:t-1}, v_t)$, π_t is updated using forward recursion through $\pi_{t+1} = F(\pi_t, \phi(\pi_t), a_t, b_t)$ and equilibrium strategies are generated as $g_t^*(a_{1:t-1}, b_{1:t-1}, v_t) = \phi[\pi_t](v_t)$. Finally the beliefs μ_t^* can be easily derived from π_t and private information v_t through Bayes rule.

In order to compare the team optimal and game equilibrium policies, we numerically solve (5.7) using value iteration to find team optimal policy, shown in Figure 5.1, for parameters $p = 0.2, \epsilon = 0.001$ and $c = 0.05$. For the same parameters, Figure 5.2 shows equilibrium policy for a strategic user that solves (5.10).

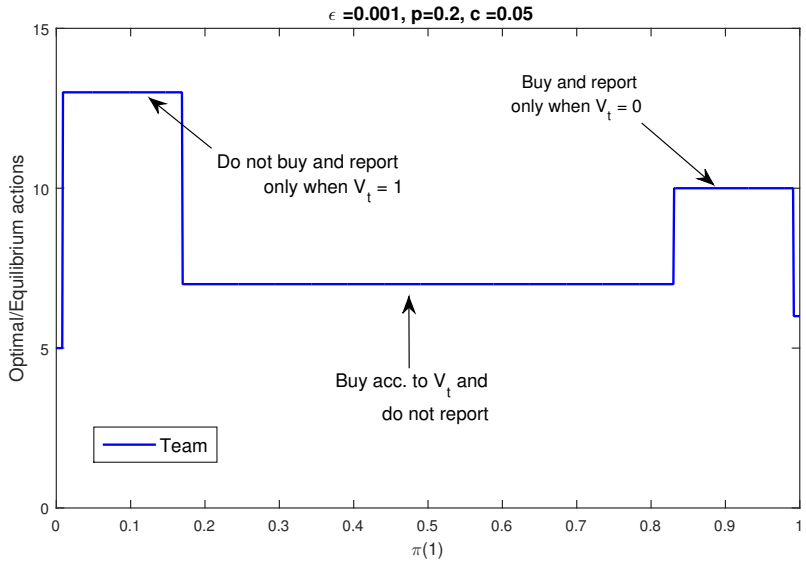


Figure 5.1: Decentralized team optimal policy

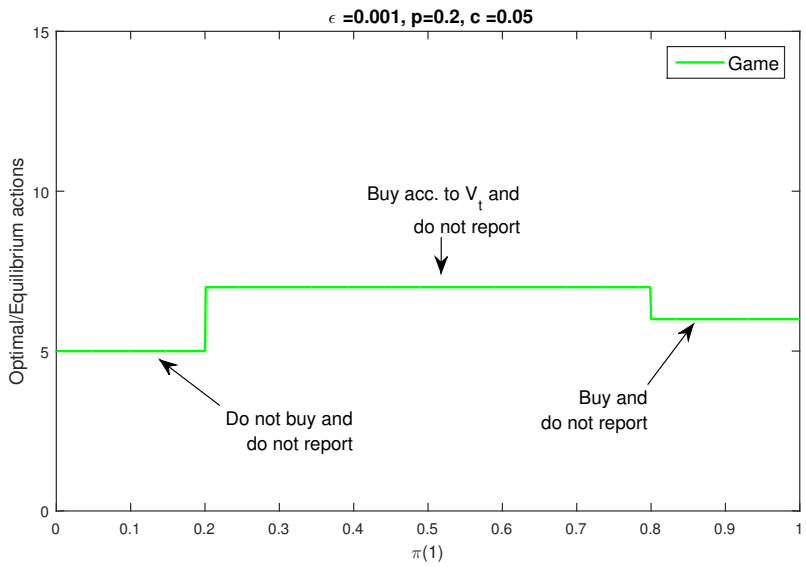


Figure 5.2: Strategic optimal policy

5.2.3.1 Incentive design for strategic users

Our goal is to align each buyers' objective with the team objective. In order to do so, we introduce incentives (tax or subsidy) for user t, $t : \mathcal{P}(\mathcal{X}) \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}$ such that its effective reward is given by $\hat{R}(\pi_t, \gamma_t) - t(\pi_t, a_t, b_t)$.

We first note that a user can not internalize social reward through incentives as is done in a pivot mechanism [5, 11, 14, 58], i.e. there does not exist an incentive mechanism such that the following equation could be true

$$\hat{R}(\pi, \gamma) - t(\pi, a, b) = \hat{R}(\pi, \gamma) + \mathbb{E}\{V(\Pi')|\pi, \gamma\} \tag{5.11}$$

$$\text{i.e.} \qquad t(\pi, a, b) = -\mathbb{E}\{V(\Pi')|\pi, \gamma\} \tag{5.12}$$

for $V(\cdot)$ defined in (5.7) and the distribution of π' is given through the kernel $P(\cdot|\pi, \gamma)$ in (5.6a). The left side of (5.11) is buyers' effective reward and right side is the objective of the team problem as in (5.7). Such a design is not feasible because while $t(\cdot)$ can depend only on public observations (π, a, b), the second term in the RHS of (5.11) depends on γ as well, which is not observed by the designer.

We observe in Figures 5.1, 5.2 that team optimal policy coincides with the strategic optimal policy for a significant range of $\pi(1)$. Let $\mathcal{S}$ be the set consisting of $\pi(1)$ where the team optimal policy coincides with the strategic optimal policy and $\mathcal{S}^c$ be the complement set. In order to align the two policies, we consider the following incentive design such that a user is paid c units by the system planner whenever the public belief $\pi(1)$ belongs to the

set $\mathcal{S}^c$, and user reports its observation,

$$t(\pi, a_t, b_t) = -c \cdot I(\pi(1) \in \mathcal{S}^c) I(b_t = 1). \tag{5.13}$$

These payments are made after any report for enforcement purposes. This is agreed upon, i.e., system planner commits to this. With these incentives, the optimal policy of the strategic user is shown in Figure 5.3. Figure 5.4 compares the time average reward achieved through these policies, found through numerical results. This shows that the gap between the team objective and the one with incentives is small. Intuitively, this occurs because the buyers learn the true state of the system relatively quickly (exponentially fast) compared to the expected time spent by the Markov process X_t in any state. Equivalently, the time spent by the process $(\Pi_t(1))_t$ in the set $\mathcal{S}^c$ is small. Yet it is crucial for the social objective that learning occurs in this region. Also in Figure 5.4, the gap between the mechanism (including incentives) and the mechanism where incentives are subtracted signifies the expected average payment made by the designer, which is relatively small.

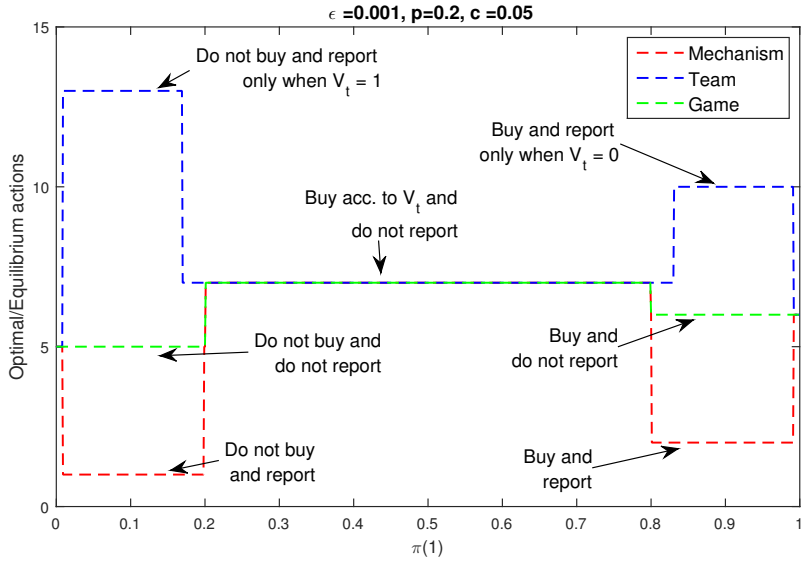


Figure 5.3: Strategic optimal policy with incentives

5.3 General framework for decentralized Bayesian learning

An indispensable tool for studying cascades is a framework for finding equilibria for these dynamical systems involving strategic players with different information sets, which are

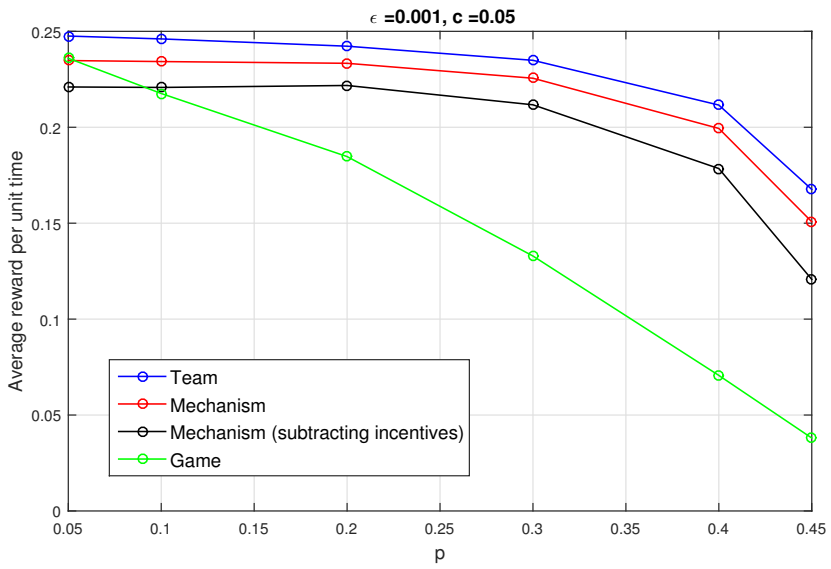


Figure 5.4: Expected time average cost comparison for different policies

modeled as dynamic games with asymmetric information. Appropriate equilibrium concepts for such games include perfect Bayesian equilibrium (PBE), sequential equilibrium, trembling hand equilibrium [13, 45]. Each of these notions of equilibrium consists of a strategy and a belief profile of all players where the equilibrium strategies are optimal given the beliefs and the beliefs are derived from the equilibrium strategy profile using Bayes' rule (whenever possible). For the games considered in the current literature including [1, 8, 32, 52], since every buyer participates only for one time period and it does not have any future individually, finding PBE reduces to solving a straightforward, one-shot optimization problem. However, for general dynamic games with asymmetric information, finding PBE is hard, since it requires solving a fixed point equation in the space of strategy and belief profiles across all users and all time periods. There is no known sequential decomposition methodology for finding PBE for such games.

In chapter 3 we presented a methodology for finding PBE for a general class of dynamic games where players types' evolve as conditionally independent Markov processes and are observed perfectly by the corresponding players. In this section, we first generalize that model to the case when players' do not perfectly observe their types; rather they make independent, noisy observations. Specifically, we consider a dynamical system where a finite number of players have different types, that evolve as conditionally independent Markov processes. Players do not observe their own types, rather make observations about them and their instantaneous rewards are a function of their current action and everyone's types. Unlike other scenarios discussed before, the proposed general framework can incorporate,

as special cases, scenarios where players participate in the game more than once, deterministically or randomly through an exogenous or endogenous process, and/or scenarios where players may be adversarial to each others' learning.

5.3.1 Model

We consider a discrete-time dynamical system with N strategic players in the set $\mathcal{N} := \{1, 2, \dots N\}$, over a finite time horizon $\mathcal{T} := \{1, 2, \dots T\}$ and with perfect recall. The system state is $x_t := (x_t^1, x_t^2, \dots x_t^N)$, where $x_t^i \in \mathcal{X}^i$ is the type of player i at time t. Players' types evolve as conditionally independent, controlled Markov processes such that

$$P(x_t|x_{1:t-1}, a_{1:t-1}) = P(x_t|x_{t-1}, a_{t-1}) \tag{5.14a}$$

$$= \prod_{i=1}^{N} Q_x^i(x_t^i|x_{t-1}^i, a_{t-1}), \tag{5.14b}$$

where $a_t = (a_t^1, \dots, a_t^N)$ and a_t^i is the action taken by player i at time t. Player i does not observe its type perfectly, rather it makes a private observation $w_t^i \in \mathcal{W}^i$ at time t, where all observations are conditionally independent across time and across players given x_t and a_{t-1}, in the following way, $\forall t \in 1, \dots T$,

$$P(w_{1:t}|x_{1:t}, a_{1:t-1}) = \prod_{n=1}^{t} \prod_{i=1}^{N} Q_w^i(w_n^i|x_n^i, a_{n-1}). \tag{5.15}$$

Player i takes action $a_t^i \in \mathcal{A}^i$ at time t upon observing $a_{1:t-1}$, which is common information among players, and $w_{1:t}^i$, which is player i's private information. The sets $\mathcal{A}^i, \mathcal{X}^i, \mathcal{W}^i$ are assumed to be finite. Let $g^i = (g_t^i)_t$ be a probabilistic strategy of player i where $g_t^i : (\times_{j=1}^{N} \mathcal{A}^j)^{t-1} \times (\mathcal{W}^i)^t \to \mathcal{P}(\mathcal{A}^i)$ such that player i plays action a_t^i according to $A_t^i \sim g_t^i(\cdot|a_{1:t-1}, w_{1:t}^i)$. Let $g := (g^i)_{i \in \mathcal{N}}$ be a strategy profile of all players. At the end of interval t, player i gets an instantaneous reward $R^i(x_t, a_t)$. The objective of player i is to maximize its total expected reward

$$J^{i,g} := \mathbb{E}^g \left[\sum_{t=1}^{T} R^i(X_t, A_t) \right]. \tag{5.16}$$

With all players being strategic, this problem is modeled as a dynamic game $\mathfrak{D}$ with imperfect and asymmetric information, and with simultaneous moves. Although this model considers all N players acting at all times, it can accommodate cases where at each time t, players are chosen through an endogenously defined (controlled) Markov process. This can

be done by introducing a nature player 0, who perfectly observes its type process $(X_t^0)_t$, has reward function zero, and plays actions $a_t^0 = w_t^0 = x_t^0$. Equivalently, all players publicly observe a controlled Markov process $(X_{t-1}^0)_t$, and a player selection process could be defined through this process. For instance, let $\mathcal{X}^0 = \mathcal{A}^0 = \mathcal{N}$, $\forall i$, $R_t^i(x_t, a_t) = 0$ if $x_t^i \neq a_t^0$, and $Q(x_{t+1}^i | x_t^i, a_t) = Q(x_{t+1}^i | x_t^i, a_t^{a_t^0})$. Here, in each period only one player acts in the game who is selected through an internal, controlled Markov process.

5.3.2 PBE of the game $\mathfrak{D}$

In this section, we provide a methodology to find PBE of the game $\mathfrak{D}$ in the domain of strategies that is time-invariant. Specifically, we seek equilibrium strategies that are structured in the sense that they depend on players' common and private information through belief states. In order to achieve this, at any time t, we summarize player i's private information, $w_{1:t}^i$, in the belief ξ_t^i, and its common information, $a_{1:t-1}$, in the belief π_t, where ξ_t^i and π_t are defined as follows. For a strategy profile g, let $\xi_t^i(x_t^i) := P^g(x_t^i | a_{1:t-1}, w_{1:t}^i)$ be the belief of player i on its current type conditioned on its information, where $\xi_t^i \in \mathcal{P}(\mathcal{X}^i)$. Also we define $\pi_t^i(\xi_t^i) := P^g(\xi_t^i | a_{1:t-1})$ as common belief on ξ_t^i based on the common information of the players, $a_{1:t-1}$, where $\pi_t^i \in \mathcal{P}(\mathcal{P}(\mathcal{X}^i))$. As it will be shown later, due to the independence of types and their evolution as independent controlled Markov processes, for any strategy profile of the players, joint beliefs on types can be factorized as product of their marginals i.e. $\pi_t(\xi_t) = \prod_{i=1}^N \pi_t^i(\xi_t^i)$. To accentuate this independence structure, we define $\underline{\pi}_t \in \times_{i \in \mathcal{N}} \mathcal{P}(\mathcal{X}^i)$ as vector of marginal beliefs where $\underline{\pi}_t := (\pi_t^i)_{i \in \mathcal{N}}$.

We now generate a player's strategy in a canonical way, as is done in decentralized team problems [43]. Using this approach, the player i's actions are generated as follows: player i at time t observes a common belief vector $\underline{\pi}_t$ and takes action γ_t^i, where $\gamma_t^i : \mathcal{P}(\mathcal{X}^i) \to \mathcal{P}(\mathcal{A}^i)$ is a partial (stochastic) function from its private belief ξ_t^i to a_t^i of the form $\gamma_t^i(a_t^i | \xi_t^i)$. These actions are generated through some policy $\theta^i = (\theta_t^i)_{t \in \mathcal{T}}$, $\theta_t^i : \times_{i \in \mathcal{N}} \mathcal{P}(\mathcal{P}(\mathcal{X}^i)) \to \{\mathcal{P}(\mathcal{X}^i) \to \mathcal{P}(\mathcal{A}^i)\}$, that operates on the common belief vector $\underline{\pi}_t$ so that $\gamma_t^i = \theta_t^i[\underline{\pi}_t]$. Then, the generated policy of the form $A_t^i \sim \theta_t^i[\underline{\pi}_t](\cdot | \xi_t^i)$ is also a policy of the form $A_t^i \sim g_t^i(\cdot | a_{1:t-1}, w_{1:t}^i)$ for an appropriately defined g. Although this is not relevant to our proofs, similar to facts 3.1 and 3.2, it can be shown that these structured policies form a sufficiently large set, which provides a good motivation for restricting attention to such equilibria. Indeed, it can be shown that policies g are outcome equivalent to policies of type θ, i.e., any expected total reward profile of the players that can be generated through a general policy profile g can also be generated through some policy profile θ. In the following Lemma, we present the update functions of the private belief ξ_t^i and the public belief π_t^i.

Lemma 5.2. There exist update functions F^i, independent of players' strategies g, such that

$$\xi^i_{t+1} = F^i(\xi^i_t, w^i_{t+1}, a_t), \tag{5.17}$$

and update functions $\bar{F}^i$, independent of θ, such that

$$\pi^i_{t+1} = \bar{F}^i(\pi^i_t, \gamma^i_t, a_t). \tag{5.18}$$

Thus $\underline{\pi}_{t+1} = \bar{F}(\underline{\pi}_t, \gamma_t, a_t)$ where $\bar{F}$ is appropriately defined through (5.18).

Proof. The proofs are straightforward using Bayes' rule and the fact that players' type and observation histories, $X^i_{1:t}, W^i_{1:t}$, are conditionally independent across players given the action history $a_{1:t-1}$, and are provided in Appendix B.

Based on (5.17), we define an update kernel of ξ^i_t in (5.56) as $Q^i(\xi^i_{t+1}|\xi^i_t, a_t) := P(\xi^i_{t+1}|\xi^i_t, a_t)$. We now present the backward-forward algorithm to find PBE of the game $\mathfrak{D}$, where strategies of the players are of type θ. The algorithm resembles the one presented in chapter 3 for perfectly observable types.

5.3.2.1 Backward recursion

In this section, we define an equilibrium generating function $\theta = (\theta^i_t)_{i\in\mathcal{N},t\in\mathcal{T}}$ and a sequence of functions
$(V^i_t)_{i\in\mathcal{N},t\in\{1,2,\dots T+1\}}$, where $V^i_t : \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{P}(\mathcal{X}^i)) \times \mathcal{P}(\mathcal{X}^i) \to \mathbb{R}$, in a backward recursive way, as follows.

1. Initialize $\forall \underline{\pi}_{T+1} \in \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{P}(\mathcal{X}^i)), \xi^i_{T+1} \in \mathcal{P}(\mathcal{X}^i)$,

$$V^i_{T+1}(\underline{\pi}_{T+1}, \xi^i_{T+1}) := 0. \tag{5.19}$$

2. For $t = T, T-1, \dots 1$, $\forall \underline{\pi}_t \in \times_{i\in\mathcal{N}}\mathcal{P}(\mathcal{P}(\mathcal{X}^i))$, let $\theta_t[\underline{\pi}_t]$ be generated as follows. Set $\tilde{\gamma}_t = \theta_t[\underline{\pi}_t]$, where $\tilde{\gamma}_t$ is the solution, if it exists[1], of the following fixed point

[1]Similar to the existence results shown in [46], it can be shown that in the special case where agent i's instantaneous reward does not depend on its private type x^i_t, and for uncontrolled types and observations, the fixed point equation always has a type-independent, myopic solution $\tilde{\gamma}^i_t(\cdot)$, since it degenerates to a best-response-like equation.

equation, $\forall i \in \mathcal{N}, \xi_t^i \in \mathcal{P}(\mathcal{X}^i)$,

$$\tilde{\gamma}_t^i(\cdot|\xi_t^i) \in \arg\max_{\gamma_t^i(\cdot|\xi_t^i)} \mathbb{E}^{\gamma_t^i(\cdot|\xi_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R^i(X_t, A_t) + V_{t+1}^i(\bar{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), \Xi_{t+1}^i)\middle|\xi_t^i\right\}, \tag{5.20}$$

where expectation in (5.20) is with respect to random variables (X_t, A_t, Ξ_{t+1}^i) through the measure
$\xi_t(x_t)\pi_t^{-i}(\xi_t^{-i})\gamma_t^i(a_t^i|\xi_t^i)\tilde{\gamma}_t^{-i}(a_t^{-i}|\xi_t^{-i})Q^i(\xi_{t+1}^i|\xi_t^i, a_t)$, F is defined in Lemma 5.5 and Q^i is defined in (5.56). Furthermore, set

$$V_t^i(\underline{\pi}_t, \xi_t^i) := \mathbb{E}^{\tilde{\gamma}_t^i(\cdot|\xi_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R^i(X_t, A_t) + V_{t+1}^i(\bar{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), \Xi_{t+1}^i)\middle|\xi_t^i\right\}. \tag{5.21}$$

It should be noted that (5.20) is a fixed point equation where the maximizer $\tilde{\gamma}_t^i$ appears in both, the left-hand-side and the right-hand-side of the equation. However, it is not the outcome of the maximization operation as in a best response equation similar to that of a Bayesian Nash equilibrium.

5.3.2.2 Forward recursion

Based on θ defined above in (5.19)–(5.21), we now construct a set of strategies β^* and beliefs μ^* for the game $\mathfrak{D}$ in a forward recursive way, as follows. As before, we will use the notation $\underline{\mu}_t^*[a_{1:t-1}] := (\mu_t^{*,i}[a_{1:t-1}])_{i\in\mathcal{N}}$ and $\mu_t^*[a_{1:t-1}]$ can be constructed from $\underline{\mu}_t^*[a_{1:t-1}]$ as $\mu_t^*[a_{1:t-1}](\xi_t) = \prod_{i=1}^N \mu_t^{*,i}[a_{1:t-1}](\xi_t^i)$ where $\mu_t^{*,i}[a_{1:t-1}]$ is a belief on ξ_t^i.

1. Initialize at time $t = 0$,

$$\mu_0^*[\phi](\xi_0) := \prod_{i=1}^N \delta_{Q_x^i}(\xi_0^i). \tag{5.22}$$

2. For $t = 1, 2 \ldots T, i \in \mathcal{N}, \forall a_{1:t}, w_{1:t}^i$

$$\beta_t^{*,i}(a_t^i|a_{1:t-1}, w_{1:t}^i) := \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]](a_t^i|\xi_t^i) \tag{5.23a}$$

$$\mu_{t+1}^{*,i}[a_{1:t}] := \bar{F}(\mu_t^{*,i}[a_{1:t-1}], \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]], a_t) \tag{5.23b}$$

where $\bar{F}$ is defined in Lemma 5.5.

Theorem 5.1. A strategy and belief profile (β^*, μ^*), constructed through backward/forward

recursive algorithm is a PBE of the game, i.e. $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, w^i_{1:t}), \beta^i$,

$$\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, w^i_{1:t} \right\}$$
$$\geq \mathbb{E}^{\beta^{i}_{t:T}\beta^{*,-i}_{t:T},\, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, w^i_{1:t} \right\}. \tag{5.24}$$

Proof. The proof relies crucially on the specific fixed point construction in (5.20) and the conditional independence structure of types and observations, and is provided in Appendix C.

5.3.3 Informational cascades

In the following definition, we define informational cascades for a dynamic game with asymmetric information, and for a given PBE of that game, as those public histories of the game for which the future actions of the players are predictable.

Definition 5.1. For a given[2] strategy and belief profile (β^*, μ^*) that constitute a PBE of the game, and for any time t and a sequence of action profile $a_{t:T}$, informational cascades can be defined as set of public histories h^c_t of the game such that at h^c_t and under (β^*, μ^*), actions $a_{t:T}$ are played almost surely, irrespective of players' future private history realizations, i.e. for a PBE (β^*, μ^*) and time t and actions $a_{t:T}$, cascades are defined by

$$\mathcal{C}^{a_{t:T}}_t := \{h^c_t \in \mathcal{H}^c_t \mid \forall i, \forall n \geq t, \forall h^i_n \text{ that are consistent with } h^c_t,$$
$$\text{that occur with non-zero probability}, \beta^{*,i}_n(a^i_n|h^i_n) = 1\}. \tag{5.25}$$

We also call an informational cascade a constant informational cascade if action profiles in the cascade are constant across time, i.e. for time t and action profile a, constant cascades are defined by

$$\mathcal{C}^{a}_t := \{h^c_t \in \mathcal{H}^c_t \mid \forall i, \forall n \geq t, \forall h^i_n \text{ that are consistent with } h^c_t,$$
$$\text{that occur with non-zero probability}, \beta^{*,i}_n(a^i|h^i_n) = 1\}. \tag{5.26}$$

For the general games considered in this section, which are dynamic game with asymmetric information and independent types, a more useful definition of cascades is the following.

[2] A stronger notion of informational cascade could be defined for *all* PBE of the game.

Definition 5.2. For a given equilibrium generating function θ, and for time t and actions $a_{t:T}$, informational cascades are defined by the sets $\{\tilde{\mathcal{C}}_t^{a_{t:T}}\}_{t=1,\dots T+1}$, which are defined as follows. For $t = T, T-1, \dots 1$,

$$\tilde{\mathcal{C}}_{T+1} := \Big\{ \text{ All possible common beliefs } \underline{\pi}_{T+1} \Big\} \tag{5.27}$$

$$\tilde{\mathcal{C}}_t^{a_{t:T}} := \Big\{ \underline{\pi}_t \mid \forall i,\ \forall \xi_t^i \in supp(\pi_t^i), \theta_t^i[\underline{\pi}_t](a_t^i|\xi_t^i) = 1 \text{ and } \bar{F}(\underline{\pi}_t, \theta_t[\underline{\pi}_t], a_t) \in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}} \Big\} \tag{5.28}$$

A constant informational cascade for time t and actions profile a is defined as,

$$\tilde{\mathcal{C}}_{T+1} := \Big\{ \text{ All possible common beliefs } \underline{\pi}_{T+1} \Big\} \tag{5.29}$$

$$\tilde{\mathcal{C}}_t^{a} := \Big\{ \underline{\pi}_t \mid \forall i,\ \forall \xi_t^i \in supp(\pi_t^i), \theta_t^i[\underline{\pi}_t](a^i|\xi_t^i) = 1 \text{ and } \bar{F}(\underline{\pi}_t, \theta_t[\underline{\pi}_t], a) \in \tilde{\mathcal{C}}_{t+1}^{a} \Big\} \tag{5.30}$$

In the following lemma, we show the connection between the two definitions.

Lemma 5.3. Let (β^*, μ^*) be an SPBE of a dynamic game with asymmetric information and independent types, generated by an equilibrium generating function θ. Then $\forall t, a_{t:T}$,

$$(\mu_t^*)^{-1}(\tilde{\mathcal{C}}_t^{a_{t:T}}) = \mathcal{C}_t^{a_{t:T}} \tag{5.31}$$

Proof. See Appendix E.

Corollary 5.1. Let (β^*, μ^*) be an SPBE of a dynamic game with asymmetric information and independent types, generated by an equilibrium generating function θ. Then $\forall t, a$,

$$(\mu_t^*)^{-1}(\tilde{\mathcal{C}}_t^{a}) = \mathcal{C}_t^{a} \tag{5.32}$$

5.3.4 Specific learning model

We now consider a specific model that captures the learning aspect in a dynamic setting with strategic agents and decentralized information. The model is similar in spirit to the model considered in [8, 52] except we consider a finite number of players who take action in every epoch and stay in the game throughout the entire duration of the game. We assume that players' types are uncontrollable and static i.e. $Q_x^i(x_{t+1}^i|x_t^i, a_t) = \delta_{x_t^i}(x_{t+1}^i)$, where $\mathcal{X}^i = \{-1, 1\}$. Since the set of types, $\mathcal{X}^i$ is has cardinality 2, the measure ξ_t^i can be sufficiently described by $\xi_t^i(1)$. Henceforth, in this section and in Appendix 5.10, with slight abuse of notation, we denote $\xi_t^i(1)$ by $\xi_t^i \in [0, 1]$. In each epoch t, player i makes independent observation w_t^i about its type where $\mathcal{W}^i = \{-1, 1\}$, through an observation kernel

of the form $Q^i_w(w^i_t|x^i_t, a^i_{t-1})$, which does not depend on a^{-i}_{t-1}. Based on its information, it takes action a^i_t, where $\mathcal{A}^i = \{0, 1\}$, and earns an instantaneous reward given by

$$R^i(x, a^i_t) = a^i_t \left(\lambda x^i + \bar{\lambda} \frac{\sum_{j \neq i} x^j}{N-1} \right), \tag{5.33}$$

where $\lambda \in [0, 1], \bar{\lambda} = 1 - \lambda$. This scenario can thought of the case when players' types represent their talent, capabilities or popularity, and a player makes a decision to either choose (action = 1) or not choose (action = 0) these players, where its instantaneous reward depends on some combination of the capabilities of all the players. We note that the instantaneous reward does not depend on other players' actions but on their types, and thus learning players' types is an important aspect of the problem.

5.3.4.1 Partially controlled observations

We consider the case where observations of the player i do depend on other players' actions, i.e. the observation kernel is of the form $Q^i_w(w^i_t|x^i_t, a^i_{t-1})$. These observations are made through a binary symmetric channel such that $Q^i_w(1|1, a^i) = Q^i_w(-1| - 1, a^i) = 1 - p_{a^i}$ and $Q^i_w(-1|1, a^i) = Q^i_w(1| - 1, a^i) = p_{a^i}$, where $p_1 \leq p_0 < 1/2$. This model implies that taking action 1 can improve the quality of a player's future private belief. In this case, the update functions of ξ^i_t and π^i_t in (5.17), (5.18) reduce to

$$\xi^i_{t+1} = F^i(\xi^i_t, w^i_{t+1}, a^i_t), \tag{5.34a}$$

$$\pi^i_{t+1} = \bar{F}^i(\pi^i_t, \gamma^i_t, a^i_t), \tag{5.34b}$$

and (5.20) in the backward recursion reduces to

$$\begin{aligned} \tilde{\gamma}^i_t(\cdot|\xi^i_t) \in \arg \max_{\gamma^i_t(\cdot|\xi^i_t)} \sum_{a^i_t} a^i_t \gamma^i_t(a^i_t|\xi^i_t)(\lambda(2\xi^i_t - 1) + \bar{\lambda}(2\hat{\xi}^{-i}_t - 1)) \\ + \mathbb{E}^{\gamma^i_t(\cdot|\xi^i_t)\tilde{\gamma}^{-i}_t, \pi_t} \left\{ V^i_{t+1}(\bar{F}(\underline{\pi}_t, \tilde{\gamma}_t, A^{-i}_t), \Xi^i_{t+1}) \middle| \xi^i_t \right\}, \end{aligned} \tag{5.35}$$

For the learning model considered in Section 5.3.4, we characterize constant informational cascades through a time invariant set $\hat{\mathcal{C}}^a$ of common beliefs $\underline{\pi}$, defined as follows.

Let

$$\hat{\mathcal{C}}^a := \Big\{ \underline{\pi} \mid \forall i, \frac{1}{2} - \frac{\bar{\lambda}}{\lambda}(\hat{\xi}^{-i} - \frac{1}{2}) \geq 1 \text{ if } a^i = 0, \\ \frac{1}{2} - \frac{\bar{\lambda}}{\lambda}(\hat{\xi}^{-i} - \frac{1}{2}) \leq 0 \text{ if } a^i = 1 \Big\} \tag{5.36}$$

where

$$\hat{\xi}^{-i} := \frac{1}{N-1} \sum_{j \neq i} \mathbb{E}^{\pi^j}[\Xi^j]. \tag{5.37}$$

In the following theorem we show that the set $\hat{\mathcal{C}}^a$ defined in (5.36) characterizes a set of constant informational cascades for this problem. Specifically, we show that $\hat{\mathcal{C}}^a \subset \tilde{\mathcal{C}}^a$.

Theorem 5.2. If, for some time t_0 and action profile a, $\underline{\pi}_{t_0} \in \hat{\mathcal{C}}^a$, then $\forall t \geq t_0, \underline{\pi}_t \in \hat{\mathcal{C}}^a$ and solutions of (5.35) satisfy $\tilde{\gamma}^i_t(a^i|\xi^i_t) = 1 \; \forall \xi^i_t \in [0,1]$. Moreover, for $t_0 \leq t \leq T$, V^i_t is given by

$$V^i_t(\underline{\pi}^{-i}_t, \xi^i_t) = (T-t+1)(\lambda(2\xi^i_t - 1) + \bar{\lambda}(2\hat{\xi}^{-i}_t - 1))a^i \quad \forall \underline{\pi}_t \in \hat{\mathcal{C}}^a. \tag{5.38}$$

Proof. See Appendix F.

5.3.5 Discussion

We characterize informational cascades by those histories of the game where learning stops for the players as a whole. Conceptually, they could be thought of as absorbing states of the system. It begets questions regarding the dynamics of the process that could lead to those states, for example hitting times of such sets and absorption probabilities. For the simplified problem considered in [8], cascades can be characterized as the fixed points of common belief update function, so that the common belief gets "stuck" once it reaches that state. It was shown that cascades eventually occur with probability 1 for that model. For the learning model considered in this section, common beliefs π_t still evolve in a cascade, although uninformatively, i.e., their evolution is directed by the primitives of the process and not on the new random variables being generated, namely, players' private observations. Also, if players' observations are informative, they asymptotically learn their true types, i.e., their private beliefs converge to their true types. One trivial case when cascades could occur for this model is if the system was born in a cascade, i.e., the initial common belief, based on the prior distributions, is in cascades, $\pi_1 \in \hat{C}^a$. In general, a cascade could occur

as follows. Suppose all players have low types (i.e. $x^i = -1$), but they get atypical observations initially, which lead them into believing that their types are high ($x^i = 1$). This information is conveyed through their actions, which leads the public belief into a cascade. Interestingly, even though players eventually learn their true types, yet they remain in a (bad) cascade, each player believing that others have high types on average.

5.4 Conclusion

In this chapter we studied Bayesian learning dynamics of specific dynamic games with asymmetric information. We first considered an ergodic sequential buyers' game where a countable number of strategic buyers buy a product exactly once in the game. We model the team problem as an instance of decentralized stochastic control problem and characterize structure of optimum policies. When users are strategic, it is modeled as a dynamic game with asymmetric information. We show that for some set $\pi_t \in \mathcal{S}$ that occurs with high probability, the strategic optimal policy coincides with the team optimal policy. Thus only outside this set, i.e., when $\pi_t \in \mathcal{S}^c$, buyers need to be incentivized to report their observations so that higher average rewards can be achieved for the whole team. Since numerically $\mathcal{S}^c$ occurs with low probability, the expected incentive payments are low. However, even though infrequent, these incentives help in the learning for the team as a whole, specifically for the future users. This suggests that using such a mechanism for the more general case could be a useful way to bridge the gap between strategic and team objectives.

In second part, we considered a more general scenario where players could participate in the game throughout the duration of the game. Players' types evolved as conditionally independent controlled Markov processes and players made noisy observations of their types. We first presented a sequential decomposition methodology to find SPBE of the game. We then studied a specific learning model and characterized information cascades using the general methodology described before. In general, the methodology presented serves as a framework for studying learning dynamics of decentralized systems with strategic agents. Some important research directions include characterization of cascades for specific classes of models, studying convergent learning behavior in such games including the probability and the rate of "falling" into a cascade, and incentive or mechanism design to avoid bad cascades.

5.5 Appendix A (Proof of Lemma 5.1)

Claim 5.1. There exists an update function F, independent of ψ such that $\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t)$.

Proof. Fix ψ

$$\pi_{t+1}(x_{t+1}) = P^{\psi}(x_{t+1}|a_{1:t}, b_{1:t}, \gamma_{1:t}) \tag{5.39a}$$

$$= \sum_{x_t} P^{\psi}(x_{t+1}, x_t|a_{1:t}, b_{1:t}, \gamma_{1:t}) \tag{5.39b}$$

$$= \sum_{x_t} P^{\psi}(x_t|a_{1:t}, b_{1:t}, \gamma_{1:t})\hat{Q}(x_{t+1}|x_t) \tag{5.39c}$$

Now,

$$P^{\psi}(x_t|a_{1:t}, b_{1:t}, \gamma_{1:t}) = \frac{P^{\psi}(x_t, a_t, b_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t})}{\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t, a_t, b_t|a_{1:t-1}, b_{1:t-1}, \gamma_{1:t})} \tag{5.40a}$$

$$= P^{\psi}(x_t|a_{1:t-1}, b_{1:t-1}, \gamma_{1:t})\times \frac{\sum_{v_t} P^{\psi}(a_t, b_t, v_t|a_{1:t-1}, b_{1:t-1}, \gamma_{1:t}, x_t)}{\sum_{\hat{x}_t} P(\hat{x}_t, a_t, b_t|a_{1:t-1}, b_{1:t-1}, \gamma_{1:t})} \tag{5.40b}$$

$$= \frac{P^{\psi}(x_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t-1})\sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t)}{\sum_{\hat{x}_t} P^{\psi}(\hat{x}_t|a_{1:t-1}b_{1:t-1}, \gamma_{1:t-1})\sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|\hat{x}_t)} \tag{5.40c}$$

where first part in numerator in (5.40c) is true since given policy ψ, γ_t can be computed as $\gamma_t = \psi_t(a_{1:t-1}, b_{1:t-1})$.

We conclude that

$$P(x_t|a_{1:t}, \gamma_{1:t}) = \frac{\pi_t(x_t)\sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|x_t)}{\sum_{\hat{x}_t} \pi_t(\hat{x}_t)\sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t)Q_v(v_t|\hat{x}_t)}, \tag{5.41}$$

thus,

$$\pi_{t+1} = F(\pi_t, \gamma_t, a_t, b_t) \tag{5.42}$$

where F is independent of policy ψ.

Claim 5.2. $(\Pi_t, \Gamma_t)_t$ is a controlled Markov process with state Π_t and action Γ_t such that

$$P^{\psi}(\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}) = P(\pi_{t+1}|\pi_t, \gamma_t) \tag{5.43}$$

$$\mathbb{E}^{\psi}\{R(X_t, A_t, B_t)|a_{1:t-1}, b_{1:t-1}, \gamma_{1:t}\} = \mathbb{E}\{R(X_t, A_t, B_t)|\pi_t, \gamma_t\} \tag{5.44}$$

$$=: \hat{R}(\pi_t, \gamma_t) \tag{5.45}$$

Proof.

$$P^{\psi}(\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}) = \sum_{a_t, b_t} P^{\psi}(\pi_{t+1}, a_t, b_t|\pi_{1:t}, \gamma_{1:t}) \tag{5.46a}$$

$$= \sum_{a_t, b_t} \mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) \sum_{v_t} P^{\psi}(a_t, b_t, v_t|\pi_{1:t}, \gamma_{1:t}) \tag{5.46b}$$

$$= \sum_{a_t, b_t, x_t} \mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) P^{\psi}(x_t|\pi_{1:t}, \gamma_{1:t}) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t) Q_v(v_t|x_t) \tag{5.46c}$$

$$= \sum_{a_t, b_t, x_t} \pi_t(x_t) \mathbf{1}_{\{F(\pi_t, \gamma_t, a_t, b_t)\}}(\pi_{t+1}) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t) Q_v(v_t|x_t) \tag{5.46d}$$

$$= P(\pi_{t+1}|\pi_t, \gamma_t) \tag{5.46e}$$

$$\mathbb{E}(R(X_t, A_t, B_t)|\pi_{1:t}, \gamma_{1:t}) = \sum_{x_t, a_t, b_t v_t} R(x_t, a_t, b_t) P(x_t, a_t, b_t, v_t|\pi_{1:t}, \gamma_{1:t}) \tag{5.47a}$$

$$= \sum_{x_t, a_t, b_t} R(x_t, a_t, b_t) P(x_t|\pi_{1:t}, \gamma_{1:t}) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t) Q_v(v_t|x_t) \tag{5.47b}$$

$$= \sum_{x_t, a_t, b_t} R(x_t, a_t, b_t) \pi_t(x_t) \sum_{v_t} I_{\{\gamma_t(v_t)\}}(a_t, b_t) Q_v(v_t|x_t) \tag{5.47c}$$

$$= \hat{R}(\pi_t, \gamma_t) \tag{5.47d}$$

5.6 Appendix B (Proof of Lemma 5.2)

Lemma 5.4. There exists an update function F^i, independent of g, such that

$$\xi^i_{t+1} = F^i(\xi^i_t, w^i_{t+1}, a_t). \tag{5.48}$$

Proof. We first prove the following Claim on conditional independence of $x_{1:t}, w_{1:t}$ given $a_{1:t-1}$.

Claim 5.3. For any policy profile g and $\forall t$,

$$P^g(x_{1:t}, w_{1:t}|a_{1:t-1}) = \prod_{i=1}^{N} P^{g^i}(x^i_{1:t}, w^i_{1:t}|a_{1:t-1}) \tag{5.49}$$

Proof.

$$\begin{aligned}
&P^g(x_{1:t}, w_{1:t}|a_{1:t-1}) \\
&= \frac{P^g(x_{1:t}, w_{1:t}, a_{1:t-1})}{\sum_{x_{1:t}, w_{1:t}} P^g(x_{1:t}, w_{1:t}, a_{1:t-1})} && \text{(5.50a)} \\
&= \frac{\begin{array}{r}\prod_{i=1}^{N} Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}}{\begin{array}{r}\sum_{x_{1:t}, w_{1:t}} \prod_{i=1}^{N} Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}} && \text{(5.50b)} \\
&= \frac{\begin{array}{r}\prod_{i=1}^{N} Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}}{\begin{array}{r}\prod_{i=1}^{N} \sum_{x^i_{1:t}, w^i_{1:t}} Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}} && \text{(5.50c)} \\
&= \prod_{i=1}^{N} \frac{\begin{array}{r}Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}}{\begin{array}{r}\sum_{x^i_{1:t}, w^i_{1:t}} Q^i_x(x^i_1) Q^i_w(w^i_1|x^i_1) \prod_{n=1}^{t-1} g^i_n(a^i_n|a_{1:n-1}, w^i_{1:n-1}) Q^i_x(x^i_{n+1}|a_n, x^i_n) \\ Q^i_w(w^i_{n+1}|x^i_{n+1}, a_n)\end{array}} && \text{(5.50d)} \\
&= \prod_{i=1}^{N} P^{g^i}(x^i_{1:t}, w^i_{1:t}|a_{1:t-1}) && \text{(5.50e)}
\end{aligned}$$

Now for any g we have,

$$\xi^i_{t+1}(x^i_{t+1}) \stackrel{\triangle}{=} P^g(x^i_{t+1}|a_{1:t}, w^i_{1:t+1}) \tag{5.51a}$$

$$= \frac{\sum_{x^i_t} P^g(x^i_t, a_t, x^i_{t+1}, w^i_{t+1}|a_{1:t-1}, w^i_{1:t})}{\sum_{\tilde{x}^i_{t+1}\tilde{x}^i_t} P^g(\tilde{x}^i_t, a_t, w^i_{t+1}, \tilde{x}^i_{t+1}|a_{1:t-1}, w^i_{1:t})} \tag{5.51b}$$

$$= \frac{\sum_{x^i_t} \xi^i_t(x^i_t) P^g(a^{-i}_t|a_{1:t-1}, w^i_{1:t}, x^i_t) Q^i_x(x^i_{t+1}|a_t, x^i_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t)}{\sum_{\tilde{x}^i_{t+1}\tilde{x}^i_t} \xi^i_t(\tilde{x}^i_t) P^g(a^{-i}_t|a_{1:t-1}, w^i_{1:t}, \tilde{x}^i_t) Q^i_x(\tilde{x}^i_{t+1}|a_t, \tilde{x}^i_t) Q^i_w(w^i_{t+1}|\tilde{x}^i_{t+1}, a_t)}, \tag{5.51c}$$

where (5.51c) is true because a^i_t is a function of $(a_{1:t-1}, w^i_{1:t})$ and thus term involving can be cancelled in numerator and denominator. We now consider the quantity $P^g(a^{-i}_t|a_{1:t-1}w^i_{1:t}x^i_t)$

$$P^g(a^{-i}_t|a_{1:t-1}, w^i_{1:t}, x^i_t) = \sum_{w^{-i}_{1:t}} P^g(a^{-i}_t, w^{-i}_{1:t}|a_{1:t-1}, w^i_{1:t}, x^i_t) \tag{5.52a}$$

$$= \sum_{w^{-i}_{1:t}} P^g(w^{-i}_{1:t}|a_{1:t-1}, w^i_{1:t}, x^i_t) \prod_{j\neq i} g^j_t(a^j_t|a_{1:t-1}, w^j_{1:t}) \tag{5.52b}$$

$$= \sum_{w^{-i}_{1:t}} P^{g^{-i}}(w^{-i}_{1:t}|a_{1:t-1}) \prod_{j\neq i} g^j_t(a^j_t|a_{1:t-1}, w^j_{1:t}) \tag{5.52c}$$

$$= P^{g^{-i}}(a^{-i}_t|a_{1:t-1}) \tag{5.52d}$$

where (5.52c) follows from Claim 5.3 in Appendix A since $w^{-i}_{1:t}$ is conditionally independent of $(w^i_{1:t}, x^i_t)$ given $a_{1:t-1}$ and is only a function of g^{-i}. Since this term does not depend on x^i_t, it gets cancelled in the final expression of ξ^i_{t+1}

$$\xi^i_{t+1}(x^i_{t+1}) = \frac{\sum_{x^i_t} \xi^i_t(x^i_t) Q^i_x(x^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|x^i_{t+1}, a_t)}{\sum_{\tilde{x}^i_{t+1}} \sum_{x^i_t} \xi^i_t(x^i_t) Q^i_x(\tilde{x}^i_{t+1}|x^i_t, a_t) Q^i_w(w^i_{t+1}|\tilde{x}^i_{t+1}, a_t)}. \tag{5.53}$$

Thus the claim of the Lemma follows. Based on this claim, we can conclude that

$$\xi^i_t(x^i_t) \;=\; P^g(x^i_t|a_{1:t-1}, w^i_{1:t}) \;=\; P(x^i_t|a_{1:t-1}, w^i_{1:t}). \tag{5.54}$$

Also, based on the update of ξ_t^i in (5.48), we define an update kernel

$$Q^i(\xi_{t+1}^i|\xi_t^i, a_t) := P(\xi_{t+1}^i|\xi_t^i, a_t) \tag{5.55}$$

$$= \sum_{x_t^i, x_{t+1}^i, w_{t+1}^i} \xi_t^i(x_t^i) Q_x^i(x_{t+1}^i|x_t^i, a_t) Q_w^i(w_{t+1}^i|x_{t+1}^i, a_t) I_{F(\xi_t^i, w_{t+1}^i, a_t^i)}(\xi_{t+1}^i) \tag{5.56}$$

Lemma 5.5. There exists an update function $\bar{F}$ of π_t, independent of ψ

$$\pi_{t+1}^i = \bar{F}(\pi_t^i, \gamma_t^i, a_t) \tag{5.57}$$

Proof.

$$\pi_{t+1}(\xi_{t+1})$$

$$= P^\psi(\xi_{t+1}|a_{1:t}, \gamma_{1:t+1}) \tag{5.58a}$$

$$= P^\psi(\xi_{t+1}|a_{1:t}, \gamma_{1:t}) \tag{5.58b}$$

$$= \frac{\sum_{\xi_t, x_t, x_{t+1}, w_{t+1}} P^\psi(\xi_t, x_t, a_t, x_{t+1}, w_{t+1}, \xi_{t+1}|a_{1:t-1}, \gamma_{1:t})}{\sum_{\xi_t} P^\psi(\xi_t, a_t|a_{1:t-1}, \gamma_{1:t})} \tag{5.58c}$$

$$= \frac{\sum_{\xi_t, x_t, x_{t+1}, w_{t+1}} \prod_{i=1}^N \pi_t^i(\xi_t^i) \xi_t^i(x_t^i) \gamma_t^i(a_t^i|\xi_t^i) Q_x^i(x_{t+1}^i|x_t^i, a_t) Q_w^i(w_{t+1}^i|x_{t+1}^i, a_t) I_{F^i(\xi_t^i, w_{t+1}^i, a_t)}(\xi_{t+1}^i)}{\sum_{\xi_t} \prod_{i=1}^N \pi_t^i(\xi_t^i) \gamma_t^i(a_t^i|\xi_t^i)} \tag{5.58d}$$

$$= \prod_{i=1}^N \frac{\sum_{\xi_t^i, x_t^i, x_{t+1}^i, w_{t+1}^i} \pi_t^i(\xi_t^i) \xi_t^i(x_t^i) \gamma_t^i(a_t^i|\xi_t^i) Q_x^i(x_{t+1}^i|x_t^i, a_t) Q_w^i(w_{t+1}^i|x_{t+1}^i, a_t) I_{F^i(\xi_t^i, w_{t+1}^i, a_t)}(\xi_{t+1}^i)}{\sum_{\xi_t^i} \pi_t^i(\xi_t^i) \gamma_t^i(a_t^i|\xi_t^i)} \tag{5.58e}$$

Thus we have,

$$\pi_{t+1} = \prod_{i=1}^N \bar{F}(\pi_t^i, \gamma_t^i, a_t) \tag{5.58f}$$

5.7 Appendix C (Proof of Theorem 5.1)

Proof. We prove (5.24) using induction and from results in Lemma 5.6, 5.7 and 5.8 proved in Appendix D. For base case at $t = T$, $\forall i \in \mathcal{N}, (a_{1:T-1}, w^i_{1:T}) \in \mathcal{H}^i_T, \beta^i$

$$\mathbb{E}^{\beta^{*,i}_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i(X_T, A_T) \middle| a_{1:T-1}, w^i_{1:T} \right\} = V^i_T(\underline{\mu}^*_T[a_{1:T-1}], \xi^i_T) \tag{5.59a}$$

$$\geq \mathbb{E}^{\beta^i_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i(X_T, A_T) \middle| a_{1:T-1}, w^i_{1:T} \right\} \tag{5.59b}$$

where (5.59a) follows from Lemma 5.8 and (5.59b) follows from Lemma 5.6 in Appendix D.

Let the induction hypothesis be that for $t+1$, $\forall i \in \mathcal{N}, (a_{1:t}, w^i_{1:t+1}) \in \mathcal{H}^i_{t+1}, \beta^i$,

$$\mathbb{E}^{\beta^{*,i}_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, w^i_{1:t+1} \right\}$$

$$\geq \mathbb{E}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, w^i_{1:t+1} \right\}. \tag{5.60a}$$

Then $\forall i \in \mathcal{N}, (a_{1:t-1}, w^i_{1:t}) \in \mathcal{H}^i_t, \beta^i$, we have

$$\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, w^i_{1:t} \right\}$$

$$= V^i_t(\underline{\mu}^*_t[a_{1:t-1}], \xi^i_t) \tag{5.61a}$$

$$\geq \mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \left\{ R^i(X_t, A_t) + W^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], \Xi^i_{t+1}) \big| a_{1:t-1}, w^i_{1:t} \right\} \tag{5.61b}$$

$$= \mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \big\{ R^i(X_t, A_t) + \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t-1}, A_t]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, w^i_{1:t} W^i_{t+1} \right\} \big| a_{1:t-1}, w^i_{1:t} \right\} \tag{5.61c}$$

$$\geq \mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \big\{ R^i(X_t, A_t) + \mathbb{E}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T}\mu^*_{t+1}[a_{1:t-1}, A_t]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, w^i_{1:t}, W^i_{t+1} \right\} \big| a_{1:t-1}, w^i_{1:t} \right\} \tag{5.61d}$$

$$= \mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \big\{ R^i(X_t, A_t) + \mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}\mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, A_t, w^i_{1:t}, W^i_{t+1} \right\} \big| a_{1:t-1}, w^i_{1:t} \right\} \tag{5.61e}$$

$$= \mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i(X_n, A_n) \big| a_{1:t-1}, w^i_{1:t} \right\}, \tag{5.61f}$$

where (5.61a) follows from Lemma 5.8, (5.61b) follows from Lemma 5.6, (5.61c) follows from Lemma 5.8, (5.61d) follows from induction hypothesis in (5.60a) and (5.61e) follows from Lemma 5.7. Moreover, construction of θ in (5.20), and consequently definition of β^* in (5.23a) are pivotal for (5.61e) to follow from (5.61d).

We note that μ^* satisfies the consistency condition of [13, p. 331] from the fact that (a) for all t and for every common history $a_{1:t-1}$, all players use the same belief $\mu^*_t[a_{1:t-1}]$ on x_t and (b) the belief μ^*_t can be factorized as $\mu^*_t[a_{1:t-1}] = \prod_{i=1}^{N} \mu^{*,i}_t[a_{1:t-1}] \ \forall a_{1:t-1} \in \mathcal{H}^c_t$ where $\mu^{*,i}_t$ is updated through Bayes' rule ($\bar{F}$) as in Lemma 5.5 in Appendix A.

5.8 Appendix D

Lemma 5.6. $\forall t \in \mathcal{T}, i \in \mathcal{N}, (a_{1:t-1}, w^i_{1:t}) \in \mathcal{H}^i_t, \beta^i_t$

$$V^i_t(\underline{\mu}^*_t[a_{1:t-1}], \xi^i_t) \geq$$
$$\mathbb{E}^{\beta^i_t \beta^{*,-i}_t, \mu^*_t[a_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[a_{1:t-1}], \beta^*_t(\cdot|a_{1:t-1}, \cdot), A_t), \Xi^i_{t+1}) \middle| a_{1:t-1}, w^i_{1:t} \right\}. \tag{5.62}$$

Proof. We prove this Lemma by contradiction.

Suppose the claim is not true for t. This implies $\exists i, \hat{\beta}^i_t, \hat{a}_{1:t-1}, \hat{w}^i_{1:t}$ such that

$$\mathbb{E}^{\hat{\beta}^i_t \beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), \Xi^i_{t+1}) \middle| \hat{a}_{1:t-1}, \hat{w}^i_{1:t} \right\}$$
$$> V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{\xi}^i_t). \tag{5.63}$$

We will show that this contradicts the definition of W^i_t in (5.21).

Construct $\hat{\gamma}^i_t(a^i_t|\xi^i_t) = \begin{cases} \hat{\beta}^i_t(a^i_t|\hat{a}_{1:t-1}, \hat{w}^i_{1:t}) & \xi^i_t = \hat{\xi}^i_t \\ \text{arbitrary} & \text{otherwise.} \end{cases}$

Then for $\hat{a}_{1:t-1}, \hat{w}^i_{1:t}$, we have

$$V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{\xi}^i_t) =$$
$$\max_{\gamma^i_t(\cdot|\hat{\xi}^i_t)} \mathbb{E}^{\gamma^i_t(\cdot|\hat{\xi}^i_t)\beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), \Xi^i_{t+1}) \middle| \hat{\xi}^i_t \right\} \tag{5.64a}$$
$$\geq \mathbb{E}^{\hat{\gamma}^i_t(\cdot|\hat{\xi}^i_t)\beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), \Xi^i_{t+1}) \middle| \hat{\xi}^i_t \right\} \tag{5.64b}$$
$$= \sum_{\xi^{-i}_t, a_t, \xi_{t+1}} \left\{ R^i(x_t, a_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), a_t), \xi^i_{t+1}) \right\} \times$$
$$\hat{\xi}^i_t(x^i_t) \xi^{-i}_t(x^{-i}_t) \mu^{*,-i}_t[\hat{a}_{1:t-1}](\xi^{-i}_t) \hat{\gamma}^i_t(a^i_t|\hat{\xi}^i_t) \beta^{*,-i}_t(a^{-i}_t|\hat{a}_{1:t-1}, \xi^{-i}_t) Q^i(\xi^i_{t+1}|\hat{\xi}^i_t, a_t) \tag{5.64c}$$
$$= \sum_{\xi^{-i}_t, a_t, \xi_{t+1}} \left\{ R^i(x_t, a_t) + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), a_t), \xi^i_{t+1}) \right\} \times$$
$$\hat{\xi}^i_t(x^i_t) \xi^{-i}_t(x^{-i}_t) \mu^{*,-i}_t[\hat{a}_{1:t-1}](\xi^{-i}_t) \hat{\beta}^i_t(a^i_t|\hat{a}_{1:t-1}, \hat{w}^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t|\hat{a}_{1:t-1}, \xi^{-i}_t) Q^i(\xi^i_{t+1}|\hat{\xi}^i_t, a_t) \tag{5.64d}$$
$$= \mathbb{E}^{\hat{\beta}^i_t \beta^{*,-i}_t, \mu^*_t[\hat{a}_{1:t-1}]} \left\{ R^i(X_t, A_t) \right. \tag{5.64e}$$
$$\left. + V^i_{t+1}(F(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \beta^*_t(\cdot|\hat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \middle| \hat{a}_{1:t-1}, \hat{w}^i_{1:t} \right\} \tag{5.64f}$$
$$> V^i_t(\underline{\mu}^*_t[\hat{a}_{1:t-1}], \hat{\xi}^i_t) \tag{5.64g}$$

where (5.64a) follows from the definition of V_t^i in (5.21), (5.64d) follows from definition of $\hat{\gamma}_t^i$ and (5.64g) follows from (5.63). However this leads to a contradiction.

Lemma 5.7. $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, w_{1:t+1}^i) \in \mathcal{H}_{t+1}^i$ and β_t^i

$$\begin{aligned}
&\mathbb{E}^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, w_{1:t+1}^i \right\} \\
&= \mathbb{E}^{\beta_{t+1:T}^i \beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i(X_n, A_n) \middle| a_{1:t}, w_{1:t+1}^i \right\}. \qquad (5.65)
\end{aligned}$$

Thus the above quantities do not depend on β_t^i.

Proof. Essentially this claim stands on the fact that $\mu_{t+1}^{*,-i}[a_{1:t}]$ can be updated from $\mu_t^{*,-i}[a_{1:t-1}], \beta_t^{*,-i}$ and a_t, as $\mu_{t+1}^{*,-i}[a_{1:t}] = \prod_{j \neq i} \bar{F}(\mu_t^{*,-i}[a_{1:t-1}], \beta_t^{*,-i}, a_t)$ as in Lemma 5.5. Since the above expectations involve random variables $X_{t+1:T}, A_{t+1:T},$, we consider $P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(x_{t+1:T}, a_{t+1:T} \big| a_{1:t}, w_{1:t+1}^i)$.

$$\begin{aligned}
&P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(x_{t+1:T}, a_{t+1:T} \big| a_{1:t}, w_{1:t+1}^i) \\
&= \frac{P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(a_t, x_{t+1}, w_{t+1}^i, a_{t+1:T}, x_{t+2:T} \big| a_{1:t-1}, w_{1:t}^i)}{P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(a_t, w_{t+1}^i \big| a_{1:t-1}, w_{1:t}^i)} \qquad (5.66a) \\
&= \frac{Nr_1}{Dr_1} \qquad (5.66b)
\end{aligned}$$

We consider the numerator and the denominator separately. The numerator in (5.66a) is given by

$$\begin{aligned}
Nr_1 = &\sum_{x_t, \xi_t^{-i}} P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(x_t, \xi_t^{-i} \big| a_{1:t-1}, w_{1:t}^i) \beta_t^i(a_t^i | a_{1:t-1}, w_{1:t}^i) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \xi_t^{-i}) \\
&Q_x(x_{t+1} | x_t, a_t) Q_w^i(w_{t+1}^i | x_{t+1}^i, a_t) P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*[a_{1:t-1}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, w_{1:t+1}^i, x_{t:t+1}) \qquad (5.66c) \\
= &\sum_{x_t, \xi_t^{-i}} \xi_t(x_t) \mu_t^{*,-i}[a_{1:t-1}](\xi_t^{-i}) \beta_t^i(a_t^i | a_{1:t-1}, w_{1:t}^i) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \xi_t^{-i}) Q_x(x_{t+1} | x_t, a_t) \\
&Q_w^i(w_{t+1}^i | x_{t+1}^i, a_t) P^{\beta_{t+1:T}^i \beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, w_{1:t+1}^i, x_{t+1}) \qquad (5.66d)
\end{aligned}$$

where (5.66d) follows from the fact that probability on $(a_{t+1:T}, x_{2+t:T})$ given $a_{1:t}, w_{1:t+1}^i$, $x_{t:t+1}$, $\mu_t^*[a_{1:t-1}]$ depends on $a_{1:t}, w_{1:t+1}^i$, x_{t+1}, $\mu_{t+1}^*[a_{1:t}]$ through $\beta_{t+1:T}^i \beta_{t+1:T}^{*,-i}$. Similarly, the denominator in (5.66a) is given by

$$Dr_1 = \sum_{\tilde{x}_t, \tilde{\xi}_t^{-i}, \tilde{x}_{t+1}^i} P^{\beta_{t:T}^i \beta_{t:T}^{*,-i}, \mu_t^*}(\tilde{x}_t, \xi_t^{-i} | a_{1:t-1}, w_{1:t}^i) \beta_t^i(a_t^i | a_{1:t-1}, w_{1:t}^i) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \tilde{\xi}_t^{-i})$$
$$Q_x^i(\tilde{x}_{t+1}^i | \tilde{x}_t^i, a_t) Q_w^i(w_{t+1}^i | \tilde{x}_{t+1}^i, a_t) \tag{5.66e}$$
$$= \sum_{\tilde{x}_t, \tilde{\xi}_t^{-i}, \tilde{x}_{t+1}^i} \xi_t^i(\tilde{x}_t^i) \tilde{\xi}_t^{-i}(\tilde{x}_t^{-i}) \mu_t^{*,-i}[a_{1:t-1}](\tilde{\xi}_t^{-i}) \beta_t^i(a_t^i | a_{1:t-1}, w_{1:t}^i) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \tilde{\xi}_t^{-i})$$
$$Q_x^i(\tilde{x}_{t+1}^i | \tilde{x}_t^i, a_t) Q_w^i(w_{t+1}^i | \tilde{x}_{t+1}^i, a_t) \tag{5.66f}$$

By canceling the terms $\beta_t^i(\cdot)$ in the numerator and the denominator, (5.66a) is given by

$$\frac{Nr_2}{Dr_2} \times P^{\beta_{t+1:T}^i \beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, w_{1:t+1}^i, x_{t+1}) \tag{5.66g}$$

where

$$Nr_2 = \sum_{x_t, \xi_t^{-i}} \xi_t(x_t) \mu_t^{*,-i}[a_{1:t-1}](\xi_t^{-i}) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \xi_t^{-i}) Q_x(x_{t+1} | x_t, a_t) Q_w^i(w_{t+1}^i | x_{t+1}^i, a_t) \tag{5.66h}$$

$$Dr_2 = \sum_{\tilde{x}_t, \tilde{\xi}_t^{-i}, \tilde{x}_{t+1}^i} \xi_t^i(\tilde{x}_t^i) \tilde{\xi}_t^{-i}(\tilde{x}_t^{-i}) \mu_t^{*,-i}[a_{1:t-1}](\tilde{\xi}_t^{-i}) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \tilde{\xi}_t^{-i}) Q_x^i(\tilde{x}_{t+1}^i | \tilde{x}_t^i, a_t)$$
$$Q_w(w_{t+1}^i | \tilde{x}_{t+1}^i, a_t) \tag{5.66i}$$

which can be written as

$$\frac{Nr_3}{Dr_3} \times P^{\beta_{t+1:T}^i \beta_{t+1:T}^{*,-i}, \mu_{t+1}^*[a_{1:t}]}(a_{t+1:T}, x_{t+2:T} | a_{1:t}, w_{1:t+1}^i, x_{t+1}) \tag{5.66j}$$

where

$$Nr_3 = \sum_{x_t^i} \xi_t^i(x_t^i) Q_x^i(x_{t+1}^i | x_t^i, a_t^i) Q_w^i(w_{t+1}^i | x_{t+1}^i, a_t) \times$$
$$\sum_{x_t^{-i}, \xi_t^{-i}} \xi_t^{-i}(x_t^{-i}) \mu_t^{*,-i}[a_{1:t-1}](\xi_t^{-i}) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \xi_t^{-i}) Q_x^{-i}(x_{t+1}^{-i} | x_t^{-i}, a_t)$$
$$Dr_3 = \sum_{\tilde{x}_t^i, \tilde{x}_{t+1}^i} \xi_t^i(\tilde{x}_t^i) Q_x^i(\tilde{x}_{t+1}^i | \tilde{x}_t^i, a_t) Q_w(w_{t+1}^i | \tilde{x}_{t+1}^i, a_t) \times$$
$$\sum_{\tilde{x}_t^{-i}, \tilde{\xi}_t^{-i}} \tilde{\xi}_t^{-i}(\tilde{x}_t^{-i}) \mu_t^{*,-i}[a_{1:t-1}](\tilde{\xi}_t^{-i}) \beta_t^{*,-i}(a_t^{-i} | a_{1:t-1}, \tilde{\xi}_t^{-i}) \tag{5.66k}$$

which is equal to

$$= \xi_{t+1}(x_{t+1})\mu^{*,-i}_{t+1}[a_{1:t}](\xi^{-i}_{t+1})P^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(a_{t+1:T}, x_{t+2:T}|a_{1:t}, w^i_{1:t}, x_{t+1}) \quad (5.66l)$$

$$= P^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(x_{t+1}, a_{t+1:T}, x_{t+2:T}|a_{1:t}, w^i_{1:t+1}), \quad (5.66m)$$

Lemma 5.8. $\forall i \in \mathcal{N}, t \in \mathcal{T}, a_{1:t-1} \in \mathcal{H}^c_t, w^i_{1:t} \in (\mathcal{W}^i)^t$

$$V^i_t(\underline{\mu}^*_t[a_{1:t-1}], \xi^i_t) = \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, w^i_{1:t}\right\}. \quad (5.67)$$

Proof. We prove the Lemma by induction. For $t = T$,

$$\mathbb{E}^{\beta^{*,i}_T\beta^{*,-i}_T,\mu^*_T[a_{1:T-1}]}\left\{R^i(X_T, A_T)\big|a_{1:T-1}, w^i_{1:T}\right\}$$
$$= \sum_{x^{-i}_T a_T} R^i(x_T, a_T)\xi_T(x_T)\mu^*_T[a_{1:T-1}](\xi^{-i}_T)\beta^{*,i}_T(a^i_T|a_{1:T-1}, \xi^i_T)\beta^{*,-i}_T(a^{-i}_T|a_{1:T-1}, \xi^{-i}_T) \quad (5.68a)$$

$$= V^i_T(\underline{\mu}^*_T[a_{1:T-1}], \xi^i_T), \quad (5.68b)$$

where (5.68b) follows from the definition of V^i_t in (5.21) and the definition of β^*_T in the forward recursion in (5.23a).

Suppose the claim is true for $t+1$, i.e., $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, w^i_{1:t+1}) \in \mathcal{H}^i_{t+1}$

$$V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t}], \xi^i_{t+1}) = \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t}, w^i_{1:t+1}\right\}. \quad (5.69)$$

Then $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, w^i_{1:t}) \in \mathcal{H}^i_t$, we have

$$\begin{aligned}
&\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, w^i_{1:t}\right\} \\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{R^i(X_t, A_t)+\right. \\
&\qquad \left.\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, A_t, w^i_{1:t}, W^i_{t+1}\right\}\big|a_{1:t-1}, w^i_{1:t}\right\} \qquad (5.70a)\\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{R^i(X_t, A_t)+\right. \\
&\qquad \left.\mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i(X_n, A_n)\big|a_{1:t-1}, A_t, w^i_{1:t}, W^i_{t+1}\right\}\big|a_{1:t-1}, w^i_{1:t}\right\} \qquad (5.70b)\\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{R^i(X_t, A_t) + V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], \Xi^i_{t+1})\big|a_{1:t-1}, w^i_{1:t}\right\} \qquad (5.70c)\\
&= \mathbb{E}^{\beta^{*,i}_{t}\beta^{*,-i}_{t},\mu^*_t[a_{1:t-1}]}\left\{R^i(X_t, A_t) + V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], \Xi^i_{t+1})\big|a_{1:t-1}, w^i_{1:t}\right\} \qquad (5.70d)\\
&= V^i_t(\underline{\mu}^*_t[a_{1:t-1}], \xi^i_t), \qquad (5.70e)
\end{aligned}$$

where (5.70b) follows from Lemma 5.7 in Appendix D, (5.70c) follows from the induction hypothesis in (5.69), (5.70d) follows because the random variables involved in expectation, X^{-i}_t, A_t, X^i_{t+1} do not depend on $\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T}$ and (5.70e) follows from the definition of β^*_t in the forward recursion in (5.23a), the definition of μ^*_{t+1} in (5.23b) and the definition of V^i_t in (5.21).

5.9 Appendix E (Proof of Lemma 5.3)

Proof. We will prove the result by induction on t. The result is vacuously true for $T+1$. Suppose it is also true for $t+1$, i.e.

$$(\mu^*_{t+1})^{-1}(\tilde{\mathcal{C}}^{a_{t+1:T}}_{t+1}) = \mathcal{C}^{a_{t+1:T}}_{t+1}. \qquad (5.71)$$

We show that the result holds true for t. In the following two cases, we show that if there exists an element in one set, it also belongs to the other. From the contrapositive of the statement, if one is empty, so is the other.

Case 1. We prove $(\mu^*_t)^{-1}(\tilde{\mathcal{C}}^{a_{t:T}}_t) \subset \mathcal{C}^{a_{t:T}}_t$

Let $h^c_t \in (\mu^*_t)^{-1}(\tilde{\mathcal{C}}^{a_{t:T}}_t)$. We will show that $h^c_t \in \mathcal{C}^{a_{t:T}}_t$.

Since $h_t^c \in (\mu_t^*)^{-1}(\tilde{\mathcal{C}}_t^{a_{t:T}})$, this implies $\mu_t^*[h_t^c] \in \tilde{\mathcal{C}}_t^{a_{t:T}}$. Then by the definition of $\tilde{\mathcal{C}}_t^{a_{t:T}}, \forall i,\ \forall \xi_t^i \in supp(\mu_t^{*,i}[h_t^c]),\ \theta_t^i[\mu_t^*[h_t^c]](a_t^i|\xi_t^i) = 1$. Since $\xi_t^i(x_t^i) = P(x_t^i|h_t^i)\ \forall x_t^i$, $\mu_t^{*,i}[h_t^c](\xi_t^i) = P^\theta(\xi_t^i|h_t^c)\ \forall \xi_t^i$ and $\beta_t^{*,i}(a_t^i|h_t^i) = \theta_t^i[\mu_t^*[h_t^c]](a_t^i|\xi_t^i)$ by the definition of β^*, this implies $\forall i, \beta_t^{*,i}(a_t^i|h_t^i) = 1,\ \forall h_t^i$ that are consistent with h_t^c and occur with non-zero probability.

Also since $\mu_t^*[h_t^c] \in \tilde{\mathcal{C}}_t^{a_{t:T}}$, this implies $\bar{F}([\mu_t^*[h_t^c], \theta_t[\mu_t^*[h_t^c]], a_t) \in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}}$ by definition of $\tilde{\mathcal{C}}_t^{a_{t:T}}$. Thus $\mu_{t+1}^*[h_t^c, a_t] \in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}}$, since $\mu_{t+1}^*[h_t^c, a_t] =$
$\bar{F}([\mu_t^*[h_t^c], \theta_t[\mu_t^*[h_t^c]], a_t)$ by definition. Using the induction hypothesis, $(h_t^c, a_t) \in \mathcal{C}_{t+1}^{a_{t+1:T}}$, which implies $\forall i, \beta_n^{*,i}(a_n^i|h_n^i) = 1,\ \forall n \geq t+1, \forall h_n^i$ that are consistent with (h_t^c, a_t) and occur with non-zero probability.

The above two facts conclude that $\forall i, \beta_n^{*,i}(a_n^i|h_n^i) = 1,\ \forall n \geq t, \forall h_n^i$ that are consistent with h_t^c and occur with non-zero probability, which implies $h_t^c \in \mathcal{C}_t^{a_{t:T}}$ by the definition of $\mathcal{C}_t^{a_{t:T}}$.

Case 2. We prove $(\mu_t^*)^{-1}(\tilde{\mathcal{C}}_t^{a_{t:T}}) \supset \mathcal{C}_t^{a_{t:T}}$.

Let $h_t^c \in \mathcal{C}_t^{a_{t:T}}$. We will show that $\mu_t^*[h_t^c] \in \tilde{\mathcal{C}}_t^{a_{t:T}}$.

Since $h_t^c \in \mathcal{C}_t^{a_{t:T}}$, this implies $\forall i, \beta_t^{*,i}(a_t^i|h_t^i) = 1, \forall h_t^i$ that are consistent with h_t^c and occur with non-zero probability. Since $\beta_t^{*,i}(a_t^i|h_t^i) = \theta_t^i[\mu_t^*[h_t^c]](a_t^i|\xi_t^i)$, by the definition of β^*, where $\xi_t^i(x_t^i) = P(x_t^i|h_t^i)\ \forall x_t^i$, this implies $\forall i, \theta_t^i[\mu_t^*[h_t^c]](a_t^i|\xi_t^i) = 1, \forall \xi_t^i \in supp(\mu_t^{*,i}[h_t^c])$, where $\mu_t^{*,i}[h_t^c](\xi_t^i) = P^\theta(\xi_t^i|h_t^c)\ \forall \xi_t^i$.

Also, since $h_t^c \in \mathcal{C}_t^{a_{t:T}}$, it is implied by the definition of $\mathcal{C}_t^{a_{t:T}}$ that $(h_t^c, a_t) \in \mathcal{C}_{t+1}^{a_{t+1:T}}$. This implies $\mu_{t+1}^*[h_t^c, a_t] \in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}}$ by the induction hypothesis. Since, by definition, $\mu_{t+1}^*[h_t^c, a_t] = \bar{F}([\mu_t^*[h_t^c], \theta_t[\mu_t^*[h_t^c]], a_t)$, this implies $\bar{F}([\mu_t^*[h_t^c], \theta_t[\mu_t^*[h_t^c]], a_t)$ $\in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}}$.

Since we have shown that $\forall i, \theta_t^i[\mu_t^*[h_t^c]](a_t^i|\xi_t^i) = 1, \forall \xi_t^i \in supp(\mu_t^*[h_t^c])$ and $\bar{F}([\mu_t^*[h_t^c], \theta_t[\mu_t^*[h_t^c]], a_t) \in \tilde{\mathcal{C}}_{t+1}^{a_{t+1:T}}$, this implies $\mu_t^*[h_t^c] \in \tilde{\mathcal{C}}_t^{a_{t:T}}$ by the definition of $\tilde{\mathcal{C}}_t^{a_{t:T}}$.

The above two cases complete the induction step.

5.10 Appendix F (Proof of Theorem 5.2)

Proof. We prove this by induction on t_0. For $t_0 = T$, (5.35) reduces to

$$\tilde{\gamma}_T^i(\cdot|\xi_T^i) \in \arg\max_{\gamma_T^i(\cdot|\xi_T^i)} \sum_{a_T^i} a_T^i \gamma_T^i(a_T^i|\xi_T^i)(\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{\xi}_T^{-i} - 1)), \tag{5.72}$$

and since $\pi_T \in \hat{\mathcal{C}}^a$, it is easy to verify that $\tilde{\gamma}_T^i(a^i|\xi_T^i) = 1,\ \forall \xi_T^i \in [0,1]$ and thus $V_T^i(\pi_T^{-i}, \xi_T^i)$ $= (\lambda(2\xi_T^i - 1) + \bar{\lambda}(2\hat{\xi}_T^{-i} - 1))a^i$. This establishes the base case.

Now, suppose the claim is true for $t_0 = \tau + 1$ i.e. if $\pi_{\tau+1} \in \hat{\mathcal{C}}^a$, then $\forall t \geq \tau+1, \pi_t \in \hat{\mathcal{C}}^a$ and $\tilde{\gamma}_t^i(a^i|\xi_t^i) = 1\ \forall \xi_t^i \in [0,1]$. Moreover, for $\tau + 1 \leq t \leq T$, V_t^i is given by

$$V_t^i(\pi_t^{-i}, \xi_t^i) = (T - t + 1)(\lambda(2\xi_t^i - 1) + \bar{\lambda}(2\hat{\xi}_t^{-i} - 1))a^i \quad \forall \pi_t \in \hat{\mathcal{C}}^a. \tag{5.73}$$

Then if $\pi_\tau \in \hat{\mathcal{C}}^a$, then $\tilde{\gamma}_\tau^i(a^i|\xi_\tau^i) = 1\ \forall \xi_\tau^i \in [0,1]$ satisfies (5.35) since,

$$\begin{aligned}
\tilde{\gamma}_\tau^i(\cdot|\xi_\tau^i) \in \arg\max_{\gamma_\tau^i(\cdot|\xi_\tau^i)} & \sum_{a_\tau^i} a_\tau^i \gamma_\tau^i(a_\tau^i|\xi_\tau^i)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1)) \\
& + \mathbb{E}^{\gamma_\tau^i(\cdot|\xi_\tau^i)\tilde{\gamma}_\tau^{-i}, \pi_\tau}\left\{V_{\tau+1}^i(F(\underline{\pi}_\tau^{-i}, \tilde{\gamma}_\tau^{-i}, A_\tau^{-i}), \Xi_{\tau+1}^i)\middle|\xi_\tau^i\right\} && (5.74) \\
\in \arg\max_{\gamma_\tau^i(\cdot|\xi_\tau^i)} & \sum_{a_\tau^i} a_\tau^i \gamma_\tau^i(a_\tau^i|\xi_\tau^i)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1)) \\
& + \mathbb{E}^{\gamma_\tau^i(\cdot|\xi_\tau^i)\tilde{\gamma}_\tau^{-i}, \pi_\tau}\left\{(T - \tau)(\lambda(2\Xi_{\tau+1}^i - 1) + \bar{\lambda}(2\hat{\Xi}_{\tau+1}^{-i} - 1))a^i|\xi_\tau^i\right\} && (5.75) \\
\in \arg\max_{\gamma_\tau^i(\cdot|\xi_\tau^i)} & \sum_{a_\tau^i} a_\tau^i \gamma_\tau^i(a_\tau^i|\xi_\tau^i)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1)) \\
& + (T - \tau)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1))a^i && (5.76) \\
\in \arg\max_{\gamma_\tau^i(\cdot|\xi_\tau^i)} & \sum_{a_\tau^i} a_\tau^i \gamma_\tau^i(a_\tau^i|\xi_\tau^i)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1)), && (5.77)
\end{aligned}$$

where (5.75) follows from the fact that $F(\underline{\pi}_\tau, \tilde{\gamma}_\tau, a_\tau) \in C^a\ \forall a_\tau$, as shown in Claim 5.4, and induction hypothesis, (5.76) follows from Claim 5.4 and Claim 5.5 and (5.77) follows from the fact that the second term does not depend on $\gamma_\tau^i(\cdot|\xi_\tau^i)$. This also shows that

$$V_\tau^i(\pi_\tau^{-i}, \xi_\tau^i) = (T - \tau + 1)(\lambda(2\xi_\tau^i - 1) + \bar{\lambda}(2\hat{\xi}_\tau^{-i} - 1))a^i, \tag{5.78}$$

which completes the induction step.

Claim 5.4. Expectation of π_{t+1}^i under non-informative $\tilde{\gamma}_t^i$ of the form $\tilde{\gamma}_t^i(a^i|\xi_t^i) = 1\ \forall \xi_t^i \in [0,1]$, remains the same as expectation of π_t^i, i.e.,

$$\mathbb{E}\{\Xi_{t+1}^i(1)|\pi_t^i, \tilde{\gamma}_t^i, a^i\} = \mathbb{E}\{\Xi_t^i(1)|\pi_t^i\} \tag{5.79}$$

Proof.

$$\mathbb{E}\{\Xi^i_{t+1}(1)|\pi^i_t,\tilde{\gamma}^i_t,a^i\}$$

$$= \sum_{\xi^i_{t+1}(1)} \xi^i_{t+1}(1)\bar{F}^i(\pi^i_t,\tilde{\gamma}^i_t,a^i)(\xi^i_{t+1}(1)) \tag{5.80}$$

$$= \frac{\sum_{\xi^i_t,x^i,\xi^i_{t+1}(1)} \xi^i_{t+1}(1)\pi^i_t(\xi^i_t)\xi^i_t(x^i)\tilde{\gamma}^i_t(a^i_t|\xi^i_t)Q^i_w(w^i_{t+1}|x^i,a_t)I_{F^i(\xi^i_t,w^i_{t+1},a_t)(1)}(\xi^i_{t+1}(1))}{\sum_{\xi^i_t,x^i,w^i_{t+1}} \pi^i_t(\xi^i_t)\xi^i_t(x^i)\tilde{\gamma}^i_t(a^i_t|\xi^i_t)} \tag{5.81}$$

$$= \frac{\sum_{\xi^i_t,x^i,w^i_{t+1},\xi^i_{t+1}(1)} \xi^i_{t+1}(1)\pi^i_t(\xi^i_t)\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i)I_{F^i(\xi^i_t,w^i_{t+1},a^i)(1)}(\xi^i_{t+1}(1))}{\sum_{\xi^i_t,x^i} \pi^i_t(\xi^i_t)\xi^i_t(x^i)} \tag{5.82}$$

$$= \sum_{\xi^i_t,x^i,w^i_{t+1}} F^i(\xi^i_t,w^i_{t+1},a^i)(1)\pi^i_t(\xi^i_t)\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i) \tag{5.83}$$

$$= \sum_{\xi^i_t,w^i_{t+1}} \frac{\xi^i_t(1)Q^i_w(w^i_{t+1}|1,a^i)}{\sum_{\tilde{x}^i}\xi^i_t(\tilde{x}^i)Q^i_w(w^i_{t+1}|\tilde{x}^i,a^i)}\pi^i_t(\xi^i_t)\sum_{x^i}\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i) \tag{5.84}$$

$$= \sum_{\xi^i_t} \xi^i_t(1)\pi^i_t(\xi^i_t(1)) \tag{5.85}$$

$$= \mathbb{E}\{\Xi^i_t(1)|\pi^i_t\} \tag{5.86}$$

Claim 5.5. For any γ^i_t,

$$\mathbb{E}\{\Xi^i_{t+1}(1)|\xi^i_t,\gamma^i_t\} = \xi^i_t(1) \tag{5.87}$$

Proof.

$$\mathbb{E}\{\Xi^i_{t+1}(1)|\xi^i_t,\gamma^i_t\}$$

$$= \sum_{x^i,w^i_{t+1},a^i_t,\xi^i_{t+1}(1)} \xi^i_{t+1}(1)I_{\bar{F}^i(\xi^i_t,w^i_{t+1},a^i_t)(1)}(\xi^i_{t+1}(1))\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i_t)\gamma^i_t(a^i_t|\xi^i_t) \tag{5.88}$$

$$= \sum_{x^i,w^i_{t+1},a^i_t} \bar{F}^i(\xi^i_t,w^i_{t+1},a^i_t)(1)\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i_t)\gamma^i_t(a^i_t|\xi^i_t) \tag{5.89}$$

$$= \sum_{a^i_t,w^i_{t+1}} \frac{\xi^i_t(1)Q^i_w(w^i_{t+1}|1,a^i_t)}{\sum_{\tilde{x}^i}\xi^i_t(\tilde{x}^i)Q^i_w(w^i_{t+1}|\tilde{x}^i,a^i_t)}\gamma^i_t(a^i_t|\xi^i_t)\sum_{x^i}\xi^i_t(x^i)Q^i_w(w^i_{t+1}|x^i,a^i_t) \tag{5.90}$$

$$= \sum_{a^i_t,w^i_{t+1}} \xi^i_t(1)Q^i_w(w^i_{t+1}|1,a^i_t)\gamma^i_t(a^i_t|\xi^i_t) \tag{5.91}$$

$$= \xi^i_t(1) \tag{5.92}$$

BIBLIOGRAPHY

[1] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[2] D. Avis and K. Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. *Discrete and Computational Geometry*, 8(1):295–313, 1992.

[3] T. Başar. Two-criteria LQG decision problems with one-step delay observation sharing pattern. *Information and Control*, 38(1):21 – 50, 1978.

[4] T. Başar and G. J. Olsder. *Dynamic noncooperative game theory*. Academic Press, 1982.

[5] D. Bergemann and J. Valimaki. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, mar 2010.

[6] D. Bertsekas. *Dynamic Programming and Stochastic Control*. Academic Press, 1976.

[7] F. Beutler and D. Teneketzis. Routing in queueing networks under imperfect information: stochastic dominance and thresholds. *Queueing Systems*, 26(2):291–314, Mar 1993.

[8] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):pp. 992–1026, 1992.

[9] T. Börgers. *An Introduction to the Theory of Mechanism Design*. 2013.

[10] A. Chakrabarti, A. Sabharwal, and B. Aazhang. *Cooperative Wireless Communications: Fundamental Techniques and Enabling Technologies*, chapter Cooperation in Wireless Networks: Principles and Applications. Springer-Verlag New York, Inc., 2007.

[11] E. H. Clarke. Multipart pricing of public goods. *Public choice*, 11(1):17–33, 1971.

[12] T. M. Cover and A. A. El Gamal. Capacity theorems for the relay channel. *IEEE Trans. Information Theory*, 25(5):572–584, Sep 1979.

[13] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, Cambridge, MA, 1991.

[14] T. Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

[15] T. Groves and J. Ledyard. Optimal allocation of public goods: A solution to the "free rider" problem. *Econometrica: Journal of the Econometric Society*, pages 783–809, 1977.

[16] A. Gupta, A. Nayyar, C. Langbort, and T. Başar. Common information based Markov perfect equilibria for linear-gaussian games with asymmetric information. *SIAM Journal on Control and Optimization*, 52(5):3228–3260, 2014.

[17] B. Hajek. Optimal control of two interacting service stations. *IEEE Trans. Automatic Control*, (6):491ñ–499, June 1984.

[18] Y.-C. Ho. Team decision theory and information structures. *Proceedings of the IEEE*, 68(6):644–654, 1980.

[19] Y. C. Ho and K.-H. Chu. Team decision theory and information structures in optimal control problems–part i. *Automatic Control, IEEE Transactions on*, 17(1):15–22, 1972.

[20] J. Huang, Z. Han, M. Chiang, and H. Poor. Auction-based resource allocation for cooperative communications. *IEEE J. Select. Areas Commun.*, 26(7):1226–1237, 2008.

[21] L. Hurwicz. Outcome functions yielding Walrasian and Lindahl allocations at Nash equilibrium points. *The Review of Economic Studies*, 46(2):217–225, 1979.

[22] O. Ileri, S.-C. Mau, and N. B. Mandayam. Pricing for enabling forwarding in self-configuring ad hoc networks. *IEEE J. Select. Areas Commun.*, 23(1):151–162, 2005.

[23] M. O. Jackson. A crash course in implementation theory. *Social choice and welfare*, 18(4):655–708, 2001.

[24] M. O. Jackson and H. F. Sonnenschein. Overcoming incentive constraints by linking decisions. *Econometrica*, 75(1):241–257, Jan 2007.

[25] K. V. Jerzy Filar. *Competitive Markov Decision Processes*. Springer-Verlag New York, Inc., New York, NY, USA, 1996.

[26] A. Kakhbod and D. Teneketzis. Power allocation and spectrum sharing in multi-user, multi-channel systems with strategic users. In *Proc. IEEE Conf. on Decision and Control*, pages 1088–1095. IEEE, 2010.

[27] A. Kakhbod and D. Teneketzis. An efficient game form for multi-rate multicast service provisioning. *IEEE J. Select. Areas Commun.*, 30(11):2093–2104, 2012.

[28] A. Kakhbod and D. Teneketzis. An efficient game form for unicast service provisioning. *IEEE Transactions on Automatic Control*, 57(2):392–404, 2012.

[29] C. Kamhoua, N. Pissinou, J. Miller, and S. Makki. Mitigating routing misbehavior in multi-hop networks using evolutionary game theory. In *GLOBECOM Workshops (GC Wkshps), 2010 IEEE*, pages 1957–1962, 2010.

[30] D. M. Kreps and J. Sobel. Chapter 25 signalling. volume 2 of *Handbook of Game Theory with Economic Applications*, pages 849 – 867. Elsevier, 1994.

[31] P. R. Kumar and P. Varaiya. *Stochastic systems: estimation, identification, and adaptive control*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[32] T. N. Le, V. Subramanian, and R. Berry. The impact of observation and action errors on informational cascades. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 1917–1922, Dec 2014.

[33] A. Mahajan. Optimal decentralized control of coupled subsystems with control sharing. *Automatic Control, IEEE Transactions on*, 58(9):2377–2382, 2013.

[34] A. Mahajan and A. Nayyar. Sufficient statistics for linear control strategies in decentralized systems with partial history sharing. *IEEE Transactions on Automatic Control*, 60(8):2046–2056, Aug 2015.

[35] A. Mahajan and D. Teneketzis. On the design of globally optimal communication strategies for real-time communcation systems with noisy feedback. *IEEE J. Select. Areas Commun.*, (4):580–595, May 2008.

[36] P. Marbach and R. Berry. Downlink resource allocation and pricing for wireless networks. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1470–1479 vol.3, 2002.

[37] S. Marti, T. J. Giuli, K. Lai, and M. Baker. Mitigating routing misbehavior in mobile ad hoc networks. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, MobiCom '00, pages 255–265, New York, NY, USA, 2000. ACM.

[38] E. Maskin and J. Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.

[39] F. Meshkati, H. Poor, and S. Schwartz. Energy-efficient resource allocation in wireless networks. *IEEE Signal Processing Magazine*, 24(3):58–68, 2007.

[40] J. Nash. Non-cooperative games. *Annals of mathematics*, pages 286–295, 1951.

[41] A. Nayyar, A. Gupta, C. Langbort, and T. Başar. Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games. *IEEE Trans. Automatic Control*, 59(3):555–570, March 2014.

[42] A. Nayyar, A. Mahajan, and D. Teneketzis. Optimal control strategies in delayed sharing information structures. *IEEE Trans. Automatic Control*, 56(7):1606–1620, July 2011.

[43] A. Nayyar, A. Mahajan, and D. Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *Automatic Control, IEEE Transactions on*, 58(7):1644–1658, 2013.

[44] A. Nayyar and D. Teneketzis. On globally optimal real-time encoding and decoding strategies in multi-terminal communication systems. In *Proc. IEEE Conf. on Decision and Control*, pages 1620–1627, Cancun, Mexico, Dec. 2008.

[45] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*, volume 1 of *MIT Press Books*. The MIT Press, 1994.

[46] Y. Ouyang, H. Tavafoghi, and D. Teneketzis. Dynamic oligopoly games with private Markovian dynamics. In *Proc. 54th IEEE Conf. Decision and Control (CDC)*, 2015.

[47] R. M. Pietro Michiardi. Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In *Advanced Communications and Multimedia Security*, pages 107–121, 2001.

[48] Y. Qiu and P. Marbach. Bandwidth allocation in ad hoc networks: a price-based approach. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 2, pages 797–807 vol.2, 2003.

[49] Y. E. Sagduyu and A. Ephremides. A game-theoretic look at simple relay channel. *ACM/Kluwer Journal of Wireless Networks*, (5):545–560, Oct 2006.

[50] S. Sharma and D. Teneketzis. Local public good provisioning in networks: A Nash implementation mechanism. *IEEE J. Select. Areas Commun.*, 30(11):2105–2116, 2012.

[51] J. E. Smith and K. F. McCardle. Structural properties of stochastic dynamic programs. *Oper. Res.*, 50(5):796–809, Sept. 2002.

[52] L. Smith and P. Sörensen. Pathological outcomes of observational learning. *Econometrica*, 68(2):371–398, 2000.

[53] V. Srinivasan, P. Nuggehalli, C. F. Chiasserini, and R. Rao. Cooperation in wireless ad hoc networks. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 2, pages 808–817 vol.2, 2003.

[54] E. C. van der Meulen. Three-terminal communication channels. *Adv. Appl. Prob.*, pages 120–154, 1971.

[55] D. Vasal and A. Anastasopoulos. Energy delay tradeoff in cooperative communication. CSPL technical report 406, University of Michigan, Ann Arbor, MI, July 2011. can be downloaded from http://www.eecs.umich.edu/systems/TechReportList.html.

[56] D. Vasal and A. Anastasopoulos. Stochastic control of relay channels with cooperative and strategic users. *Communications, IEEE Transactions on*, 62(10):3434–3446, Oct 2014.

[57] D. Vasal, V. Subramanian, and A. Anastasopoulos. A systematic process for evaluating structured perfect Bayesian equilibria in dynamic games with asymmetric information. Technical report, Aug. 2015.

[58] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[59] H. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

[60] H. S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, 1971.

[61] J. Yang and I. Brown, D.R. Energy efficient relaying games in cooperative wireless transmission systems. In *Signals, Systems and Computers, 2007. ACSSC 2007. Conference Record of the Forty-First Asilomar Conference on*, pages 835–839, 2007.

[62] S. Yüksel. Stochastic nestedness and the belief sharing information pattern. *Automatic Control, IEEE Transactions on*, 54(12):2773–2786, 2009.