A systematic process for evaluating structured perfect Bayesian equilibria in dynamic games with asymmetric information

Deepanshu Vasal, Abhinav Sinha and Achilleas Anastasopoulos

Abstract

We consider both finite-horizon and infinite-horizon versions of a dynamic game with N selfish players who observe their types privately and take actions that 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. In dynamic games with asymmetric information a widely used concept of equilibrium is perfect Bayesian equilibrium (PBE), which consists of a strategy and belief pair that simultaneously satisfy sequential rationality and belief consistency. To date there does not exist a universal algorithm that decouples the interdependence of strategies and beliefs over time in calculating PBE. For the finite-horizon game we develop a two-step backward-forward recursive algorithm that sequentially decomposes the problem (w.r.t. time) to obtain a subset of PBEs, which we refer to as *structured Bayesian perfect equilibria* (SPBE). In such equilibria, an agent's strategy depends on its history only through a common public belief and its current private type. 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. We then extend this methodology to the infinite-horizon model, where we propose a time-invariant single-shot fixed-point equation, which in conjunction with a forward recursive step, generates the SPBE. With our proposed method, we find equilibria that exhibit *signaling* behavior. This is illustrated with the help of a concrete public goods example. Finally, sufficient conditions for the existence of SPBE are provided.

I. Introduction

Several practical applications involve dynamic interaction of strategic decision-makers with private and public observations. Such applications include repeated online advertisement auctions, wireless resource sharing, and energy markets. In repeated online advertisement auctions, advertisers place bids for locations on a website to sell a product. These bids are calculated based on the value of that product, which is privately observed by the advertiser and past actions of other advertisers, which are observed publicly. Each advertiser's goal is to maximize its reward, which for an auction depends on the actions taken by others. In wireless resource sharing, players are allocated channels that interfere with each other. Each player privately observes its channel gain and takes an action, which can be the choice of modulation or coding scheme and also the transmission power. The reward it receives depends on the rate the player gets, which is a function of each player's channel gain and other players' actions (through the signal-to-interference ratio). Finally, in an energy market, different suppliers bid their estimated power outputs to an independent system operator (ISO) that formulates the market mechanism to determine the prices assessed to the different suppliers. Each supplier wants to maximize its return, which depends on its cost of production of energy, which is its private information, and the market-determined prices which depend on all the bids.

Dynamical systems with strategic players are modeled as dynamic stochastic games, introduced by Shapley in [1]. Discrete-time dynamic games with Markovian structure have been studied extensively to

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, 48105 USA e-mail: {dvasal, absi, anastas}@umich.edu.

This work is supported in part by NSF grants CIF-1111061 and ECCS-1608361.

This paper has originally appeared on arXiv.org on August 26, 2015 as working paper 1508.06269 and revised on September 14, 2016 as 1609.04221.

model many practical applications, in engineering as well as economics literature [2], [3]. In dynamic games with perfect and symmetric information, subgame perfect equilibrium (SPE) is an appropriate equilibrium concept and there exists a backward recursive algorithm to find all the SPEs of these games (refer to [4]–[6] for a more elaborate discussion). Maskin and Tirole in [7] introduced the concept of Markov perfect equilibrium (MPE) for dynamic games with symmetric information, where equilibrium strategies are dependent on some payoff relevant Markovian state of the system, rather than on the entire history. This is a refinement of the SPE. Some prominent examples of the application of MPE include [8]–[10]. Ericson and Pakes in [8] model industry dynamics for firms' entry, exit and investment participation, through a dynamic game with symmetric information, compute its MPE, and prove ergodicity of the equilibrium process. Bergemann and Välimäki in [9] study a learning process in a dynamic oligopoly with strategic sellers and a single buyer, allowing for price competition among sellers. They study MPE of the game and its convergence behavior. Acemoğlu and Robinson in [10] develop a theory of political transitions in a country by modeling it as a repeated game between the elites and the poor, and study its MPE.

In dynamic games with asymmetric information, and more generally in multi-agent, dynamic decision problems (cooperative or non-cooperative) with asymmetric information, there is a signaling phenomenon that can occur, where a player's action reveals part of its private information to other players, which in turn affects their future payoff [1] (see [14] for a survey of signaling models). In one of the first works demonstrating signaling, a two-stage dynamic game was considered by Spence [15], where a worker signals her abilities to a potential employer using the level of education as a signal. Since then, this phenomenon has been shown in many settings, e.g., warranty as a signal for better quality of a product, in [16], larger deductible or partial insurance as a signal for better health of a person, in [17], [18], and in evolutionary game theory, extra large antlers by a deer to signal fitness to a potential mate, in [19].

For dynamic games with asymmetric information, where players' observations belong to different information sets, in order to calculate expected future rewards players need to form a belief on the observations of other players (where players need not have consistent beliefs). As a result, SPE or MPE [2], are not appropriate equilibrium concepts for such settings. There are several notions of equilibrium for such games, such as perfect Bayesian equilibrium (PBE), sequential equilibrium, and trembling hand equilibrium [4], [5]. Each of these equilibrium notions consist of an assessment, i.e., a strategy and a belief profile for the entire time horizon. The equilibrium strategies are optimal given the equilibrium beliefs and the equilibrium beliefs are derived from the equilibrium strategy profile using Bayes' rule (whenever possible), with some equilibrium concepts requiring further refinements. Thus there is a cyclical requirement of beliefs being consistent with strategies, which are in turn optimal given the beliefs, and finding such equilibria can be thought of as being equivalent to solving a fixed point equation in the space of strategy and belief profiles over the entire time horizon. Furthermore, these strategies and beliefs are functions of histories and thus their domain grows exponentially in time, which makes the problem computationally intractable. To date, there is no universal algorithm that provides simplification by decomposing the aforementioned fixed-point equation for calculating PBEs.

Some practically motivated work in this category is the work in [20]–[23]. Authors in [20]–[22] study the problem of social learning with sequentially-acting selfish players who act exactly once in the game and make a decision to adopt or reject a trend based on their estimate of the system state. Players observe a private signal about the system state and publicly observe actions of past players. The authors analyze PBE of the dynamic game and study the convergent behavior of the system under an equilibrium, where

[1]There are however instances where even though actions reveal private information, at equilibrium the signaling effect is non-existent [11], [12] and [13, sec III.A]. Thus, MPE is an appropriate equilibrium concept for such games. In [11], authors extend the model of [8] where firms' set-up costs and scrap values are random and their private information. However, these are assumed to be i.i.d. across time and thus the knowledge of this private information in any period does not affect the future reward. In [12], [13, sec III.A], authors discuss games with one-step delayed information pattern, where all players get access to players' private information with delay one. In this case as well, signaling does not occur.

[2]SPE and MPE are used for games where beliefs in the game are strategy-independent and consistent among players. As a result, these beliefs are derived from basic parameters of the problems and are not part of the definition of the equilibrium concept.

they show occurrence of herding. Devanur et al. in [23] study PBE of a repeated sales game where a single buyer has a valuation of a good, which is its private information, and a seller offers to sell a fresh copy of that good in every period through a posted price.

A. Contributions

In this paper, we present a sequential decomposition methodology for calculating a subset of all PBEs for finite and infinite horizon dynamic games with asymmetric information. Our model, consists of strategic agents having types that evolve as independent Markov controlled processes. Players observe their types privately and actions taken by all players are observed publicly. Instantaneous reward for each player depends on everyone's types and actions. The proposed methodology provides a decomposition of the interdependence between beliefs and strategies in PBE and enables a systematic evaluation of a subset of PBE, namely *structured perfect Bayesian equilibria* (SPBE). For SPBE, players' strategies are based on their current private type and a set of beliefs on each player's current private type, which is common to all the players and whose domain is time-invariant. The beliefs on players' types are such that they can be updated individually for each player and sequentially w.r.t. time. The model allows for signaling amongst agents as beliefs depend on strategies.

Our motivation for considering SPBE stems from ideas in decentralized team problems and specifically the works of Ho [24] and Nayyar et al. [25]. We utilize the agent-by-agent approach in [24] to motivate a Markovian structure where players' strategies depend only on their current types. In addition, we utilize the common information based approach introduced in [25] to summarize the common information into a common belief on players' private types. Even though these ideas motivate the special structure of our equilibrium strategies, they can not be applied in games to evaluate SPBE because they have been developed for dynamic teams and are incompatible with equilibrium notions. Our main contribution is a new construction based on which SPBE can be systematically evaluated.

Specifically, for the finite horizon model, we provide a two-step algorithm involving a backward recursion followed by a forward recursion. The algorithm works as follows. In the backward recursion, for every time period, the algorithm finds an equilibrium generating function defined for all possible common beliefs at that time. This involves solving an one-step fixed point equation on the space of probability simplexes. Then, the equilibrium strategies and beliefs are obtained through a forward recursion using the equilibrium generating function obtained in the backward step and the Bayes update rule. The SPBE that are developed in this paper are analogous to MPEs (for games with symmetric information) in the sense that players choose their actions based on beliefs that depend on common information, and private types, both of which have Markovian dynamics.

For the infinite horizon model, instead of the backwards recursion step, the algorithm solves a single-shot time invariant fixed-point equation involving both an equilibrium generating function and an equilibrium reward-to-go function.

We show that using our method, existence of SPBEs in the asymmetric information dynamic game is guaranteed if the aforementioned fixed-point equation admits a solution. We provide sufficient conditions under which this is true.

We demonstrate our methodology of finding SPBEs through a multi-stage public goods game, whereby we observe the aforementioned signaling effect at equilibrium.

B. Relevant Literature

Related literature on this topic include [13], [26] and [27]. Nayyar et al. in [13], [26] consider dynamic games with asymmetric information. There is an underlying controlled Markov process and players jointly observe part of the process and whilst making additional private observations. It is shown that the considered game with asymmetric information, under certain assumptions, can be transformed to another game with symmetric information. A backward recursive algorithm is provided to find MPE of the transformed game. For this strong equivalence to hold, authors in [13], [26] 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 past strategies used by the players. This leads to all strategies being non-signaling. Our model is different from this since we assume that the underlying state of the system has independent components, each constituting a player's private type. However, we do not make any assumption regarding update of beliefs and allow the belief state to depend on players' past strategies, which in turn allows the possibility of signaling in the game.

Ouyang et al. in [27] consider a dynamic oligopoly game with strategic sellers and 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. The common belief is strategy dependent and the authors consider equilibria based on this common information belief. It is shown that if all other players play actions based on the common belief and their private information, then player i faces a Markov decision problem (MDP) with respect to its action with state being the common belief and its private type. Thus calculating equilibrium boils down to solving a fixed point equation on belief update functions and strategies of all players. Existence of such equilibrium is shown for a degenerate case where agents act myopically at equilibrium and the equilibrium itself is non-signaling.

Other than the common information based approach, Li et al. [28] consider a finite horizon zero-sum dynamic game, where at each time only one agent out of the two knows the state of the system. The value of the game is calculated by formulating an appropriate linear program. Cole et al. [29] consider an infinite horizon discounted reward dynamic game where actions are only privately observable. They provide a fixed-point equation for calculating a subset of sequential equilibrium, which is referred to as Markov private equilibrium (MPrE). In MPrE strategies depend on history only through the latest private observation.

C. Notation

We use uppercase letters for random variables and lowercase for their realizations. For any variable, subscripts represent time indices and superscripts represent player identities. We use notation $-i$ to represent all players other than player i i.e. $-i = \{1, 2, \dots i-1, i+1, \dots, N\}$. We use notation $A_{t:t'}$ to represent the vector $(A_t, A_{t+1}, \dots A_{t'})$ when $t' \geq t$ or an empty vector if $t' < t$. We use A_t^{-i} to mean $(A_t^1, A_t^2, \dots, A_t^{i-1}, A_t^{i+1} \dots, A_t^N)$. We remove superscripts or subscripts if we want to represent the whole vector, for example A_t represents $(A_t^1, \dots, A_t^N)$. 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 a set A by $I_A(\cdot)$. For any finite set $\mathcal{S}$, $\Delta(\mathcal{S})$ represents the space of probability measures on $\mathcal{S}$ and $|\mathcal{S}|$ represents its cardinality. We denote by $\mathbb{P}^g$ (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 $\mathbb{R}$. For a probabilistic strategy profile of players $(\beta_t^i)_{i\in\mathcal{N}}$ where the 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 $\prod_{j\neq i}\beta_t^j(a_t^j|a_{1:t-1}, x_{1:t}^j)$. All equalities and inequalities involving random variables are to be interpreted in the *a.s.* sense. For mappings with range function sets $f : \mathcal{A} \to (\mathcal{B} \to \mathcal{C})$ we use square brackets $f[a] \in \mathcal{B} \to \mathcal{C}$ to denote the image of $a \in \mathcal{A}$ through f and parentheses $f[a](b) \in \mathcal{C}$ to denote the image of $b \in \mathcal{B}$ through $f[a]$.

The paper is organized as follows. In Section II, we present the model for games with finite and infinite horizon. Section III serves as motivation for focusing on SPBE. In Section IV, for finite-horizon games, we present 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. In Section V, we extend that methodology to infinite-horizon games. Section VI discusses a concrete example of a public goods game with two agents and results are presented for both, finite and infinite horizon versions of the example. All proofs are presented in appendices.

II. Model and Preliminaries

We consider a discrete-time dynamical system with N strategic players in the set $\mathcal{N} \triangleq \{1, 2, \dots N\}$. We consider two cases: finite horizon $\mathcal{T} \triangleq \{1, 2, \dots T\}$ with perfect recall and infinite horizon with perfect

recall. The system state is $X_t \triangleq (X_t^1, X_t^2, \dots X_t^N)$, 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. Players' types evolve as conditionally independent, controlled Markov processes such that

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

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

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

where Q_t^i are known kernels. Player i at time t takes action $a_t^i \in \mathcal{A}^i$ on observing the actions $a_{1:t-1} = (a_k)_{k=1,\dots,t-1}$ where $a_k = \left(a_k^j\right)_{j\in\mathcal{N}}$, which is common information among players, and the types $x_{1:t}^i$, 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 $g_t^i : \mathcal{A}^{t-1} \times (\mathcal{X}^i)^t \to \Delta(\mathcal{A}^i)$ 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 $R_t^i(x_t, a_t)$. To preserve the information structure of the problem, we assume that players do not observe their rewards until the end of game[3]. The reward functions and state transition kernels are common knowledge among the players. For the finite-horizon problem, the objective of player i is to maximize its total expected reward

$$J^{i,g} \triangleq \mathbb{E}^g \left\{ \sum_{t=1}^{T} R_t^i(X_t, A_t) \right\}. \tag{2}$$

For the infinite-horizon case, the transition kernels Q_t^i are considered to not depend on time t. We also substitute $R_t^i(X_t, A_t) = \delta^t R^i(X_t, A_t)$ take $\lim_{T\to\infty}$ in the above equation, where $\delta \in [0, 1)$ is the common discount factor and R^i is the time invariant stage reward function for agent i. With all players being strategic, this problem is modeled as a dynamic game, $\mathfrak{D}_T$ for finite horizon and $\mathfrak{D}_\infty$ for infinite horizon, with asymmetric information and simultaneous moves.

A. 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, $\mathcal{H}^T \triangleq \cup_{t=0}^{T} \mathcal{H}_t$ be the set of all possible such histories in finite horizon and $\mathcal{H}^\infty \triangleq \cup_{t=0}^{\infty} \mathcal{H}_t$ for infinite horizon. 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 [5], which consists of a pair (β^*, μ^*) of strategy profile $\beta^* = (\beta_t^{*,i})_{t\in\mathcal{T}, i\in\mathcal{N}}$ where $\beta_t^{*,i} : \mathcal{H}_t^i \to \Delta(\mathcal{A}^i)$ and a belief profile $\mu^* = ({}^i\mu_t^*)_{t\in\mathcal{T}, i\in\mathcal{N}}$ where ${}^i\mu_t^* : \mathcal{H}_t^i \to \Delta(\mathcal{H}_t)$ that satisfy sequential rationality so that $\forall i \in \mathcal{N}, t \in \mathcal{T}, h_t^i \in \mathcal{H}_t^i, \beta^i$

$$W_t^{i,\beta^*,i,T}(h_t^i) \geq W_t^{i,\beta^i,T}(h_t^i) \tag{3}$$

where the reward-to-go is defined as

$$W_t^{i,\beta^i,T}(h_t^i) \triangleq \mathbb{E}^{\beta^i\beta^{*,-i},\, {}^i\mu_t^*[h_t^i]} \left\{ \sum_{n=t}^{T} R_n^i(X_n, A_n) \big| h_t^i \right\}, \tag{4}$$

[3]Alternatively, we could have assumed instantaneous reward of a player to depend only on its own type, i.e. be of the form $R_t^i(x_t^i, a_t)$, and have allowed rewards to be observed by the players during the game as this would not alter the information structure of the game

and the beliefs satisfy some consistency conditions as described in [5, p. 331]. Similarly, for the game $\mathfrak{D}_\infty$ PBE (β^*, μ^*) requires: $\forall i \in \mathcal{N}, t \geq 1, h_t^i \in \mathcal{H}_t^i, \beta^i$

$$W_t^{i,\beta^*,{}^i}(h_t^i) \geq W_t^{i,\beta^i}(h_t^i) \tag{5}$$

where the reward-to-go is

$$W_t^{i,\beta^i}(h_t^i) \triangleq \mathbb{E}^{\beta^i \beta^{*,-i},\, {}^i\mu_t^*[h_t^i]} \left\{ \sum_{n=t}^{\infty} R_n^i(X_n, A_n) \big| h_t^i \right\}. \tag{6}$$

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' types $x_{1:t}^{-i} \in \times_{n=1}^t (\times_{j \neq i} \mathcal{X}^j)$ and their future actions. In our setting, due to independence of types, and given the common history h_t^c, player i's type history $x_{1:t}^i$ does not provide any additional information about $x_{1:t}^{-i}$, 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 $\mu_t^*[a_{1:t-1}]$. Furthermore, as will be show later, this belief factorizes into a product of marginals $\prod_{j \in \mathcal{N}} \mu_t^{*,j}[a_{1:t-1}]$. Thus we can sufficiently use the system of beliefs, $\mu^* = (\underline{\mu}_t^*)_{t \in \mathcal{T}}$, where $\underline{\mu}_t^* = (\mu_t^{*,i})_{i \in \mathcal{N}}$, and $\mu_t^{*,i} : \mathcal{H}_t^c \to \Delta(\mathcal{X}^i)$, 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 [5, p. 331] are automatically satisfied.

III. 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 $\mathfrak{D}_T$. 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 players' current types, which are their 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 IV.

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 1 that given the 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 $A_t^i \sim s_t^i(\cdot | a_{1:t-1}, x_t^i)$. Then in Lemma 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, $\pi_t \in \Delta(\mathcal{X})$, is defined as $\pi_t(x_t) \stackrel{\triangle}{=} \mathbb{P}^g(X_t = x_t | a_{1:t-1}),\ \forall x_t \in \mathcal{X}$. We also define the marginals $\pi_t^i(x_t^i) \stackrel{\triangle}{=} \mathbb{P}^g(x_t^i = x_t^i | a_{1:t-1}),\ \forall x_t^i \in \mathcal{X}^i$.

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

Using the agent-by-agent approach [24], we show in Lemma 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 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

$$\mathbb{P}^{s^i g^{-i}}(x_t, a_t) = \mathbb{P}^{g^i g^{-i}}(x_t, a_t) \qquad \forall t \in \mathcal{T}, x_t \in \mathcal{X}, a_t \in \mathcal{A}, \tag{7}$$

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

Proof: Please 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 in [30]. The common agent approach is a general approach that has been used extensively in dynamic team problems [31]–[34]. 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 $\gamma_t^i : \mathcal{X}^i \to \Delta(\mathcal{A}^i)$ is a partial (stochastic) function from its private information x_t^i to a_t^i, of the form $A_t^i \sim \gamma_t^i(\cdot|x_t^i)$. These actions are generated through some policy $\psi^i = (\psi_t^i)_{t\in\mathcal{T}}$, $\psi_t^i : \mathcal{A}^{t-1} \to \{\mathcal{X}^i \to \Delta(\mathcal{A}^i)\}$, that operates on the common information $a_{1:t-1}$ such 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 $A_t^i \sim \psi_t^i[a_{1:t-1}](\cdot|x_t^i)$.

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 $\theta_t^i : \Delta(\mathcal{X}) \to \{\mathcal{X}^i \to \Delta(\mathcal{A}^i)\}$, 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 summary, the notation for the various functional form of strategies is

$$\begin{aligned} &A_t^i \sim s_t^i(\cdot|a_{1:t-1}, x_t^i) \qquad && A_t^i \sim \psi_t^i[a_{1:t-1}](\cdot|x_t^i), \\ &A_t^i \sim m_t^i(\cdot|\pi_t, x_t^i) && A_t^i \sim \theta_t^i[\pi_t](\cdot|x_t^i). \end{aligned}$$

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 2: For any given strategy profile s of all players, there exists a strategy profile m such that

$$\mathbb{P}^m(x_t, a_t) = \mathbb{P}^s(x_t, a_t) \qquad \forall t \in \mathcal{T}, x_t \in \mathcal{X}, a_t \in \mathcal{A}, \tag{9}$$

which implies $(J^{i,m})_{i\in\mathcal{N}} = (J^{i,s})_{i\in\mathcal{N}}$.

Proof: Please see Appendix B. ∎

The above two lemmas show that any reward profile that can be generated through a policy profile of type g can also be generated through a policy profile of type m. This is precisely the motivation for using SPBE which are equilibria based on policies of type m. It should be noted that the construction of s^i depends only on g^i (as shown in (44d)), while the construction of m^i depends on the whole policy profile g and not just on g^i, since the construction of θ^i depends on ψ in (56). 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.

We end this section by noting that finding general PBEs of type g of the game $\mathfrak{D}_T$ or $\mathfrak{D}_\infty$ would be a desirable goal, but due to the space of strategies growing exponentially with time, that would be computationally intractable. However Lemmas 1 and 2 suggest 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. Our goal is to devise an algorithm to find structured equilibria of type m of the dynamic game $\mathfrak{D}_T$ and $\mathfrak{D}_\infty$.

IV. A METHODOLOGY FOR SPBE COMPUTATION IN FINITE HORIZON

In this section we consider the finite horizon dynamic game $\mathfrak{D}_T$. In the previous section, in the proof of Lemma 2 and specifically in Claim 5, it was shown that due to the independence of types and their evolution as independent controlled Markov processes, for any strategy of the players, the joint common belief can be factorized as a product of its marginals i.e., $\pi_t(x_t) = \prod_{i=1}^N \pi_t^i(x_t^i), \forall x_t$. Since in this paper, we only deal with such joint beliefs, to accentuate this independence structure, we define $\underline{\pi}_t \in \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i)$ as vector of marginal beliefs where $\underline{\pi}_t := (\pi_t^i)_{i\in\mathcal{N}}$. In the rest of the paper, 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 the vector of belief updates as $\underline{F}(\underline{\pi}, \gamma, a) := (F^i(\pi^i, \gamma^i, a))_{i\in\mathcal{N}}$ where (using Bayes rule)

$$F^i(\pi^i,\gamma^i,a)(x_{t+1}^i) = \begin{cases} \frac{\sum_{x_t^i}\pi^i(x_t^i)\gamma^i(a^i|x_t^i)Q_t^i(x_{t+1}^i|x_t^i,a)}{\sum_{\tilde{x}_t^i}\pi^i(\tilde{x}_t^i)\gamma^i(a^i|\tilde{x}_t^i)} & \text{if } \sum_{\tilde{x}_t^i}\pi^i(\tilde{x}_t^i)\gamma^i(a^i|\tilde{x}_t^i) > 0 \\ \sum_{x_t^i}\pi^i(x_t^i)Q_t^i(x_{t+1}^i|x_t^i,a) & \text{if } \sum_{\tilde{x}_t^i}\pi^i(\tilde{x}_t^i)\gamma^i(a^i|\tilde{x}_t^i) = 0. \end{cases} \tag{10}$$

The update function F^i defined above depends on time t through the kernel Q_t^i (for the finite horizon model). For notational simplicity we suppress this dependence on t. We also change the notation of policies of type m and θ as follows, so they depend on $\underline{\pi}_t$ instead of π_t

$$m_t^i : \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i)\times\mathcal{X}^i \to \Delta(\mathcal{A}^i) \qquad \theta_t^i : \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i) \to \left\{\mathcal{X}^i \to \Delta(\mathcal{A}^i)\right\}. \tag{11}$$

A. Backward Recursion

In this section, we define an equilibrium generating function $\theta = (\theta_t^i)_{i\in\mathcal{N},t\in\mathcal{T}}$, where $\theta_t^i : \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i) \to \{\mathcal{X}^i \to \Delta(\mathcal{A}^i)\}$. In addition, we define a sequence of reward-to-go functions of player i at time t, $(V_t^i)_{i\in\mathcal{N},t\in\{1,2,\dots T+1\}}$, where $V_t^i : \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i)\times\mathcal{X}^i \to \mathbb{R}$. These quantities are generated through a backward recursive way, as follows.

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

$$V_{T+1}^i(\underline{\pi}_{T+1}, x_{T+1}^i) \stackrel{\triangle}{=} 0. \tag{12}$$

2. For $t = T, T-1, \dots 1$, $\forall \underline{\pi}_t \in \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i), \pi_t = \prod_{i\in\mathcal{N}}\pi_t^i$, 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[4], of the following fixed-point equation, $\forall i \in \mathcal{N}, x_t^i \in \mathcal{X}^i$,

$$\tilde{\gamma}_t^i(\cdot|x_t^i) \in \arg\max_{\gamma_t^i(\cdot|x_t^i)} \mathbb{E}^{\gamma_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R_t^i(X_t,A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t,\tilde{\gamma}_t,A_t),X_{t+1}^i)\big|x_t^i\right\}, \tag{13}$$

where expectation in (13) 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 $\underline{F}$ is defined above.
Furthermore, using the quantity $\tilde{\gamma}_t$ found above, define

$$V_t^i(\underline{\pi}_t, x_t^i) \stackrel{\triangle}{=} \mathbb{E}^{\tilde{\gamma}_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i},\pi_t}\left\{R_t^i(X_t,A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t,\tilde{\gamma}_t,A_t),X_{t+1}^i)\big|x_t^i\right\}. \tag{14}$$

It should be noted that in (13), $\tilde{\gamma}_t^i$ is not the outcome of a maximization operation as is the case in a best response equation of a Bayesian Nash equilibrium. Rather (13) is a different 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 (in the belief update $\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t) = (F^i(\pi_t^i, \tilde{\gamma}_t^i, A_t))_{i\in\mathcal{N}}$). 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 proof of Theorem 1.

To highlight the significance of the unique structure of(13), we contrast it with two alternate intuitive, but incorrect constructions.

[4]The problem of existence in this step will be discussed in Section VII.

(a) Following the common information approach as in decentralized team problems [30], instead of (13), suppose $\tilde{\gamma}_t^i$ were constructed as equilibrium actions of the common agent, i.e. for a fixed $\underline{\pi}_t$, $\forall i \in \mathcal{N}$,

$$\tilde{\gamma}_t^i \in \arg\max_{\gamma_t^i} \mathbb{E}^{\gamma_t^i \tilde{\gamma}_t^{-i}, \pi_t} \left\{ R_t^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \gamma_t^i \tilde{\gamma}_t^{-i}, A_t), X_{t+1}^i) \right\}. \tag{15}$$

It should be noted that in (15), 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, (15) is not conditioned on x_t^i, the private information of player i, as is 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 (4), which requires conditioning on the entire history $(a_{1:t-1}, x_{1:t}^i)$ 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 (13), suppose $\tilde{\gamma}_t^i$ 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 $\underline{\pi}_t, \forall i \in \mathcal{N}, x_t^i \in \mathcal{X}^i$,

$$\tilde{\gamma}_t^i \in \arg\max_{\gamma_t^i} \mathbb{E}^{\gamma_t^i(\cdot|x_t^i)\tilde{\gamma}_t^{-i}, \pi_t} \left\{ R_t^i(X_t, A_t) + V_{t+1}^i(\underline{F}(\underline{\pi}_t, \gamma_t^i \tilde{\gamma}_t^{-i}, A_t), X_{t+1}^i) \middle| x_t^i \right\}. \tag{16}$$

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 through (16). 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.

B. Forward Recursion

As discussed above, a pair of strategy and belief profile (β^*, μ^*) is a PBE if it satisfies (4). Based on θ defined above in (12)–(14), we now construct a set of strategies β^* and beliefs μ^* for the game $\mathfrak{D}_T$ in a forward recursive way, as follows[5]. 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 $\mu_t^{*,i}[a_{1:t-1}]$ is a belief on x_t^i, and $\mu_t^*[a_{1:t-1}]$ can be constructed from $\underline{\mu}_t^*[a_{1:t-1}]$ as $\mu_t^*[a_{1:t-1}](x_t) = \prod_{i=1}^N \mu_t^{*,i}[a_{1:t-1}](x_t^i),\ \forall a_{1:t-1} \in \mathcal{H}_t^c$.

1. Initialize at time $t = 1$,

$$\mu_1^*[\phi](x_1) := \prod_{i=1}^N Q_1^i(x_1^i). \tag{17}$$

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

$$\beta_t^{*,i}(a_t^i | a_{1:t-1}, x_{1:t}^i) = \beta_t^{*,i}(a_t^i | a_{1:t-1}, x_t^i) := \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]](a_t^i | x_t^i) \tag{18}$$

and

$$\mu_{t+1}^{*,i}[a_{1:t}] := F^i(\mu_t^{*,i}[a_{1:t-1}], \theta_t^i[\underline{\mu}_t^*[a_{1:t-1}]], a_t) \tag{19}$$

where F^i is defined in (10).

We now state our main result.

[5]As discussed in the preliminaries subsection on Section II, the equilibrium beliefs in SPBE, μ_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.

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

$$\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R_n^i(X_n, A_n)\big|a_{1:t-1}, x_{1:t}^i\right\} \geq \mathbb{E}^{\beta^{i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R_n^i(X_n, A_n)\big|a_{1:t-1}, x_{1:t}^i\right\}. \tag{20}$$

Proof: Please see Appendix C. ■

We emphasize that even though the backward-forward algorithm presented above finds a class of equilibrium strategies that are structured, the unilateral deviations of players in (20) are considered in the space of general strategies, i.e., the algorithm does not make any bounded rationality assumptions.

Several remarks are in order with regard to the above methodology and the result.

Remark 1: An intuitive explanation for why all players are able to use a common belief is as follows. The sequence of beliefs defined above serves 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, $x_{1:t}^i$ is conditionally independent of x_t^{-i} given the common history $a_{1:t-1}$, player i's knowledge of $x_{1:t}^i$ does not affect this belief and thus in our definition, all players can use the same belief μ^* which is independent of their private information.

Remark 2: Independence of types is a crucial assumption in proving the above result, which manifests itself in Lemma 4 in Appendix D, used in the proof of Theorem 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 equilibrium 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 $\widehat{\beta}_t^i$ 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 $(\pi_t)_{t\in\{t+1,\dots T\}}$ in the same way 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 $\widehat{\pi}_{t+1}$ on x_t which depends on strategy profile $\beta^*_{1:t-1}\widehat{\beta}_t, \beta_t^{*,-i}$. Thus at time t, player i would need both the beliefs $\pi_{t+1}, \widehat{\pi}_{t+1}$ to compute its expected future reward (as also discussed in [13]); π_{t+1} to predict other players' actions and $\widehat{\pi}_{t+1}$ to form a true belief on x_t based on its information. As it turns out, due to independence of types, $\widehat{\pi}_{t+1}$ 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 of player j till time t (as shown in Claim 5 in Appendix B); thus $\pi_{t+1}^{-i} = \widehat{\pi}_{t+1}^{-i}$. Now since player i already observes its type x_t^i, its belief $\widehat{\pi}_t^i$ on x_t^i does not provide any additional information to player i, and thus π_t (which is the same as $\mu_t^*[a_{1:t-1}]$) is sufficient to compute 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 $\widehat{\beta}_t^i$ given a_t^i. This implies player i can use the equilibrium strategy β_t^* to update its future belief, as used in (13). Then by construction of θ and specifically due to (13), player i does not gain by unilaterally deviating at time t keeping the remainder of its strategy the same.

Remark 3: 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_1 on types X_1. 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.

Remark 4: The following result shows that all SPBE can be found through the backward-forward methodology described before. An SPBE can be defined as a PBE (β^*, μ^*) of the game that is generated through forward recursion in (17)–(19), using an equilibrium generating function ϕ, where $\phi = (\phi_t^i)_{i\in\mathcal{N},t\in\mathcal{T}}$, $\phi_t^i : \times_{i\in\mathcal{N}}\Delta(\mathcal{X}^i) \to \{\mathcal{X}^i \to \Delta(\mathcal{A}^i)\}$, common belief update function $\underline{F}$ and prior distributions Q_1. 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 $\mu_t^{*,i}$ depends only on common information history $a_{1:t-1}$.

Theorem 2 (Converse): Let (β^*, μ^*) be an SPBE. Then there exists an equilibrium generating function ϕ that satisfies (13) 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 [6].

Proof: Please see Appendix E. ■

Remark 5: In this paper, we find a class of PBEs of the game, while there may exist other equilibria that are not 'structured', and can not be found by directly using the proposed methodology. The rationale for using structured equilibria over others is the same as that for using MPE over SPE for a symmetric information game; a focussing argument for using simpler strategies being one of them.

Remark 6: Using this methodology, dynamic LQG games with asymmetric information are studied in [35] where it is shown that under certain conditions, there exists an SPBE of the game with strategies being linear in players' private types. In [36], authors extend the finite-horizon model in this paper such that players do not observe their own types, rather make independent noisy observations of their types. An analogous backward-forward algorithm is presented for that model.

Finally, existence of the solution of the fixed-point equation in (13) is discussed in Section VII.

V. A Methodology for SPBE computation in Infinite Horizon

In this section we consider the infinite horizon discounted reward dynamic game $\mathfrak{D}_\infty$. We state the fixed-point equation that defines the value function and strategy mapping for the infinite horizon problem. This is analogous to the backwards recursion ((13) and (14)) that define the value function and θ mapping for the finite horizon problem.

Define the set of functions $V^i : \times_{j=1}^N \Delta(\mathcal{X}^j) \times \mathcal{X}^i \to \mathbb{R}$ and strategies $\tilde{\gamma}^i : \mathcal{X}^i \to \Delta(\mathcal{A}^i)$ (which are generated formally as $\tilde{\gamma}^i = \theta^i[\underline{\pi}]$ for given $\underline{\pi}$) via the following fixed-point equation: $\forall\ i \in \mathcal{N}$, $x^i \in \mathcal{X}^i$,

$$\tilde{\gamma}^i(\cdot \mid x^i) \in \operatorname*{argmax}_{\gamma^i(\cdot|x^i)\in\Delta(\mathcal{A}^i)} \mathbb{E}^{\gamma^i(\cdot|x^i),\tilde{\gamma}^{-i},\pi^{-i}}\big[R^i(X,A) + \delta V^i\big(\underline{F}(\underline{\pi},\tilde{\gamma},A), X'^i\big) \mid \underline{\pi}, x^i\big], \tag{21a}$$

$$V^i(\underline{\pi}, x^i) = \mathbb{E}^{\tilde{\gamma}^i(\cdot|x^i),\tilde{\gamma}^{-i},\pi^{-i}}\big[R^i(X,A) + \delta V^i\big(\underline{F}(\underline{\pi},\tilde{\gamma},A), X'^i\big) \mid \underline{\pi}, x^i\big]. \tag{21b}$$

Note that the above is a joint fixed-point equation in $(V, \tilde{\gamma})$, unlike the backwards recursive algorithm earlier which required solving a fixed-point equation only in $\tilde{\gamma}$. Here the unknown quantity is distributed as

$$(X^{-i}, A^i, A^{-i}, X'^i) \sim \pi^{-i}(x^{-i})\gamma^i(a^i \mid x^i)\tilde{\gamma}^{-i}(a^{-i} \mid x^{-i})Q^i(x'^i \mid x^i, a). \tag{22}$$

and $F^i(\cdot)$ is defined in (10).

Define the belief μ^* inductively similar to the forward recursion from Section IV-B. By construction the belief defined above satisfies the consistency condition needed for a PBE. Denote the strategy arising out of $\tilde{\gamma}$ by β^* i.e.,

$$\beta_t^{i,*}(a_t^i \mid x_{1:t}^i, a_{1:t-1}) = \theta^i\big[\mu_t^*[a_{1:t-1}]\big](a_t^i \mid x_t^i). \tag{23}$$

Note that although the mapping θ^i is stationary, the strategy $\beta_t^{i,*}$ derived from it is not so.

Below we state the central result of this section. It states that the strategy-belief pair (β^*, μ^*) constructed from the solution of the fixed-point equation (21) and the forward recursion indeed constitutes a PBE.

Theorem 3: Assuming that the fixed-point equation (21) admits an absolutely bounded solution V^i (for all $i \in \mathcal{N}$), the strategy-belief pair (β^*, μ^*) defined in (23) is a PBE of the infinite horizon discounted reward dynamic game i.e., $\forall\ i \in \mathcal{N}$, β^i, $t \geq 1$, $h_t^i \in \mathcal{H}_t^i$,

$$\mathbb{E}^{\beta^{i,*},\beta^{-i,*},\mu_t^*[h_t^c]}\big[\sum_{n=t}^{\infty}\delta^{n-t}R^i(X_n,A_n) \mid h_t^i\big] \geq \mathbb{E}^{\beta^i,\beta^{-i,*},\mu_t^*[h_t^c]}\big[\sum_{n=t}^{\infty}\delta^{n-t}R^i(X_n,A_n) \mid h_t^i\big]. \tag{24}$$

[6]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^*).

Proof: Please see Appendix F. ■

Our approach to proving Theorem 3 is as follows. We begin by noting that the standard contraction mapping arguments used in infinite horizon discounted reward MDPs/POMDPs viewed as a limit of finite horizon problems, do not apply here, since the policy equation (21a) is not a maximization, but a different fixed-point equation. So we attempt to "fit" the infinite horizon problem into the framework of finite-horizon model developed in the previous section. We do that by first introducing a terminal reward that depends on common beliefs, in the backward-forward recursion construction of Section IV for finite horizon games. We consider a finite horizon, $T > 1$, dynamic game with rewards same as in the infinite horizon version and time invariant transition kernels Q^i. For each agent i, there is a terminal reward $G^i(\pi_{T+1}, x^i_{T+1})$ that depends on the terminal type of agent i and the terminal belief. It is assumed that $G^i(\cdot)$ is absolutely bounded. We define the value functions $\left(V_t^{i,T} : \times_{j\in\mathcal{N}}\Delta(\mathcal{X}^j) \times \mathcal{X}^i \to \mathbb{R}\right)_{i\in\mathcal{N},t\in\mathcal{T}}$ and strategies $\left(\tilde{\gamma}_t^{i,T}\right)_{i\in\mathcal{N},t\in\mathcal{T}}$ backwards inductively in the same way as in Section IV-A except Step 1, where instead of (12) we set $V^{i,T}_{T+1} \equiv G^i$. This consequently results in a strategy/belief pair (β^*, μ^*), based on the forward recursion in Section IV-B. Now, due to the above construction, the value function $V_t^{i,T}$ from above and V^i from (21) are related (see Lemma 9, Appendix G). This result combined with continuity arguments as $T \to \infty$ complete the proof of Theorem 3.

VI. A Concrete Example of Multi-stage investment in Public Goods

In this section, we discuss both, a two-stage (finite) and an infinite-horizon version of a public goods example to illustrate the methodology described above for the construction of SPBEs.

A. A two stage public goods game

We consider a discrete version of Example 8.3 from [5, ch.8], which is an instance of a repeated public goods 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, which is a binary decision $a^i_t \in \{0, 1\}$ 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^i's 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

$$R^i_t(x, a_t) = \delta^t R^i(x, a_t) \quad \text{with} \quad R^i(x, a_t) = \begin{cases} a_t^{-i} & \text{if } a^i_t = 0 \\ 1 - x^i & \text{if } a^i_t = 1. \end{cases} \tag{25}$$

We set $\delta = 1$ in this two-stage case. We will use the backward recursive algorithm, defined in Section IV, to find an SPBE of this game. For period $t = 1, 2$ and for $i = 1, 2$, the partial functions γ^i_t can equivalently be defined through scalars p^{iL}_t and p^{iH}_t such that

$$\gamma^i_t(1|x^L) = p^{iL}_t, \quad \gamma^i_t(0|x^L) = 1 - p^{iL}_t, \quad \gamma^i_t(1|x^H) = p^{iH}_t, \quad \gamma^i_t(0|x^H) = 1 - p^{iH}_t, \tag{26}$$

where $p^{iL}_t, p^{iH}_t \in [0, 1]$. Henceforth, we will use p^{iL}_t and p^{iH}_t interchangeably with the corresponding γ^i_t.

For $t = 2$ and for any fixed $\underline{\pi}_2 = (\pi^1_2, \pi^2_2)$, where $\pi^i_2 = \pi^i_2(x^H) \in [0, 1]$ represents a probability measure on the event $\{X^i = x^H\}$, player i's reward is

$$\mathbb{E}^{\gamma_2}\{R^i_2(X, A_2)|\pi_2, X^i = x^L\} = (1 - p^{iL}_2)\left((1 - \pi^{-i}_2)p^{-iL}_2 + \pi^{-i}_2 p^{-iH}_2\right) + p^{iL}_2(1 - x^L), \tag{27a}$$

$$\mathbb{E}^{\gamma_2}\{R^i_2(X, A_2)|\pi_2, X^i = x^H\} = (1 - p^{iH}_2)\left((1 - \pi^{-i}_2)p^{-iL}_2 + \pi^{-i}_2 p^{-iH}_2\right) + p^{iH}_2(1 - x^H). \tag{27b}$$

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 (13). For $i = 1, 2$

$$\tilde{p}_2^{iL} \in \arg\max_{p_2^{iL}} \quad (1 - p_2^{iL}) \left((1 - \pi_2^{-i})\tilde{p}_2^{-iL} + \pi_2^{-i}\tilde{p}_2^{-iH}\right) + p_2^{iL}(1 - x^L), \tag{28a}$$

$$\tilde{p}_2^{iH} \in \arg\max_{p_2^{iH}} \quad (1 - p_2^{iH}) \left((1 - \pi_2^{-i})\tilde{p}_2^{-iL} + \pi_2^{-i}\tilde{p}_2^{-iH}\right) + p_2^{iH}(1 - x^H). \tag{28b}$$

Since $1 - x^H < 0$, $\tilde{p}_2^{iH} = 0$ achieves the maximum in (28b). Thus (28a)–(28b) can be reduced to, $\forall i \in \{1, 2\}$

$$\tilde{p}_2^{iL} \in \arg\max_{p_2^{iL}} \quad (1 - p_2^{iL})(1 - \pi_2^{-i})\tilde{p}_2^{-iL} + p_2^{iL}(1 - x^L). \tag{29}$$

This implies,

$$\tilde{p}_2^{iL} = \begin{cases} 0 & \text{if} \quad x^L > 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}, \\ 1 & \text{if} \quad x^L < 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}, \\ \text{arbitrary} & \text{if} \quad x^L = 1 - (1 - \pi_2^{-i})\tilde{p}_2^{-iL}. \end{cases} \tag{30}$$

The solutions of the fixed point equation (30) are shown in Figure 1 in the space of (π_2^1, π_2^2).

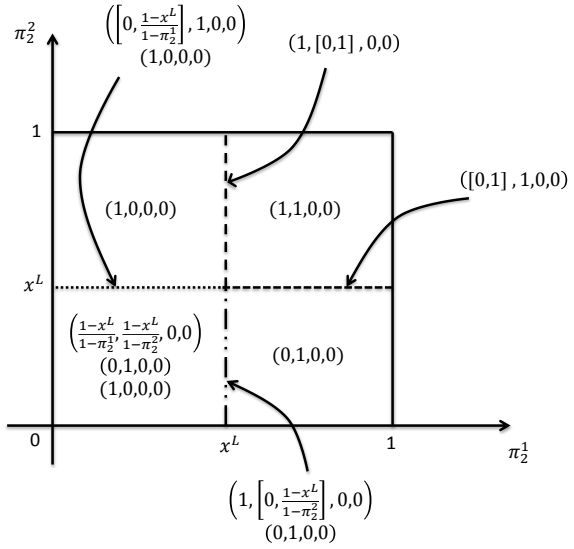

Fig. 1: Solutions of fixed point equation in (30). Solutions are shown as quadruplets $(\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H})$ with intervals in place of single values whenever the solution is not uniquely defined.

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, $\forall i \in \{1, 2\}$

$$\begin{aligned} \tilde{p}_1^{iL} \in \arg\max_{p_1^{iL}} \quad & (1 - p_1^{iL}) \left((1 - q)\tilde{p}_1^{-iL} + q\tilde{p}_1^{-iH} + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(\underline{F}(Q^2, \tilde{\gamma}_1, (0, A_1^{-i})), x^L)\}\right) \\ & + p_1^{iL} \left(1 - x^L + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(\underline{F}(Q^2, \tilde{\gamma}_1, (1, A_1^{-i})), x^L)\}\right). \end{aligned} \tag{31a}$$

$$\begin{aligned} \tilde{p}_1^{iH} \in \arg\max_{p_1^{iH}} \quad & (1 - p_1^{iH}) \left((1 - q)\tilde{p}_1^{-iL} + q\tilde{p}_1^{-iH} + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(\underline{F}(Q^2, \tilde{\gamma}_1, (0, A_1^{-i})), x^H)\}\right) \\ & + p_1^{iH} \left(1 - x^H + \mathbb{E}^{\tilde{\gamma}_1}\{V_2^i(\underline{F}(Q^2, \tilde{\gamma}_1, (1, A_1^{-i})), x^H)\}\right). \end{aligned} \tag{31b}$$

where $\underline{F}(Q^2, \tilde{\gamma}, (A^1, A^2)) \triangleq (F^1(Q, \tilde{\gamma}^1, A^1), F^2(Q, \tilde{\gamma}^2, A^2))$ (for this example $F^i(\cdot, \cdot, A)$ depends only on A^i, hence this expression is written with slight abuse of notation) and

$$F^i(Q, \tilde{\gamma}_1^i, 0) = \frac{q(1 - \tilde{p}_1^{iH})}{q(1 - \tilde{p}_1^{iH}) + (1 - q)(1 - \tilde{p}_1^{iL})}, \tag{32a}$$

$$F^i(Q, \tilde{\gamma}_1^i, 1) = \frac{q\tilde{p}_1^{iH}}{q\tilde{p}_1^{iH} + (1 - q)\tilde{p}_1^{iL}}, \tag{32b}$$

if the denominators in (32a)–(32b) are strictly positive, else $F^i(Q, \tilde{\gamma}_1^i, A^i) = Q$ as in the proof of Lemma 2, and in particular Claim 5. A solution of the fixed point equation in (31a)-(31b) 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 (30), shown in Figure 2 and described below,

$$\theta_2[\pi_2] = (\tilde{p}_2^{1L}, \tilde{p}_2^{2L}, \tilde{p}_2^{1H}, \tilde{p}_2^{2H}) = \begin{cases} (\frac{1-x^L}{1-\pi_2^1}, \frac{1-x^L}{1-\pi_2^2}, 0, 0) & \pi_2^1 \in [0, x^L), \pi_2^2 \in [0, x^L) \\ (1, 0, 0, 0) & \pi_2^1 \in [0, x^L], \pi_2^2 \in [x^L, 1] \\ (0, 1, 0, 0) & \pi_2^1 \in [x^L, 1], \pi_2^2 \in [0, x^L] \\ (1, 1, 0, 0) & \pi_2^1 \in (x^L, 1], \pi_2^2 \in (x^L, 1]. \end{cases} \tag{33}$$

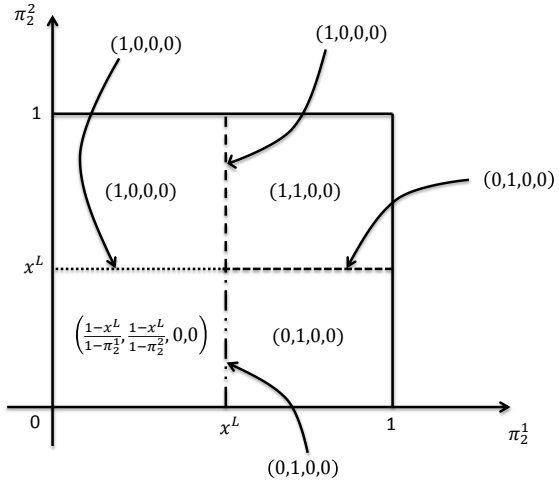

Fig. 2: Specific solution $\theta_2[\underline{\pi}_2]$ described in (33).

Then, through iteration on the fixed point equation (31a)-(31b) 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 $(\beta_2^i(\cdot | a_1, \cdot))_{i \in \{1,2\}} = \theta_2[\mu_2^*[a_1]]$ 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 can 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 $p = \frac{q(1+x^L)}{q(1+x^L)+(1-x^L)}$.

B. Infinite horizon version

In this section, we consider an infinite horizon version of the above public goods dynamic game. We consider three values $\delta = 0, 0.5, 0.95$. We solve the corresponding fixed point equation (arising out of (21)) numerically to calculate the mapping θ (which in turn generates the perfect Bayesian equilibrium (β^*, μ^*)).

We solve the fixed-point equation numerically by discretizing the $\underline{\pi}-$space $[0, 1]^2$ and all solutions that we find are symmetric w.r.t. agents i.e., $\tilde{p}^{1L}$ for $\underline{\pi} = (\pi_1, \pi_2)$ is the same as $\tilde{p}^{2L}$ for $\underline{\pi}' = (\pi_2, \pi_1)$ and similarly for $\tilde{p}^{1H}, \tilde{p}^{2H}$.

For $\delta = 0$, the game is instantaneous and actually corresponds to the second round $t = 2$ play in the finite horizon two-stage version above. Thus whenever agent 1's type is x^H, it is instantaneously profitable not to contribute. This gives $\tilde{p}^{1H} = 0$, for all $\underline{\pi}$. Thus we only plot $\tilde{p}^{1L}$; in Fig. 3 (this can be inferred from the discussion and Fig. 1 above). Intuitively, with type x^L the only values of $\underline{\pi}$ for which agent 1 would not wish to contribute is if he anticipates agent 2's type to be x^L with high probability and rely on agent 2 to contribute. This is why for lower values of π_2 (i.e., agent 2's type likely to be x^L) we see $\tilde{p}^{1L} = 0$ in Fig. 3.

Now consider $\tilde{p}^{1L}$ plotted in Fig. 3, 4 and 6. As δ increases, future rewards attain more priority and signaling comes into play. So while taking an action, agents not only look for their instantaneous reward but also how their action affects the future public belief π about their private type. It is evident in the figures that as δ increases, at high π_1, up to larger values of π_2 agent 1 chooses not to contribute when his type is x^L. This way he intends to send a "wrong" signal to agent 2 i.e., that his type is x^H and subsequently force agent 2 to invest. This way agent 1 can free-ride on agent 2's investment.

Now consider Fig. 5 and 7, where $\tilde{p}^{1H}$ is plotted. For $\delta = 0$ we know that it is profitable to not contribute, however as δ increases from 0, agents are mindful of future rewards and thus are willing to contribute at certain beliefs. Specifically, coordination via signaling is evident here. Although it is instantaneously not profitable to contribute if agent 1's type is x^H, by contributing at higher values of π_2 (i.e., agent 2's type is likely x^H) and low π_1, agent 1 coordinates with agent 2 to achieve net profit greater than 0 (reward when no one contributes). This can be done since the loss of contributing is -0.2 whereas profit from free-riding on agent 2's contribution is 1.

Under the equilibrium strategy, beliefs $\underline{\Pi}_t$ form a Markov chain. One can trace this Markov chain to study the signaling effect at equilibrium. On numerically simulating this Markov chain for the above example (at $\delta = 0.95$) we observe that for almost all initial beliefs, within a few rounds agents completely learn each other's private type truthfully (or at least with very high probability). In other words, agents manage to reveal their private type via their actions at equilibrium and to such an extent that it negates any possibly incorrect initial belief about their type.

As a measure of cooperative coordination at equilibrium one can perform the following calculation. Compare the value function $V^1(\cdot, x)$ of agent 1 arising out of the fixed-point equation, for $\delta = 0.95$ and $x \in \{x^H, x^L\}$ (normalize it by multiplying with $1 - \delta$ so that it represents per-round value) with the best possible attainable single-round reward under a symmetric mixed strategy with a) full coordination and b) no coordination. Note that the two cases need not be equilibrium themselves, which is why this will result in a bound on the efficiency of the evaluated equilibria.

In case a), assuming both agents have the same type x, full coordination can lead to the best possible reward of $\frac{1+1-x}{2} = 1 - \frac{x}{2}$ i.e., agent 1 contributes with probability 0.5 and agent 2 contributes with probability 0.5 but in a coordinated manner so that it doesn't overlap with agent 1 contributing.

In case b) when agents do not coordinate and invest with probability p each, then the expected single-round reward is $p(1-x)+p(1-p)$. The maximum possible value of this expression is $(1-\frac{x}{2})^2$.

For $x = x^L = 0.2$, the range of values of $V^1(\pi_1, \pi_2, x^L)$ over $(\pi_1, \pi_2) \in [0,1]^2$ is $[0.865, 0.894]$. Whereas full coordination produces 0.9 and no coordination 0.81. It is thus evident that agents at equilibrium end up achieving reward close to the best possible and gain significantly compared to the strategy of no coordination.

Similarly for $x = x^H = 1.2$ the range is $[0.3, 0.395]$. Whereas full coordination produces 0.4 and no coordination 0.16. The gain via coordination is evident here too.

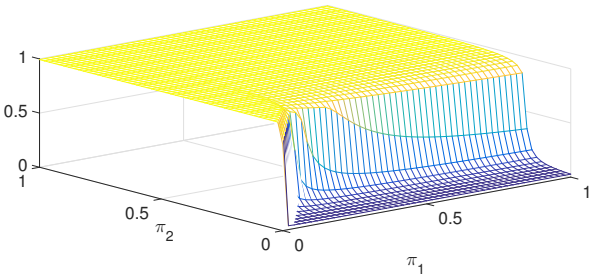

Fig. 3: $\tilde{p}^{1L}$ vs. (π_1, π_2) at $\delta = 0$ where $(\tilde{p}^{1L}, \tilde{p}^{1H}) = \theta^1[\pi_1, \pi_2]$.

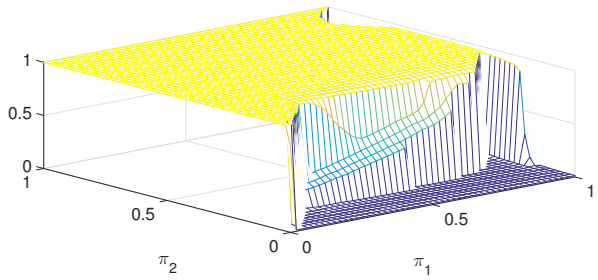

Fig. 4: $\tilde{p}^{1L}$ vs. (π_1, π_2) at $\delta = 0.5$.

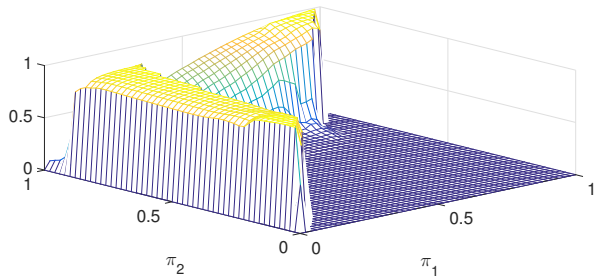

Fig. 5: $\tilde{p}^{1H}$ vs. (π_1, π_2) at $\delta = 0.5$.

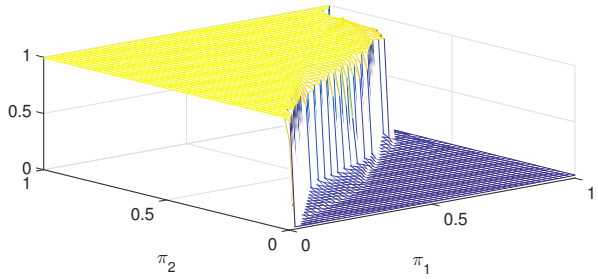

Fig. 6: $\tilde{p}^{1L}$ vs. (π_1, π_2) at $\delta = 0.95$.

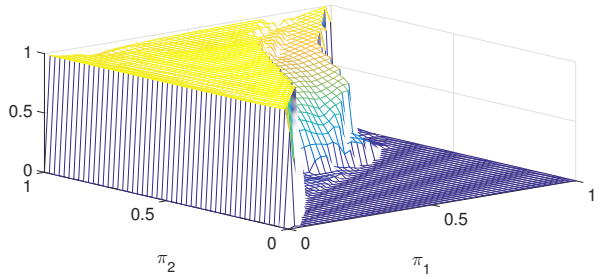

Fig. 7: $\tilde{p}^{1H}$ vs. (π_1, π_2) at $\delta = 0.95$.

VII. An Existence Result for the Fixed-Point Equation

In this section, we discuss the problem of existence of signaling equilibria[7]. While it is known that for any finite dynamic game with asymmetric information and perfect recall, there always exists a PBE [4, Prop. 249.1], existence of SPBE is not guaranteed. It is clear from our algorithm that existence of SPBEs boils down to existence of a solution to the fixed-point equation (13) in finite horizon and (21) in infinite horizon. Specifically, for the finite horizon, at each time t given the functions V^i_{t+1} for all $i \in \mathcal{N}$ from the previous round (in the backwards recursion) equation (13) must have a solution $\tilde{\gamma}^i_t$ for all $i \in \mathcal{N}$. Generally,

[7]In the special case of uncontrolled types where agent i's instantaneous reward does not depend on its private type x^i_t, 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 similar to the one for computing Nash equilibrium. This result is shown in [27].

existence of equilibria is shown through Kakutani's fixed point theorem, as is done in proving existence of a mixed strategy Nash equilibrium of a finite game [4], [37]. This is done by showing existence of fixed point of the best-response correspondences of the game. Among other conditions, it requires the "closed graph" property of the correspondences, which is usually implied by the continuity property of the utility functions involved. For (13) establishing existence is not straightforward due to: (a) potential discontinuity of the π_t update function F when the denominator in the Bayesian update is 0 and (b) potential discontinuity of the value functions, V^i_{t+1}. In the following we provide sufficient conditions that can be checked at each time t to establish the existence of a solution.

We consider a generic fixed-point equation similar to the one encountered in Section IV and Section V and state conditions under which they are guaranteed to have a solution. To concentrate on the essential aspects of the problem we consider a simple case with $N = 2$, type sets $\mathcal{X}^i = \{x^H, x^L\}$ and action sets $\mathcal{A}^i = \{0, 1\}$. Furthermore, types are static and instantaneous rewards $R^i(x, a)$ do not depend on x^{-i}.

Given public belief $\underline{\pi} = (\pi^1, \pi^2) \in \times_{i=1}^2 \Delta(A^i)$, value functions V^1, V^2, one wishes to solve the following system of equations for $\big(\tilde\gamma^i(\cdot \mid x^i)\big)_{x^i\in\{x^H,x^L\}, i\in\{1,2\}}$.

$$\tilde\gamma^i(\cdot \mid x^i) \in \underset{\gamma^i(\cdot|x^i)\in\Delta(\mathcal{A}^i)}{\operatorname{argmax}} \mathbb{E}^{\gamma^i(\cdot|x^i),\tilde\gamma^{-i}}\Big[R^i(x^i, A) + V^i\Big(\big(F^1(\pi^1, \tilde\gamma^1, A^1), F^2(\pi^2, \tilde\gamma^2, A^2)\big), x^i\Big) \mid x^i, \underline{\pi}\Big] \tag{35}$$

where the expectation is evaluated using the probability distribution

$$(A^1, A^2) \sim \gamma^i(a^i \mid x^i)\big[\pi^j(x^H)\tilde\gamma^j(a^j \mid x^H) + \pi^j(x^L)\tilde\gamma^j(a^j \mid x^L)\big]. \tag{36}$$

The probabilistic policy $\tilde\gamma$ can be represented by the 4-tuple $\underline{p} = \big(\tilde p^{1L}, \tilde p^{2L}, \tilde p^{1H}, \tilde p^{2H}\big)$ where $\tilde p^{iH} = \gamma^i(a^i = 1 \mid x^H)$ and $\tilde p^{iL} = \gamma^i(a^i = 1 \mid x^L)$, $i = 1, 2$.

The fixed-point equation of interest reduces to

$$\begin{aligned}\tilde p^{1H} \in \underset{a\in[0,1]}{\operatorname{argmax}}\, a\Big[&\big(\pi^2\tilde p^{2H}+(1-\pi^2)\tilde p^{2L}\big)\big(V^1(F_1(\pi^1,\tilde p^1),F_1(\pi^2,\tilde p^2),x^H)-V^1(F_0(\pi^1,\tilde p^1),F_1(\pi^2,\tilde p^2),x^H)\big)\\
&+\big(1-\pi^2\tilde p^{2H}-(1-\pi^2)\tilde p^{2L}\big)\big(V^1(F_1(\pi^1,\tilde p^1),F_0(\pi^2,\tilde p^2),x^H)-V^1(F_0(\pi^1,\tilde p^1),F_0(\pi^2,\tilde p^2),x^H)\big)\\
&+\big(\pi^2\tilde p^{2H}+(1-\pi^2)\tilde p^{2L}\big)\big(R^1(x^H,1,1)-R^1(x^H,0,1)\big)\\
&+\big(1-\pi^2\tilde p^{2H}-(1-\pi^2)\tilde p^{2L}\big)\big(R^1(x^H,1,0)-R^1(x^H,0,0)\big)\Big]\end{aligned} \tag{37}$$

and three other similar equations for $\tilde p^{1L}, \tilde p^{2H}, \tilde p^{2L}$. Here

$$F_1(\pi, (p^H, p^L)) \triangleq \frac{\pi p^H}{\pi p^H + \overline{\pi} p^L} \tag{38a}$$

$$F_0(\pi, (p^H, p^L)) \triangleq \frac{\pi(1 - p^H)}{\pi(1 - p^H) + \overline{\pi}(1 - p^L)} \tag{38b}$$

and in both definitions, if the denominator is 0 then the RHS is taken as π.

A. Points of Discontinuities and the Closed graph result

Equation (37) and the other three similar equations are essentially of the form (for a given $\underline{\pi}$)

$$x \in \underset{a\in[0,1]}{\operatorname{argmax}}\ af_1(x, y, w, z) \tag{39a}$$

$$y \in \underset{b\in[0,1]}{\operatorname{argmax}}\ bf_2(x, y, w, z) \tag{39b}$$

$$w \in \underset{c\in[0,1]}{\operatorname{argmax}}\ cf_3(x, y, w, z) \tag{39c}$$

$$z \in \underset{d\in[0,1]}{\operatorname{argmax}}\ df_4(x, y, w, z) \tag{39d}$$

with x, y, z, w as $\tilde{p}^{1H}, \tilde{p}^{1L}, \tilde{p}^{2H}, \tilde{p}^{2L}$, respectively.

Define $\mathcal{D}_i \subseteq [0,1]^4$ as the set of discontinuity points of f_i and $\mathcal{D} \triangleq \cup_{i=1}^4 \mathcal{D}_i$.

For any point $\underline{x}_0 \in \mathcal{D}$, define $S(\underline{x}_0)$ as the subset of indexes $i \in \{1,2,3,4\}$ for which $f_i(\underline{x})$ is discontinuous at $\underline{x}_0$.

Assumption (E1): At any point $\underline{x}_0 \in \mathcal{D}$, $\forall\ i \in S(\underline{x}_0)$ one of the following is satisfied:

1) $f_i(\underline{x}_0) = 0$, or
2) $\exists\ \epsilon > 0$ such that $\forall\ \underline{x} \in B_\epsilon(\underline{x}_0)$ (inside an ϵ-ball of $\underline{x}_0$) the sign of $f_i(\underline{x})$ is same as the sign of $f_i(\underline{x}_0)$.

In the following we provide a sufficient condition for existence.

Theorem 4: Under Assumption (E1), there exists a solution to the fixed-point equation (39).

Proof: Please see Appendix H. ∎

The above set of results provide us with an analytical tool for establishing existence of a solution to the concerned fixed-point equation.

While the above analytical result is useful in understanding a theoretical basis for existence, it doesn't cover all instances. For instance, fixed-point equation (31) from Section VI-A, does not satisfy assumption (E1). In the following we provide a more computationally orientated approach to establishing existence and/or solving the generic fixed-point equation (39).

We motivate this case-by-case approach with the help of an example. Suppose we hypothesize that the solution to (39) is such that $x = 0, w = 0$ and $y, z \in (0,1)$. Then (39) effectively reduces to checking if there exists $y^*, z^* \in (0,1)$ such that

$$y^* \in \operatorname*{argmax}_b b f_2(0, y^*, 0, z^*) \tag{40a}$$

$$z^* \in \operatorname*{argmax}_d d f_4(0, y^*, 0, z^*) \tag{40b}$$

$$f_1(0, y^*, 0, z^*) \leq 0 \tag{40c}$$

$$f_3(0, y^*, 0, z^*) \leq 0. \tag{40d}$$

Thus the 4-variable system reduces to solving a 2-variable system and 2 conditions to verify. For instance, if $f_2(0, y, 0, z)$, $f_4(0, y, 0, z)$ as functions of y, z satisfy the conditions of Theorem 4 then the sub-system (40a), (40b) has a solution. If one of these solution is also consistent with (40c), (40d) then this sub-case indeed provides a solution to (39).

Generalizing the simplification provided in the above example, we divide solutions into $3^4 = 81$ cases based on whether each of x, y, w, z are in $\{0\}, (0,1), \{1\}$. There are (1) 16 corner cases where none are in the strict interior $(0,1)$; (2) 32 cases where exactly one is in the strict interior $(0,1)$; (3) 24 cases where 2 variables are in the strict interior $(0,1)$; (4) 8 cases where 3 variables are in the strict interior $(0,1)$; and (5) 1 case where all 4 variables are in the strict interior $(0,1)$.

Similar to the calculations above, for each of the 81 cases one can write a sub-system to which the problem (39) effectively reduces to. Clearly, if any one of the 81 sub-systems has a solution then the problem (39) has a solution. Furthermore, searching for a solution reduces to an appropriate sub-problem depending on the case.

The approach then is to enumerate each of these 81 cases (as stated above) and check them in order. The computational simplification arises out of the fact that whenever a variable, say y, is not in the strict interior $(0,1)$ then the corresponding equation (39b) need not be solved, since one only needs to verify the sign at a specific point. Hence, all sub-cases of (1) reduce to simply checking the value of functions f_i at corner points - no need for solving a fixed-point equation. All sub-cases of (2) reduce to solving a 1-variable fixed-point equation and three corresponding conditions to verify, etc.

VIII. Conclusion

In this paper, we study both finite and infinite-horizon models of 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 $\underline{\pi}_t$, 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 θ. 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. For the infinite-horizon model, the equilibrium generating function is defined using a single-shot fixed-point equation, which in conjunction with the forward recursion defines equilibrium strategies and beliefs. Finally, we demonstrate our methodology by a concrete example of a two agent symmetric public goods game where the signaling effect is observed at equilibrium. The signaling effect implies that agents, at equilibrium, take into account how their actions affect future public beliefs $\underline{\pi}_t$ about their private type.

In general, this methodology enables a systematic study of PBE for many applications, analytically or numerically, which was not feasible before. Some interesting future directions include: (a) finding structural results for games modeling real-life systems like communication systems, industry dynamics, labor markets; (b) finding sufficient conditions for existence of solution of per-stage fixed point equation of finite-horizon and the single-shot fixed point equation for the infinite-horizon model; and (c) dynamic mechanism design for such games, among others.

ACKNOWLEDGMENT

The authors wish to acknowledge Vijay Subramanian for his contribution to the paper. Achilleas Anastasopoulos wishes to acknowledge Ashutosh Nayyar for the fruitful discussion and criticism of an early draft of this work presented during the ITA 2012 conference.

APPENDIX A
PROOF OF LEMMA 1

We prove this Lemma in the following steps.

(a) In Claim 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 2, using Claim 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 $s^i_t(a^i_t|a_{1:t-1}, x^i_t) \stackrel{\triangle}{=} \mathbb{P}^g(a^i_t|a_{1:t-1}, x^i_t)$.
(d) In Claim 3, we prove that the dynamics of this controlled Markov process $\big((A_{1:t-1}, X^i_t), A^i_t\big)_{t\in\mathcal{T}}$ under $(s^i g^{-i})$ are same as under g i.e. $\mathbb{P}^{s^i g^{-i}}(x^i_t, x^i_{t+1}, a_{1:t}) = \mathbb{P}^g(x^i_t, x^i_{t+1}, a_{1:t})$.
(e) In Claim 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. $\mathbb{P}^g(x_t, a_t|a_{1:t-1}, x^i_{1:t}, a^i_t) = \mathbb{P}^{g^{-i}}(x_t, a_t|a_{1:t-1}, x^i_t, a^i_t)$.
(f) From (c), (d) and (e) we then prove the result of the lemma that $\mathbb{P}^{s^i g^{-i}}(x_t, a_t) = \mathbb{P}^g(x_t, a_t)$.

Claim 1: For any policy profile g and $\forall t$,

$$\mathbb{P}^g(x_{1:t}|a_{1:t-1}) = \prod_{i=1}^{N} \mathbb{P}^{g^i}(x^i_{1:t}|a_{1:t-1}) \tag{41}$$

Proof:

$$\mathbb{P}^g(x_{1:t}|a_{1:t-1}) = \frac{\mathbb{P}^g(x_{1:t}, a_{1:t-1})}{\sum_{\overline{x}_{1:t}} \mathbb{P}^g(\overline{x}_{1:t}, a_{1:t-1})} \tag{42a}$$

$$= \frac{\prod_{i=1}^N \left(Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)\right)}{\sum_{\overline{x}_{1:t}} \prod_{i=1}^N \left(Q^i(\overline{x}_1^i) g_1^i(a_1^i|\overline{x}_1^i) \prod_{n=2}^t Q_n^i(\overline{x}_n^i|\overline{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \overline{x}_{1:n}^i)\right)} \tag{42b}$$

$$= \frac{\prod_{i=1}^N \left(Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)\right)}{\prod_{i=1}^N \left(\sum_{\overline{x}_{1:t}^i} Q^i(\overline{x}_1^i) g_1^i(a_1^i|\overline{x}_1^i) \prod_{n=2}^t Q_n^i(\overline{x}_n^i|\overline{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \overline{x}_{1:n}^i)\right)} \tag{42c}$$

$$= \prod_{i=1}^N \frac{Q_1^i(x_1^i) g_1^i(a_1^i|x_1^i) \prod_{n=2}^t Q_n^i(x_n^i|x_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, x_{1:n}^i)}{\sum_{\overline{x}_{1:t}^i} Q^i(\overline{x}_1^i) g_1^i(a_1^i|\overline{x}_1^i) \prod_{n=2}^t Q_n^i(\overline{x}_n^i|\overline{x}_{n-1}^i, a_{n-1}) g_n^i(a_n^i|a_{1:n-1}, \overline{x}_{1:n}^i)} \tag{42d}$$

$$= \prod_{i=1}^N \mathbb{P}^{g^i}(x_{1:t}^i|a_{1:t-1}) \tag{42e}$$

■

Claim 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:

$$\mathbb{P}^g(\tilde{a}_{1:t}, x_{t+1}^i|a_{1:t-1}, x_{1:t}^i, a_{1:t}^i)$$

$$= \sum_{x_{1:t}^{-i}} \mathbb{P}^g(\tilde{a}_{1:t}, x_{t+1}^i, x_{1:t}^{-i}|a_{1:t-1}, x_{1:t}^i, a_t^i) \tag{43a}$$

$$= \sum_{x_{1:t}^{-i}} \mathbb{P}^g(\tilde{a}_t^{-i}, x_{t+1}^i, x_{1:t}^{-i}|a_{1:t-1}, x_{1:t}^i, a_t^i) I_{(a_{1:t-1}, a_t^i)}(\tilde{a}_{1:t-1}, \tilde{a}_t^i) \tag{43b}$$

$$= \sum_{x_{1:t}^{-i}} \mathbb{P}^{g^{-i}}(x_{1:t}^{-i}|a_{1:t-1}) \left(\prod_{j \neq i} g_t^j(\tilde{a}_t^j|a_{1:t-1}, x_{1:t}^j)\right) Q_t^i(x_{t+1}^i|x_t^i, a_t^i, \tilde{a}_t^{-i}) I_{(a_{1:t-1}, a_t^i)}(\tilde{a}_{1:t-1}, \tilde{a}_t^i) \tag{43c}$$

$$= \mathbb{P}^{g^{-i}}(\tilde{a}_{1:t}, x_{t+1}^i|a_{1:t-1}, x_t^i, a_t^i), \tag{43d}$$

where (43c) follows from Claim 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,

$$s_t^i(a_t^i|a_{1:t-1}, x_t^i) \stackrel{\triangle}{=} \mathbb{P}^g(a_t^i|a_{1:t-1}, x_t^i) \tag{44a}$$

$$= \frac{\sum_{x_{1:t-1}^i} \mathbb{P}^g(a_t^i, x_{1:t}^i|a_{1:t-1})}{\sum_{\tilde{a}_t^i} \sum_{\tilde{x}_{1:t-1}^i} \mathbb{P}^g(\tilde{a}_t^i, \tilde{x}_{1:t-1}^i x_t^i|a_{1:t-1})} \tag{44b}$$

$$= \frac{\sum_{x_{1:t-1}^i} \mathbb{P}^{g^i}(x_{1:t}^i|a_{1:t-1}) g_t^i(a_t^i|a_{1:t-1}, x_{1:t}^i)}{\sum_{\tilde{a}_t^i} \sum_{\tilde{x}_{1:t-1}^i} \mathbb{P}^{g^i}(\tilde{x}_{1:t-1}^i x_t^i|a_{1:t-1}) g_t^i(\tilde{a}_t^i|a_{1:t-1}, \tilde{x}_{1:t-1}^i x_t^i)} \tag{44c}$$

$$= \mathbb{P}^{g^i}(a_t^i|a_{1:t-1}, x_t^i), \tag{44d}$$

where dependence of (44c) on only g^i is due to Claim 1.

Claim 3: The dynamics of the Markov process $\{(A_{1:t-1}, X_t^i), A_t^i\}_t$ under $(s^i g^{-i})$ are the same as under g i.e.,

$$\mathbb{P}^{s^i g^{-i}}(x_t^i, x_{t+1}^i, a_{1:t}) = \mathbb{P}^g(x_t^i, x_{t+1}^i, a_{1:t}) \quad \forall t \tag{45}$$

Proof: We prove this by induction. Clearly,

$$\mathbb{P}^g(x_1^i) = \mathbb{P}^{s^i g^{-i}}(x_1^i) = Q_1^i(x_1^i). \tag{46}$$

Now suppose (45) is true for $t-1$ which also implies that the marginals $\mathbb{P}^g(x_t^i, a_{1:t-1}) = \mathbb{P}^{s^i g^{-i}}(x_t^i, a_{1:t-1})$. Then

$$\begin{aligned}
\mathbb{P}^g(x_t^i, a_{1:t-1}, x_{t+1}^i, a_t) &= \mathbb{P}^g(x_t^i, a_{1:t-1})\mathbb{P}^g(a_t^i|a_{1:t-1}, x_t^i)\mathbb{P}^g(x_{t+1}^i, a_{1:t}|x_t^i, a_{1:t-1}, a_t^i) && \text{(47a)}\\
&= \mathbb{P}^{s^i g^{-i}}(x_t^i, a_{1:t-1})s_t^i(a_t^i|a_{1:t-1}, x_t^i)\mathbb{P}^{g^{-i}}(x_{t+1}^i, a_{1:t}|x_t^i, a_{1:t-1}, a_t^i) && \text{(47b)}\\
&= \mathbb{P}^{s^i g^{-i}}(x_t^i, a_{1:t-1}, x_{t+1}^i, a_t), && \text{(47c)}
\end{aligned}$$

where (47b) is true from induction hypothesis, definition of s^i in (44d) and since $\{(a_{1:t-1}, x_t^i), a_t^i\}_t$ is a controlled Markov process as proved in Claim 2 and its update kernel does not depend on policy g^i.This completes the induction step.

■

Claim 4: For any policy g,

$$\mathbb{P}^g(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_{1:t}^i, a_t^i) = \mathbb{P}^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i). \tag{48}$$

Proof:

$$\mathbb{P}^g(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_{1:t}^i, a_t^i) = I_{x_t^i, a_t^i}(\tilde{x}_t^i, \tilde{a}_t^i)\mathbb{P}^g(\tilde{x}_t^{-i}, \tilde{a}_t^{-i}|a_{1:t-1}, x_{1:t}^i). \tag{49}$$

Now

$$\begin{aligned}
\mathbb{P}^g(\tilde{x}_t^{-i}, \tilde{a}_t^{-i}|a_{1:t-1}, x_{1:t}^i) &= \sum_{\tilde{x}_{1:t-1}^{-i}} \mathbb{P}^g(\tilde{x}_{1:t}^{-i}, \tilde{a}_t^{-i}|a_{1:t-1}, x_{1:t}^i) && \text{(50a)}\\
&= \sum_{\tilde{x}_{1:t-1}^{-i}} \mathbb{P}^g(\tilde{x}_{1:t}^{-i}|a_{1:t-1}, x_{1:t}^i)\left(\prod_{j\neq i} g_t^j(\tilde{a}_t^j|a_{1:t-1}, \tilde{x}_{1:t}^j)\right) && \text{(50b)}\\
&= \sum_{\tilde{x}_{1:t}^{-i}} \mathbb{P}^{g^{-i}}(\tilde{x}_{1:t}^{-i}|a_{1:t-1})\left(\prod_{j\neq i} g_t^j(\tilde{a}_t^j|a_{1:t-1}, \tilde{x}_{1:t}^j)\right) && \text{(50c)}\\
&= \mathbb{P}^{g^{-i}}(\tilde{x}_t^{-i}, \tilde{a}_t^{-i}|a_{1:t-1}) && \text{(50d)}
\end{aligned}$$

where (50c) follows from Claim 1.

Hence

$$\begin{aligned}
\mathbb{P}^g(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_{1:t}^i, a_t^i) &= I_{x_t^i, a_t^i}(\tilde{x}_t^i, \tilde{a}_t^i)\mathbb{P}^{g^{-i}}(\tilde{x}_t^{-i}, \tilde{a}_t^{-i}|a_{1:t-1}) && \text{(51a)}\\
&= \mathbb{P}^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i) && \text{(51b)}
\end{aligned}$$

■

Finally,

$$\begin{aligned}
\mathbb{P}^g(\tilde{x}_t, \tilde{a}_t) &= \sum_{a_{1:t-1}x_{1:t}^i a_t^i} \mathbb{P}^g(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_{1:t}^i, a_t^i)\mathbb{P}^g(a_{1:t-1}, x_{1:t}^i, a_t^i) && \text{(52a)}\\
&= \sum_{a_{1:t-1}x_{1:t}^i, a_t^i} \mathbb{P}^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i)\mathbb{P}^g(a_{1:t-1}, x_{1:t}^i, a_t^i) && \text{(52b)}\\
&= \sum_{a_{1:t-1}x_t^i, a_t^i} \mathbb{P}^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i)\mathbb{P}^g(a_{1:t-1}, x_t^i, a_t^i) && \text{(52c)}\\
&= \sum_{a_{1:t-1}x_t^i, a_t^i} \mathbb{P}^{g^{-i}}(\tilde{x}_t, \tilde{a}_t|a_{1:t-1}, x_t^i, a_t^i)\mathbb{P}^{s^i g^{-i}}(a_{1:t-1}, x_t^i, a_t^i) && \text{(52d)}\\
&= \mathbb{P}^{s^i g^{-i}}(\tilde{x}_t, \tilde{a}_t). && \text{(52e)}
\end{aligned}$$

where (52b) follows from (48) in Claim 4 and (52d) from (45) in Claim 3.

APPENDIX B
PROOF OF LEMMA 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 III. This means actions of the common agent, γ_t^i's are generated probabilistically from ψ^i as $\Gamma_t^i \sim \psi_t^i(\cdot|a_{1:t-1})$, 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 $\pi_t^i(x_t^i) = \mathbb{P}^{\psi^i}(x_t^i|a_{1:t-1})$.

(a) In Claim 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 $\pi_{t+1}^i = F^i(\pi_t^i, \gamma_t^i, a_t)$ and F^i is independent of common agent's policy ψ.
(b) In Claim 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 $\theta_t(d\gamma_t|\pi_t) \triangleq \mathbb{P}^{\psi}(d\gamma_t|\pi_t)$.
(d) In Claim 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. $\mathbb{P}^{\theta}(d\pi_t, d\gamma_t, d\pi_{t+1}) = \mathbb{P}^{\psi}(d\pi_t, d\gamma_t, d\pi_{t+1})$.
(e) In Claim 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. $\mathbb{P}^{\psi}(x_t, a_t|a_{1:t-1}, \gamma_t) = \mathbb{P}(x_t, a_t|\pi_t, \gamma_t)$.
(f) From (c), (d) and (e) we then prove the result of the lemma that $\mathbb{P}^{\psi}(x_t, a_t) = \mathbb{P}^{\theta}(x_t, a_t)$ which is equivalent to $\mathbb{P}^{g}(x_t, a_t) = \mathbb{P}^{m}(x_t, a_t)$, where m is the policy profile of players corresponding to θ.

Claim 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 $\pi_{t+1}^i = F^i(\pi_t^i, \gamma_t^i, a_t)$ and F^i is independent of common agent's policy ψ. We also say $\underline{\pi}_{t+1} = \underline{F}(\underline{\pi}_t, \gamma_t, a_t)$.

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,

$$\pi_{t+1}(x_{t+1}) = \mathbb{P}^{\psi}(x_{t+1}|a_{1:t}, \gamma_{1:t+1}) \tag{53a}$$
$$= \mathbb{P}^{\psi}(x_{t+1}|a_{1:t}, \gamma_{1:t}) \tag{53b}$$
$$= \frac{\sum_{x_t} \mathbb{P}^{\psi}(x_t, a_t, x_{t+1}|a_{1:t-1}, \gamma_{1:t})}{\sum_{\tilde{x}_{t+1}\tilde{x}_t} \mathbb{P}^{\psi}(\tilde{x}_t, \tilde{x}_{t+1}, a_t|a_{1:t-1}, \gamma_{1:t})} \tag{53c}$$
$$= \frac{\sum_{x_t} \pi_t(x_t) \prod_{i=1}^N \gamma_t^i(a_t^i|x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t)}{\sum_{\tilde{x}_t\tilde{x}_{t+1}} \pi_t(\tilde{x}_t) \prod_{i=1}^N \gamma_t^i(a_t^i|\tilde{x}_t^i) Q_t^i(\tilde{x}_{t+1}^i|\tilde{x}_t^i, a_t)} \tag{53d}$$
$$= \prod_{i=1}^N \frac{\sum_{x_t^i} \pi_t^i(x_t^i)\gamma_t^i(a_t^i|x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t)}{\sum_{\tilde{x}_t^i} \pi_t^i(\tilde{x}_t^i)\gamma_t^i(a_t^i|\tilde{x}_t^i)} \tag{53e}$$
$$= \prod_{i=1}^N \pi_{t+1}^i(x_{t+1}^i), \tag{53f}$$

where (53e) follows from induction hypothesis. It is assumed in (53c)-(53e) that the denominator is not 0. If denominator corresponding to any γ_t^i is zero, we define

$$\pi_{t+1}^i(x_{t+1}^i) = \sum_{x_t^i} \pi_t^i(x_t^i) Q_t^i(x_{t+1}^i|x_t^i, a_t), \tag{54}$$

where π_{t+1} still satisfies (53f). Thus $\pi^i_{t+1} = F^i(\pi^i_t, \gamma^i_t, a_t)$ and $\underline{\pi}_{t+1} = \underline{F}(\underline{\pi}_t, \gamma_t, a_1)$ where F^i and $\underline{F}$ are appropriately defined from above. ■

Claim 6: $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ is a controlled Markov process with state Π_t and control action Γ_t

Proof:

$$\mathbb{P}^\psi(d\pi_{t+1}|\pi_{1:t}, \gamma_{1:t}) = \sum_{a_t, x_t} \mathbb{P}^\psi(d\pi_{t+1}, a_t, x_t|\pi_{1:t}, \gamma_{1:t}) \tag{55a}$$

$$= \sum_{a_t, x_t} \mathbb{P}^\psi(x_t|\pi_{1:t}, \gamma_{1:t}) \left\{ \prod_{i=1}^N \gamma^i_t(a^i_t|x^i_t) \right\} I_{F(\pi_t, \gamma_t, a_t)}(\pi_{t+1}) \tag{55b}$$

$$= \sum_{a_t, x_t} \pi_t(x_t) \left\{ \prod_{i=1}^N \gamma^i_t(a^i_t|x^i_t) \right\} I_{F(\pi_t, \gamma_t, a_t)}(\pi_{t+1}) \tag{55c}$$

$$= \mathbb{P}(d\pi_{t+1}|\pi_t, \gamma_t). \tag{55d}$$

■

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

$$\theta_t(d\gamma_t|\pi_t) \triangleq \mathbb{P}^\psi(d\gamma_t|\pi_t). \tag{56}$$

Claim 7:

$$\mathbb{P}^\psi(d\pi_t, d\gamma_t, d\pi_{t+1}) = \mathbb{P}^\theta(d\pi_t, d\gamma_t, d\pi_{t+1}) \quad \forall t \in \mathcal{T}. \tag{57}$$

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

$$\mathbb{P}^\psi(d\pi_1) = \mathbb{P}^\theta(d\pi_1) = I_Q(\pi_1). \tag{58}$$

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

$$\mathbb{P}^\psi(d\pi_t, d\gamma_t, d\pi_{t+1}) = \mathbb{P}^\psi(d\pi_t) P^\psi(d\gamma_t|\pi_t) \mathbb{P}^\psi(d\pi_{t+1}|\pi_t\gamma_t) \tag{59a}$$

$$= \mathbb{P}^\theta(d\pi_t) \theta_t(d\gamma_t|\pi_t) P(d\pi_{t+1}|\pi_t, \gamma_t) \tag{59b}$$

$$= \mathbb{P}^\theta(d\pi_t, d\gamma_t, d\pi_{t+1}). \tag{59c}$$

where (59b) is true from induction hypothesis, definition of θ in (56) and since $(\Pi_t, \Gamma_t)_{t\in\mathcal{T}}$ is a controlled Markov process as proved in Claim 6 and thus its update kernel does not depend on policy ψ. This completes the induction step. ■

Claim 8: For any policy ψ,

$$\mathbb{P}^\psi(x_t, a_t|a_{1:t-1}, \gamma_t) = \mathbb{P}(x_t, a_t|\pi_t, \gamma_t). \tag{60}$$

Proof:

$$\mathbb{P}^\psi(x_t, a_t|a_{1:t-1}, \gamma_t) = \mathbb{P}^\psi(x_t|a_{1:t-1}, \gamma_t) \prod_{i\in\mathcal{N}} \gamma^i_t(a^i_t|x^i_t) \tag{61a}$$

$$= \pi_t(x_t) \prod_{i\in\mathcal{N}} \gamma^i_t(a^i_t|x^i_t) \tag{61b}$$

$$= \mathbb{P}(x_t, a_t|\pi_t, \gamma_t). \tag{61c}$$

■

Finally,

$$\begin{aligned}
\mathbb{P}^{\psi}(x_t, a_t) &= \sum_{a_{1:t-1}, \gamma_t} \mathbb{P}^{\psi}(x_t, a_t | a_{1:t-1}, \gamma_t) \mathbb{P}^{\psi}(a_{1:t-1}, \gamma_t) && \text{(62a)} \\
&= \sum_{a_{1:t-1} \gamma_t} \mathbb{P}(x_t, a_t | \pi_t, \gamma_t) \mathbb{P}^{\psi}(a_{1:t-1}, \gamma_t) && \text{(62b)} \\
&= \sum_{\pi_t, \gamma_t} \mathbb{P}(x_t, a_t | \pi_t, \gamma_t) \mathbb{P}^{\psi}(\pi_t, \gamma_t) && \text{(62c)} \\
&= \sum_{\pi_t, \gamma_t} \mathbb{P}(x_t, a_t | \pi_t, \gamma_t) \mathbb{P}^{\theta}(\pi_t, \gamma_t) && \text{(62d)} \\
&= \mathbb{P}^{\theta}(x_t, a_t). && \text{(62e)}
\end{aligned}$$

where (62b) follows from (60), (62c) is due to change of measure and (62d) follows from (57).

■

APPENDIX C
PROOF OF THEOREM 1

Proof: We prove (20) using induction and the results in Lemma 3, 4 and 5 proved in Appendix D. For base case at $t = T$, $\forall i \in \mathcal{N}, (a_{1:T-1}, x^i_{1:T}) \in \mathcal{H}^i_T, \beta^i$

$$\begin{aligned}
\mathbb{E}^{\beta^{*,i}_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i_T(X_T, A_T) \middle| a_{1:T-1}, x^i_{1:T} \right\} &= V^i_T(\underline{\mu}^*_T[a_{1:T-1}], x^i_T) && \text{(63a)} \\
&\geq \mathbb{E}^{\beta^i_T \beta^{*,-i}_T, \mu^*_T[a_{1:T-1}]} \left\{ R^i_T(X_T, A_T) \middle| a_{1:T-1}, x^i_{1:T} \right\}, && \text{(63b)}
\end{aligned}$$

where (63a) follows from Lemma 5 and (63b) follows from Lemma 3 in Appendix D.

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

$$\begin{aligned}
&\mathbb{E}^{\beta^{*,i}_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \middle| a_{1:t}, x^i_{1:t+1} \right\} && \text{(64a)} \\
&\geq \mathbb{E}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \middle| a_{1:t}, x^i_{1:t+1} \right\}. && \text{(64b)}
\end{aligned}$$

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

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

$$= V^i_t(\underline{\mu}^*_t[a_{1:t-1}],x^i_t) \tag{65a}$$

$$\geq \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\left\{R^i_t(X_t,A_t)+V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t],X^i_{t+1})\big|a_{1:t-1},x^i_{1:t}\right\} \tag{65b}$$

$$= \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\Big\{R^i_t(X_t,A_t)+ \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\Big\} \tag{65c}$$

$$\geq \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\Big\{R^i_t(X_t,A_t)+ \mathbb{E}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T}\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t},X^i_{t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\Big\} \tag{65d}$$

$$= \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\left\{R^i_t(X_t,A_t)+\mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t},X^i_{t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\right\} \tag{65e}$$

$$= \mathbb{E}^{\beta^i_{t:T}\beta^{*,-i}_{t:T}\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},x^i_{1:t}\right\}, \tag{65f}$$

where (65a) follows from Lemma 5, (65b) follows from Lemma 3, (65c) follows from Lemma 5, (65d) follows from induction hypothesis in (64b) and (65e) follows from Lemma 4. Moreover, construction of θ in (13), and consequently definition of β^* in (18) are pivotal for (65e) to follow from (65d).

We note that μ^* satisfies the consistency condition of [5, 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 F^i as in Claim 5 in Appendix B.

■

APPENDIX D
INTERMEDIATE LEMMAS USED IN PROOF OF THEOREM 1

Lemma 3: $\forall t \in \mathcal{T}, i \in \mathcal{N}, (a_{1:t-1}, x^i_{1:t}) \in \mathcal{H}^i_t, \beta^i_t$

$$V^i_t(\underline{\mu}^*_t[a_{1:t-1}],x^i_t) \geq \mathbb{E}^{\beta^i_t\beta^{*,-i}_t,\mu^*_t[a_{1:t-1}]}\left\{R^i_t(X_t,A_t)+V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[a_{1:t-1}],\beta^*_t(\cdot|a_{1:t-1},\cdot),A_t),X^i_{t+1})\big|a_{1:t-1},x^i_{1:t}\right\}. \tag{66}$$

Proof: We prove this Lemma by contradiction.

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

$$\mathbb{E}^{\widehat{\beta}^i_t\beta^{*,-i}_t,\mu^*_t[\widehat{a}_{1:t-1}]}\left\{R^i_t(X_t,A_t)+V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[\widehat{a}_{1:t-1}],\beta^*_t(\cdot|\widehat{a}_{1:t-1},\cdot),A_t),X^i_{t+1})\big|\widehat{a}_{1:t-1},\widehat{x}^i_{1:t}\right\} > V^i_t(\underline{\mu}^*_t[\widehat{a}_{1:t-1}],\widehat{x}^i_t). \tag{67}$$

We will show that this leads to a contradiction.

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

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

$$V^i_t(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \widehat{x}^i_t) \tag{68a}$$

$$= \max_{\gamma^i_t(\cdot|\widehat{x}^i_t)} \mathbb{E}^{\gamma^i_t(\cdot|\widehat{x}^i_t)\beta^{*,-i}_t, \mu^*_t[\widehat{a}_{1:t-1}]} \left\{ R^i_t(\widehat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \beta^*_t(\cdot|\widehat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \big| \widehat{x}^i_t \right\}, \tag{68b}$$

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

$$= \sum_{x^{-i}_t, a_t, x_{t+1}} \left\{ R^i_t(\widehat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \beta^*_t(\cdot|\widehat{a}_{1:t-1}, \cdot), a_t), x^i_{t+1}) \right\} \times$$

$$\mu^{*,-i}_t[\widehat{a}_{1:t-1}](x^{-i}_t) \widehat{\gamma}^i_t(a^i_t|\widehat{x}^i_t) \beta^{*,-i}_t(a^{-i}_t|\widehat{a}_{1:t-1}, x^{-i}_t) Q^i_t(x^i_{t+1}|\widehat{x}^i_t, a_t) \tag{68c}$$

$$= \sum_{x^{-i}_t, a_t, x_{t+1}} \left\{ R^i_t(\widehat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \beta^*_t(\cdot|\widehat{a}_{1:t-1}, \cdot), a_t), x^i_{t+1}) \right\} \times$$

$$\mu^{*,-i}_t[\widehat{a}_{1:t-1}](x^{-i}_t) \widehat{\beta}^i_t(a^i_t|\widehat{a}_{1:t-1}, \widehat{x}^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t|\widehat{a}_{1:t-1}, x^{-i}_t) Q^i_t(x^i_{t+1}|\widehat{x}^i_t, a_t) \tag{68d}$$

$$= \mathbb{E}^{\widehat{\beta}^i_t \beta^{*,-i}_t, \mu^*_t[\widehat{a}_{1:t-1}]} \left\{ R^i_t(\widehat{x}^i_t x^{-i}_t, a_t) + V^i_{t+1}(\underline{F}(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \beta^*_t(\cdot|\widehat{a}_{1:t-1}, \cdot), A_t), X^i_{t+1}) \big| \widehat{a}_{1:t-1}, \widehat{x}^i_{1:t} \right\} \tag{68e}$$

$$> V^i_t(\underline{\mu}^*_t[\widehat{a}_{1:t-1}], \widehat{x}^i_t), \tag{68f}$$

where (68b) follows from definition of V^i_t in (14), (68d) follows from definition of $\widehat{\gamma}^i_t$ and (68f) follows from (67). However this leads to a contradiction. ■

Lemma 4: $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t}, x^i_{1:t+1}) \in \mathcal{H}^i_{t+1}$ and β^i_t

$$\mathbb{E}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \big| a_{1:t}, x^i_{1:t+1} \right\} = \mathbb{E}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \big| a_{1:t}, x^i_{1:t+1} \right\}. \tag{69}$$

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 $\mu^{*,-i}_{t+1}[a_{1:t}] = \prod_{j \neq i} F(\mu^{*,j}_t[a_{1:t-1}], \beta^{*,j}_t, a_t)$ as in Claim 5. Since the above expectations involve random variables $X^{-i}_{t+1}, A_{t+1:T}, X_{t+2:T}$, we consider the probability $\mathbb{P}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t}, x^i_{1:t+1})$.

$$\mathbb{P}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t}, x^i_{1:t+1})$$
$$= \frac{\sum_{x^{-i}_t} \mathbb{P}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(x^{-i}_t, a_t, x_{t+1}, a_{t+1:T}, x_{t+2:T} \big| a_{1:t-1}, x^i_{1:t})}{\sum_{\tilde{x}^{-i}_t} \mathbb{P}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(\tilde{x}^{-i}_t, a_t, x^i_{t+1} \big| a_{1:t-1}, x^i_{1:t})}. \tag{70a}$$

We consider the numerator and the denominator separately. The numerator in (70a) is given by

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

$$\mathbb{P}^{\beta^i_{t:T} \beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]}(a_{t+1:T}, x_{t+2:T}|a_{1:t}, x^i_{1:t-1}, x_{t:t+1}) \tag{70b}$$

$$= \sum_{x^{-i}_t} \mu^{*,-i}_t[a_{1:t-1}](x^{-i}_t) \beta^i_t(a^i_t|a_{1:t-1}, x^i_{1:t}) \beta^{*,-i}_t(a^{-i}_t|a_{1:t-1}, x^{-i}_t) Q^i(x^i_{t+1}|x^i_t, a_t)$$

$$Q^{-i}(x^{-i}_{t+1}|x^{-i}_t, a_t) \mathbb{P}^{\beta^i_{t+1:T} \beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t}]}(a_{t+1:T}, x_{t+2:T}|a_{1:t}, x^i_{1:t}, x_{t+1}), \tag{70c}$$

where (70c) follows from the conditional independence of types given common information, as shown in Claim 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 (70a) is given by

$$Dr = \sum_{\tilde{x}_t^{-i}} \mathbb{P}^{\beta^i_{t:T}\beta^{*,-i}_{t:T},\mu^*_t}(\tilde{x}_t^{-i}|a_{1:t-1},x^i_{1:t})\beta^i_t(a^i_t|a_{1:t-1},x^i_{1:t})\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1},\tilde{x}^{-i}_t)Q^i(x^i_{t+1}|x^i_t,a_t) \tag{70d}$$

$$= \sum_{\tilde{x}_t^{-i}} \mu^{*,-i}_t[a_{1:t-1}](\tilde{x}^{-i}_t)\beta^i_t(a^i_t|a_{1:t-1},x^i_{1:t})\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1},\tilde{x}^{-i}_t)Q^i(x^i_{t+1}|x^i_t,a_t). \tag{70e}$$

By canceling the terms $\beta^i_t(\cdot)$ and $Q^i(\cdot)$ in the numerator and the denominator, (70a) is given by

$$\frac{\sum_{x_t^{-i}} \mu^{*,-i}_t[a_{1:t-1}](x^{-i}_t)\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1},x^{-i}_t)Q^{-i}_{t+1}(x^{-i}_{t+1}|x^{-i}_t,a_t)}{\sum_{\tilde{x}_t^{-i}} \mu^{*,-i}_t[a_{1:t-1}](\tilde{x}^{-i}_t)\beta^{*,-i}_t(a^{-i}_t|a_{1:t-1},\tilde{x}^{-i}_t)}\times$$

$$\mathbb{P}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(a_{t+1:T},x_{t+2:T}|a_{1:t},x^i_{1:t},x_{t+1}) \tag{70f}$$

$$= \mu^{*,-i}_{t+1}[a_{1:t}](x^{-i}_{t+1})\mathbb{P}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(a_{t+1:T},x_{t+2:T}|a_{1:t},x^i_{1:t},x_{t+1}) \tag{70g}$$

$$= \mathbb{P}^{\beta^i_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}(x^{-i}_{t+1},a_{t+1:T},x_{t+2:T}|a_{1:t},x^i_{1:t+1}), \tag{70h}$$

where (70g) follows from using the definition of $\mu^{*,-i}_{t+1}[a_{1:t}](x^{-i}_t)$ in the forward recursive step in (19) and the definition of the belief update in (53).

■

Lemma 5: $\forall i \in \mathcal{N}, t \in \mathcal{T}, (a_{1:t-1}, x^i_{1:t}) \in \mathcal{H}^i_t$,

$$V^i_t(\underline{\mu}^*_t[a_{1:t-1}], x^i_t) = \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},x^i_{1:t}\right\}. \tag{71}$$

Proof:

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

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

$$= \sum_{x^{-i}_T a_T} R^i_T(x_T,a_T)\mu^*_T[a_{1:T-1}](x^{-i}_T)\beta^{*,i}_T(a^i_T|a_{1:T-1},x^i_T)\beta^{*,-i}_T(a^{-i}_T|a_{1:T-1},x^{-i}_T) \tag{72a}$$

$$= V^i_T(\underline{\mu}^*_T[a_{1:T-1}], x^i_T), \tag{72b}$$

where (72b) follows from the definition of V^i_t in (14) and the definition of β^*_T in the forward recursion in (18).

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

$$V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t}], x^i_{t+1}) = \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\mu^*_{t+1}[a_{1:t}]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t},x^i_{1:t+1}\right\}. \tag{73}$$

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

$$\begin{aligned}
&\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t}^{T} R^i_n(X_n, A_n) \big| a_{1:t-1}, x^i_{1:t} \right\} \\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \Big\{ R^i_t(X_t, A_t) + \\
&\qquad \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \big| a_{1:t-1}, A_t, x^i_{1:t}, X^i_{t+1} \right\} \big| a_{1:t-1}, x^i_{1:t} \right\} && \text{(74a)} \\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \Big\{ R^i_t(X_t, A_t) + \\
&\qquad \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T}, \mu^*_{t+1}[a_{1:t-1}, A_t]} \left\{ \sum_{n=t+1}^{T} R^i_n(X_n, A_n) \big| a_{1:t-1}, A_t, x^i_{1:t}, X^i_{t+1} \right\} \big| a_{1:t-1}, x^i_{1:t} \right\} && \text{(74b)} \\
&= \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T}, \mu^*_t[a_{1:t-1}]} \left\{ R^i_t(X_t, A_t) + V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], X^i_{t+1}) \big| a_{1:t-1}, x^i_{1:t} \right\} && \text{(74c)} \\
&= \mathbb{E}^{\beta^{*,i}_{t}\beta^{*,-i}_{t}, \mu^*_t[a_{1:t-1}]} \left\{ R^i_t(X_t, A_t) + V^i_{t+1}(\underline{\mu}^*_{t+1}[a_{1:t-1}A_t], X^i_{t+1}) \big| a_{1:t-1}, x^i_{1:t} \right\} && \text{(74d)} \\
&= V^i_t(\underline{\mu}^*_t[a_{1:t-1}], x^i_t), && \text{(74e)}
\end{aligned}$$

where (74b) follows from Lemma 4 in Appendix D, (74c) follows from the induction hypothesis in (73), (74d) 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 (74e) follows from the definition of β^*_t in the forward recursion in (18), the definition of μ^*_{t+1} in (19) and the definition of V^i_t in (14). ■

APPENDIX E
PROOF OF THEOREM 2

Proof: We prove this by contradiction. Suppose for any equilibrium generating function ϕ that generates (β^*, μ^*) through forward recursion, there exists $t \in \mathcal{T}, i \in \mathcal{N}, a_{1:t-1} \in \mathcal{H}^c_t$, such that for $\underline{\pi}_t = \underline{\mu}_t[a_{1:t-1}]$, (13) is not satisfied for ϕ i.e. for $\tilde{\gamma}^i_t = \phi^i[\underline{\pi}_t] = \beta^{*,i}_t(\cdot|\underline{\mu}_t[a_{1:t-1}], x^i_t)$,

$$\tilde{\gamma}^i_t \notin \arg\max_{\gamma^i_t} \mathbb{E}^{\gamma^i_t(\cdot|x^i)\tilde{\gamma}^{-i}_t, \pi_t} \left\{ R^i_t(X_t, A_t) + V^i_{t+1}(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X^i_{t+1}) \big| x^i_t \right\}. \tag{75}$$

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

$$\begin{aligned}
&\mathbb{E}^{\widehat{\gamma}^i_t(\cdot|x^i)\tilde{\gamma}^{-i}_t, \pi_t} \left\{ R^i_t(X_t, A_t) + V^i_{t+1}(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X^i_{t+1}) \big| x^i_t \right\} \\
&> \mathbb{E}^{\tilde{\gamma}^i_t(\cdot|x^i)\tilde{\gamma}^{-i}_t, \pi_t} \left\{ R^i_t(X_t, A_t) + V^i_{t+1}(\underline{F}(\underline{\pi}_t, \tilde{\gamma}_t, A_t), X^i_{t+1}) \big| x^i_t \right\}
\end{aligned} \tag{76}$$

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

$$\begin{aligned}
&\mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\,\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},x^i_{1:t}\right\}\\
&= \mathbb{E}^{\beta^{*,i}_t\beta^{*,-i}_t,\,\mu^*_t[a_{1:t-1}]}\Big\{R^i_t(X_t,A_t)+\\
&\qquad \mathbb{E}^{\beta^{*,i}_{t:T}\beta^{*,-i}_{t:T},\,\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\right\} &(77)\\
&= \mathbb{E}^{\beta^{*,i}_t\beta^{*,-i}_t,\,\mu^*_t[a_{1:t-1}]}\Big\{R^i_t(X_t,A_t)+\\
&\qquad \mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T},\,\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\right\} &(78)\\
&= \mathbb{E}^{\tilde{\gamma}^i_t(\cdot|x^i_t)\tilde{\gamma}^{-i}_t,\,\pi_t}\left\{R^i_t(X_t,A_t)+V^i_{t+1}(\underline{F}(\underline{\pi}_t,\tilde{\gamma}_t,A_t),X^i_{t+1})\big|x^i_t\right\} &(79)\\
&< \mathbb{E}^{\widehat{\beta}^i_t(\cdot|\underline{\mu}_t[a_{1:t-1}],x^i_t)\tilde{\gamma}^{-i}_t,\,\pi_t}\left\{R^i_t(X_t,A_t)+V^i_{t+1}(\underline{F}(\underline{\pi}_t,\tilde{\gamma}_t,A_t),X^i_{t+1})\big|x^i_t\right\} &(80)\\
&= \mathbb{E}^{\widehat{\beta}^i_t\beta^{*,-i}_t,\,\mu^*_t[a_{1:t-1}]}\Big\{R^i_t(X_t,A_t)\\
&\qquad +\mathbb{E}^{\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t+1:T}\mu^*_{t+1}[a_{1:t-1},A_t]}\left\{\sum_{n=t+1}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},A_t,x^i_{1:t},X^i_{t+1}\right\}\Big|a_{1:t-1},x^i_{1:t}\right\} &(81)\\
&= \mathbb{E}^{\widehat{\beta}^i_t,\beta^{*,i}_{t+1:T}\beta^{*,-i}_{t:T},\,\mu^*_t[a_{1:t-1}]}\left\{\sum_{n=t}^{T} R^i_n(X_n,A_n)\big|a_{1:t-1},x^i_{1:t}\right\}, &(82)
\end{aligned}$$

where (78) follows from Lemma 4, (79) follows from the definitions of $\tilde{\gamma}^i_t$ and $\mu_{t+1}[a_{1:t-1},A_t]$ and Lemma 5, (80) follows from (76) and the definition of $\widehat{\beta}^i_t$, (81) follows from Lemma 3, (82) follows from Lemma 4. However, this leads to a contradiction since (β^*,μ^*) is a PBE of the game. ■

APPENDIX F
PROOF OF THEOREM 3

We divide the proof into two parts: first we show that the value function V^i is at least as big as any reward-to-go function; secondly we show that under the strategy β^*_i, reward-to-go is V^i.

Part 1: For any $i\in\mathcal{N}$, β^i define the following reward-to-go functions

$$W^{i,\beta^i}_t(h^i_t) = \mathbb{E}^{\beta^i,\beta^{-i,*},\mu^*_t[h^c_t]}\big[\sum_{n=t}^{\infty}\delta^{n-t}R^i(X_n,A_n)\mid h^i_t\big] \tag{83a}$$

$$W^{i,\beta^i,T}_t(h^i_t) = \mathbb{E}^{\beta^i,\beta^{-i,*},\mu^*_t[h^c_t]}\big[\sum_{n=t}^{T}\delta^{n-t}R^i(X_n,A_n)+\delta^{T+1-t}V^i(\underline{\Pi}_{T+1},X^i_{T+1})\mid h^i_t\big]. \tag{83b}$$

Since $\mathcal{X}^i,\mathcal{A}^i$ are finite sets the reward R^i is absolutely bounded, the reward-to-go $W^{i,\beta^i}_t(h^i_t)$ is finite $\forall$ i,t,β^i,h^i_t.

For any $i\in\mathcal{N}$, $h^i_t\in\mathcal{H}^i_t$,

$$V^i\big(\underline{\mu}^*_t[h^c_t],x^i_t\big) - W^{i,\beta^i}_t(h^i_t) = \Big(V^i\big(\underline{\mu}^*_t[h^c_t],x^i_t\big) - W^{i,\beta^i,T}_t(h^i_t)\Big) + \Big(W^{i,\beta^i,T}_t(h^i_t) - W^{i,\beta^i}_t(h^i_t)\Big) \tag{84}$$

Combining results from Lemma 8 and 9 (Appendix G), the term in the first bracket in RHS of (84) is non-negative. Using (83), the term in the second bracket is

$$\big(\delta^{T+1-t}\big)\,\mathbb{E}^{\beta^i,\beta^{-i,*},\mu^*_t[h^c_t]}\big[-\sum_{n=T+1}^{\infty}\delta^{n-(T+1)}R^i(X_n,A_n)+V^i(\underline{\Pi}_{T+1},X^i_{T+1})\mid h^i_t\big]. \tag{85}$$

The summation in the expression above is bounded by a convergent geometric series. Also, V^i is bounded. Hence the above quantity can be made arbitrarily small by choosing T appropriately large. Since the LHS of (84) does not depend on T, this results in

$$V^i\big(\underline{\mu}_t^*[h_t^c], x_t^i\big) \geq W_t^{i,\beta^i}(h_t^i). \tag{86}$$

Part 2: Since the strategy β^* generated in (23) is such that $\beta_t^{i,*}$ depends on h_t^i only through $\underline{\mu}_t^*[h_t^c]$ and x_t^i, the reward-to-go $W_t^{i,\beta^{i,*}}$, at strategy β^*, can be written (with abuse of notation) as

$$W_t^{i,\beta^{i,*}}(h_t^i) = W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) = \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[\sum_{n=t}^{\infty} \delta^{n-t} R^i(X_n, A_n) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big]. \tag{87}$$

For any $h_t^i \in \mathcal{H}_t^i$,

$$W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) = \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[R^i(X_t, A_t) + \delta W_{t+1}^{i,\beta^{i,*}}\big(\underline{F}(\underline{\mu}_t^*[h_t^c], \theta[\underline{\mu}_t^*[h_t^c]], A_{t+1}), X_{t+1}^i\big) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big] \tag{88a}$$

$$V^i(\underline{\mu}_t^*[h_t^c], x_t^i) = \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[R^i(X_t, A_t) + \delta V^i\big(\underline{F}(\underline{\mu}_t^*[h_t^c], \theta[\underline{\mu}_t^*[h_t^c]], A_{t+1}), X_{t+1}^i\big) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big]. \tag{88b}$$

Repeated application of the above for the first n time periods gives

$$W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) = \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[\sum_{m=t}^{t+n-1} \delta^{m-t} R^i(X_t, A_t) + \delta^n W_{t+n}^{i,\beta^{i,*}}\big(\underline{\Pi}_{t+n}, X_{t+n}^i\big) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big] \tag{89a}$$

$$V^i(\underline{\mu}_t^*[h_t^c], x_t^i) = \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[\sum_{m=t}^{t+n-1} \delta^{m-t} R^i(X_t, A_t) + \delta^n V^i\big(\underline{\Pi}_{t+n}, X_{t+n}^i\big) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big]. \tag{89b}$$

Here $\underline{\Pi}_{t+n}$ is the $n-$step belief update under strategy and belief prescribed by β^*, μ^*.

Taking differences results in

$$\begin{aligned} W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) - V^i(\underline{\mu}_t^*[h_t^c], x_t^i) \\ = \delta^n \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[W_{t+n}^{i,\beta^{i,*}}\big(\underline{\Pi}_{t+n}, X_{t+n}^i\big) - V^i\big(\underline{\Pi}_{t+n}, X_{t+n}^i\big) \mid \underline{\mu}_t^*[h_t^c], x_t^i\big]. \end{aligned} \tag{90a}$$

Taking absolute value of both sides then using Jensen's inequality for $f(x) = |x|$ and finally taking supremum over h_t^i reduces to

$$\begin{aligned} \sup_{h_t^i} \big|W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) - V^i(\underline{\mu}_t^*[h_t^c], x_t^i)\big| \\ \leq \delta^n \sup_{h_t^i} \mathbb{E}^{\beta^*,\mu_t^*[h_t^c]}\big[\big|W_{t+n}^{i,\beta^{i,*}}(\underline{\Pi}_{t+n}, X_{t+n}^i) - V^i(\underline{\mu}_t^*[h_t^c], x_t^i)\big| \mid \underline{\mu}_t^*[h_t^c], x_t^i\big]. \end{aligned} \tag{91}$$

Now using the fact that W_{t+n}, V^i are bounded and that we can choose n arbitrarily large, we get $\sup_{h_t^i} |W_t^{i,\beta^{i,*}}(\underline{\mu}_t^*[h_t^c], x_t^i) - V^i(\underline{\mu}_t^*[h_t^c], x_t^i)| = 0$.

APPENDIX G
INTERMEDIATE LEMMA USED IN PROOF OF THEOREM 3

In this section, we present four lemmas. Lemma 6 and 7 are intermediate technical results needed in the proof of Lemma 8. Then the results in Lemma 8 and 9 are used in Appendix F for the proof of Theorem 3. The proofs for Lemma 6 and 7 below aren't stated as they are analogous (the only difference being a non-zero terminal reward in the finite horizon model) to the proofs of Lemma 3 and 4, from Appendix D, used in the proof of Theorem 1.

Define the reward-to-go $W_t^{i,\beta^i,T}$ for any agent i and strategy β^i as

$$W_t^{i,\beta^i,T}(h_t^i) = \mathbb{E}^{\beta^i,\beta^{-i,*},\mu_t^*[h_t^c]}\big[\sum_{n=t}^{T} \delta^{n-t} R^i(X_n, A_n) + \delta^{T+1-t} G^i(\underline{\Pi}_{T+1}, X_{T+1}^i) \mid h_t^i\big]. \tag{92}$$

Here agent i's strategy is β^i whereas all other agents use strategy $\beta^{-i,*}$ defined above. Since $\mathcal{X}^i, \mathcal{A}^i$ are assumed to be finite and G^i absolutely bounded, the reward-to-go is finite $\forall\ i, t, \beta^i, h_t^i$.

In the following, any quantity with a T in the superscript refers the finite horizon model with terminal reward G^i. For further discussion, please refer to the comments after the statement of Theorem 3.

Lemma 6: For any $t \in \mathcal{T}$, $i \in \mathcal{N}$, h_t^i and β^i,

$$V_t^{i,T}(\underline{\mu}_t^*[h_t^c], x_t^i) \geq \mathbb{E}^{\beta^i,\beta^{-i,*},\mu_t^*[h_t^c]}\big[R^i(X_t, A_t) + \delta V_{t+1}^{i,T}\big(\underline{F}(\underline{\mu}_t^*[h_t^c], \beta_t^*(\cdot|\underline{\mu}_t^*[h_t^c], \cdot), A_t), X_{t+1}^i\big) \mid h_t^i\big]. \tag{93}$$

Lemma 7:

$$\begin{aligned}
&\mathbb{E}^{\beta_{t+1:T}^i,\beta_{t+1:T}^{-i,*},\mu_{t+1}^*[h_t^c,a_t]}\big[\sum_{n=t+1}^{T} \delta^{n-(t+1)} R^i(X_n, A_n) + \delta^{T+1-t} G^i(\underline{\Pi}_{T+1}, X_{T+1}^i) \mid h_t^i, a_t, x_{t+1}^i\big] \\
&\qquad = \mathbb{E}^{\beta_{t:T}^i,\beta_{t:T}^{-i,*},\mu_t^*[h_t^c]}\big[\sum_{n=t+1}^{T} \delta^{n-(t+1)} R^i(X_n, A_n) + \delta^{T+1-t} G^i(\underline{\Pi}_{T+1}, X_{T+1}^i) \mid h_t^i, a_t, x_{t+1}^i\big].
\end{aligned} \tag{94}$$

The result below shows that the value function from the backwards recursive algorithm is higher than any reward-to-go.

Lemma 8: For any $t \in \mathcal{T}$, $i \in \mathcal{N}$, h_t^i and β^i,

$$V_t^{i,T}(\underline{\mu}_t^*[h_t^c], x_t^i) \geq W_t^{i,\beta^i,T}(h_t^i). \tag{95}$$

Proof: We use backward induction for this. At time T, using the maximization property from (13) (modified with terminal reward G^i),

$$\begin{aligned}
&V_T^{i,T}(\underline{\mu}_T^*[h_T^c], x_T^i) && \text{(96a)}\\
&\triangleq \mathbb{E}^{\tilde{\gamma}_T^{i,T}(\cdot|x_T^i),\tilde{\gamma}_T^{-i,T},\mu_T^*[h_t^c]}\big[R^i(X_T, A_T) + \delta G^i\big(\underline{F}(\underline{\mu}_T^*[h_T^c], \tilde{\gamma}_T^T, A_T), X_{T+1}^i\big) \mid \underline{\mu}_T^*[h_T^c], x_T^i\big] && \text{(96b)}\\
&\geq \mathbb{E}^{\gamma_T^{i,T}(\cdot|x_T^i),\tilde{\gamma}_T^{-i,T},\mu_T^*[h_t^c]}\big[R^i(X_T, A_T) + \delta G^i\big(\underline{F}(\underline{\mu}_T^*[h_T^c], \tilde{\gamma}_T^T, A_T), X_{T+1}^i\big) \mid \underline{\mu}_T^*[h_T^c], x_T^i\big] && \text{(96c)}\\
&= W_T^{i,\beta^i,T}(h_T^i) && \text{(96d)}
\end{aligned}$$

Here the second inequality follows from (13) and (14) and the final equality is by definition in (92).

Assume that the result holds for all $n \in \{t+1, \ldots, T\}$, then at time t we have

$$\begin{aligned}
&V_t^{i,T}(\underline{\mu}_t^*[h_t^c], x_t^i) && \text{(97a)}\\
&\geq \mathbb{E}^{\beta_t^i,\beta_t^{-i,*},\mu_t^*[h_t^c]}\big[R^i(X_t, A_t) + \delta V_{t+1}^{i,T}\big(\underline{F}(\underline{\mu}_t^*[h_t^c], \beta_t^*(\cdot|\underline{\mu}_t^*[h_t^c], \cdot), A_t), X_{t+1}^i\big) \mid h_t^i\big] && \text{(97b)}\\
&\geq \mathbb{E}^{\beta_t^i,\beta_t^{-i,*},\mu_t^*[h_t^c]}\big[R^i(X_t, A_t) + \delta \mathbb{E}^{\beta_{t+1:T}^i,\beta_{t+1:T}^{-i,*},\mu_{t+1}^*[h_t^c,A_t]}\big[\sum_{n=t+1}^{T} \delta^{n-(t+1)} R^i(X_n, A_n) && \text{(97c)}\\
&+ \delta^{T-t} G^i(\underline{\Pi}_{T+1}, X_{T+1}^i) \mid h_t^i, A_t, X_{t+1}^i\big] \mid h_t^i\big] \\
&= \mathbb{E}^{\beta_{t:T}^i,\beta_{t:T}^{-i,*},\mu_t^*[h_t^c]}\big[\sum_{n=t}^{T} \delta^{n-t} R^i(X_n, A_n) + \delta^{T+1-t} G^i(\underline{\Pi}_{T+1}, X_{T+1}^i) \mid h_t^i\big] && \text{(97d)}\\
&= W_t^{i,\beta^i,T}(h_t^i) && \text{(97e)}
\end{aligned}$$

Here the first inequality follows from Lemma 6, the second inequality from the induction hypothesis, the third equality follows from Lemma 7 and the final equality by definition (92). ■

The following result highlights the similarities between the fixed-point equation in infinite horizon and the backwards recursion in the finite horizon.

Lemma 9: Consider the finite horizon game with $G^i \equiv V^i$. Then $V_t^{i,T} = V^i$, $\forall\ i \in \mathcal{N}$, $t \in \{1, \ldots, T\}$ satisfies the backwards recursive construction stated above (adapted from (13) and (14)).

Proof: Use backward induction for this. Consider the finite horizon algorithm at time $t = T$, noting that $V_{T+1}^{i,T} \equiv G^i \equiv V^i$,

$$\tilde{\gamma}_T^{i,T}(\cdot \mid x_T^i) \in \underset{\gamma_T^i(\cdot|x_T^i)\in\Delta(\mathcal{A}^i)}{\operatorname{argmax}}\ \mathbb{E}^{\gamma_T^i(\cdot|x_T^i),\tilde{\gamma}_T^{-i,T},\underline{\pi}_T^{-i}}\big[R^i(X_T, A_T) + \delta V^i\big(\underline{F}(\underline{\pi}_T, \tilde{\gamma}_T^T, A_T), X_{T+1}^i\big) \mid \underline{\pi}_T, x_T^i\big] \tag{98a}$$

$$V_T^{i,T}(\underline{\pi}_T, x_T^i) = \mathbb{E}^{\tilde{\gamma}_T^{i,T}(\cdot|x_T^i),\tilde{\gamma}_T^{-i,T},\underline{\pi}_T^{-i}}\big[R^i(X_T, A_T) + \delta V^i\big(\underline{F}(\underline{\pi}_T, \tilde{\gamma}_T^T, A_T), X_{T+1}^i\big) \mid \underline{\pi}_T, x_T^i\big]. \tag{98b}$$

Comparing the above set of equations with (21), we can see that the pair $(V, \tilde{\gamma})$ arising out of (21) satisfies the above. Now assume that $V_n^{i,T} \equiv V^i$ for all $n \in \{t+1, \ldots, T\}$. At time t, in the finite horizon construction from (13), (14), substituting V^i in place of $V_{t+1}^{i,T}$ from the induction hypothesis, we get the same set of equations as (98). Thus $V_t^{i,T} \equiv V^i$ satisfies it. ∎

APPENDIX H
PROOF OF THEOREM 4

Proof: Denote the vector correspondence defined by the RHS of (39) by

$$\phi(\underline{x}) = \begin{pmatrix} \phi_1(\underline{x}) \\ \vdots \\ \phi_4(\underline{x}) \end{pmatrix} = \begin{pmatrix} \operatorname{argmax}_a\, a f_1(\underline{x}) \\ \vdots \\ \operatorname{argmax}_d\, d f_4(\underline{x}) \end{pmatrix} \tag{99}$$

where $\underline{x} = (x, y, w, z)$. For any $\underline{x} \in [0,1]^4$, $\phi(\underline{x})$ is non-empty and closed, since the argmax solution always exists and is one of $\{0\}, \{1\}, [0,1]$. If in addition ϕ also has a *closed graph* then by Kakutani Fixed Point Theorem there exists a solution to (39).

Consider any sequence $(\underline{x}_n, a_n, b_n, c_n, d_n) \to (\underline{x}_0, a_0, b_0, c_0, d_0)$ such that $\forall\ n \geq 1$,

$$a_n \in \underset{a\in[0,1]}{\operatorname{argmax}}\, a f_1(\underline{x}_n) \tag{100a}$$

$$b_n \in \underset{b\in[0,1]}{\operatorname{argmax}}\, b f_2(\underline{x}_n) \tag{100b}$$

$$c_n \in \underset{c\in[0,1]}{\operatorname{argmax}}\, c f_3(\underline{x}_n) \tag{100c}$$

$$d_n \in \underset{d\in[0,1]}{\operatorname{argmax}}\, d f_4(\underline{x}_n). \tag{100d}$$

We need to verify that (100) also holds for the limit $(\underline{x}_0, a_0, b_0, c_0, d_0)$.

If $\underline{x}_0 \notin \mathcal{D}$ then due to continuity, (100) indeed holds at the limit.

For $\underline{x}_0 \in \mathcal{D}$, for any $i \in S(\underline{x}_0)$ if $f_i(\underline{x}_0) = 0$ then in the relation to be verified, the requirement is either of $a_0, b_0, c_0, d_0 \in [0,1]$, which is always true.

For $\underline{x}_0 \in \mathcal{D}_1 \cap \mathcal{D}_2^{\complement} \cap \mathcal{D}_3^{\complement} \cap \mathcal{D}_4^{\complement}$, if $f_1(\underline{x}_0) > 0$ then for any sequence $\underline{x}_n \to \underline{x}_0$, for large n the points in the sequence are within $B_\epsilon(\underline{x}_0)$ and thus $f_1(\underline{x}_n) > 0$ for large n. This means that the relation from (100) holds at the limit (noting that f_2, f_3, f_4 are continuous at $\underline{x}_0$ in this case).

Similarly if $f_1(\underline{x}_0) < 0$ and for any $\underline{x}_0 \in \mathcal{D}_1^{\complement} \cap \mathcal{D}_2 \cap \mathcal{D}_3^{\complement} \cap \mathcal{D}_4^{\complement}$.

For $\underline{x}_0 \in \mathcal{D}_1 \cap \mathcal{D}_2 \cap \mathcal{D}_3^{\complement} \cap \mathcal{D}_4^{\complement}$ if $f_1(\underline{x}_0) > 0$ and $f_2(\underline{x}_0) < 0$ then there exists an $\epsilon > 0$ such that $\forall\ \underline{x} \in B_\epsilon(\underline{x}_0)$ we have $f_1(\underline{x}) > 0$ and $f_2(\underline{x}) < 0$. From this it follows that the relation (100) holds at the limit. Similar argument works for any other sign combination of f_1, f_2, f_3, f_4.

The two arguments above cover all cases. ∎

REFERENCES

[1] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[2] T. Başar and G. Olsder, *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.

[3] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[4] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*, ser. MIT Press Books. The MIT Press, 1994, vol. 1.

[5] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.

[6] G. J. Mailath and L. Samuelson, *Repeated games and reputations: long-run relationships*. Oxford university press, 2006.

[7] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. observable actions," *Journal of Economic Theory*, vol. 100, no. 2, pp. 191–219, 2001.

[8] R. Ericson and A. Pakes, "Markov-perfect industry dynamics: A framework for empirical work," *The Review of Economic Studies*, vol. 62, no. 1, pp. 53–82, 1995.

[9] D. Bergemann and J. Välimäki, "Learning and strategic pricing," *Econometrica: Journal of the Econometric Society*, pp. 1125–1149, 1996.

[10] D. Acemoğlu and J. A. Robinson, "A theory of political transitions," *American Economic Review*, pp. 938–963, 2001.

[11] U. Doraszelski and A. Pakes, "A framework for applied dynamic analysis in IO," *Handbook of industrial organization*, vol. 3, pp. 1887–1966, 2007.

[12] E. Altman, V. Kambley, and A. Silva, "Stochastic games with one step delay sharing information pattern with application to power control," in *Game Theory for Networks, 2009. GameNets' 09. International Conference on*. IEEE, 2009, pp. 124–129.

[13] 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*, vol. 59, no. 3, pp. 555–570, March 2014.

[14] D. M. Kreps and J. Sobel, "Chapter 25 signalling," ser. Handbook of Game Theory with Economic Applications. Elsevier, 1994, vol. 2, pp. 849 – 867.

[15] M. Spence, "Job market signaling," *The quarterly journal of Economics*, pp. 355–374, 1973.

[16] S. J. Grossman, "The informational role of warranties and private disclosure about product quality," *The Journal of Law & Economics*, vol. 24, no. 3, pp. 461–483, 1981.

[17] C. Wilson, "A model of insurance markets with incomplete information," *Journal of Economic theory*, vol. 16, no. 2, pp. 167–207, 1977.

[18] M. Rothschild and J. Stiglitz, "Equilibrium in competitive insurance markets: An essay on the economics of imperfect information," in *Foundations of Insurance Economics*. Springer, 1976, pp. 355–375.

[19] A. Zahavi, "Mate selection–a selection for a handicap," *Journal of theoretical Biology*, vol. 53, no. 1, pp. 205–214, 1975.

[20] A. V. Banerjee, "A simple model of herd behavior," *The Quarterly Journal of Economics*, pp. 797–817, 1992.

[21] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of Political Economy*, vol. 100, no. 5, pp. pp. 992–1026, 1992. [Online]. Available: http://www.jstor.org/stable/2138632

[22] L. Smith and P. Sörensen, "Pathological outcomes of observational learning," *Econometrica*, vol. 68, no. 2, pp. 371–398, 2000. [Online]. Available: http://dx.doi.org/10.1111/1468-0262.00113

[23] N. R. Devanur, Y. Peres, and B. Sivan, "Perfect Bayesian equilibria in repeated sales," in *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2015, pp. 983–1002.

[24] Y.-C. Ho, "Team decision theory and information structures," *Proceedings of the IEEE*, vol. 68, no. 6, pp. 644–654, 1980.

[25] A. Nayyar, A. Mahajan, and D. Teneketzis, "Optimal control strategies in delayed sharing information structures," *IEEE Trans. Automatic Control*, vol. 56, no. 7, pp. 1606–1620, July 2011.

[26] 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*, vol. 52, no. 5, pp. 3228–3260, 2014.

[27] Y. Ouyang, H. Tavafoghi, and D. Teneketzis, "Dynamic oligopoly games with private Markovian dynamics," in *Proc. 54th IEEE Conf. Decision and Control (CDC)*, 2015.

[28] L. Li and J. Shamma, "Lp formulation of asymmetric zero-sum stochastic games," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 1930–1935.

[29] H. L. Cole and N. Kocherlakota, "Dynamic games with hidden actions and hidden states," *Journal of Economic Theory*, vol. 98, no. 1, pp. 114–126, 2001.

[30] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Transactions on Automatic Control*, vol. 58, no. 7, pp. 1644–1658, July 2013.

[31] A. Mahajan, "Optimal decentralized control of coupled subsystems with control sharing," *Automatic Control, IEEE Transactions on*, vol. 58, no. 9, pp. 2377–2382, 2013.

[32] 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.*, no. 4, pp. 580–595, May 2008.

[33] 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*, Cancun, Mexico, Dec. 2008, pp. 1620–1627.

[34] D. Vasal and A. Anastasopoulos, "Stochastic control of relay channels with cooperative and strategic users," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3434–3446, Oct 2014.

[35] ——, "Signaling equilibria for dynamic LQG games with asymmetric information," in *Proc. IEEE Conf. on Decision and Control*, Dec. 2016, pp. 6901–6908.

[36] ——, "Decentralized Bayesian learning in dynamic games," in *Proc. Allerton Conf. Commun., Control, Comp.*, Sept. 2016.

[37] J. Nash, "Non-cooperative games," *Annals of mathematics*, pp. 286–295, 1951.