# Stochastic Control Approach to Reputation Games

Nuh Aygün Dalkıran and Serdar Yüksel, *Member, IEEE*

***Abstract*—Through a stochastic-control-theoretic approach, we analyze reputation games, where a strategic long-lived player acts in a sequential repeated game against a collection of short-lived players. The key assumption in our model is that the information of the short-lived players is nested in that of the long-lived player. This nested information structure is obtained through an appropriate monitoring structure. Under this monitoring structure, we show that, given mild assumptions, the set of perfect Bayesian equilibrium payoffs coincides with Markov perfect equilibrium payoffs, and hence, a dynamic programming formulation can be obtained for the computation of equilibrium strategies of the strategic long-lived player in the discounted setup. We also consider the undiscounted average-payoff setup, where we obtain an optimal equilibrium strategy of the strategic long-lived player under further technical conditions. We then use this optimal strategy in the undiscounted setup as a tool to obtain a tight upper payoff bound for the arbitrarily patient long-lived player in the discounted setup. Finally, by using measure concentration techniques, we obtain a refined lower payoff bound on the value of reputation in the discounted setup. We also study the continuity of equilibrium payoffs in the prior beliefs.**

***Index Terms*—Game theory, repeated games, incomplete information, signaling games.**

## I. Introduction

REPUTATION plays an important role in long-run relationships. When one considers buying a product from a particular firm, his action (buy/not buy) depends on his belief about this firm, i.e., the firm's reputation, which he has formed based on previous experiences (of himself and of others). Many interactions among rational agents are repeated and are in the form of long-run relationships. This is why game theorists have been extensively studying the role of reputation in long-run relationships and repeated games [37]. By defining reputation as a conceptual as well as a mathematical quantitative variable, game theorists have been able to explain how reputation can rationalize intuitive equilibria, as in the expectation of cooperation in early rounds of a finitely repeated prisoners' dilemma [31], and entry deterrence in the early rounds of the chain store game [32], [39].

Recently, there has been an emergence of use of tools from information and control theory in the reputation literature (see e.g., [15], [16], and [24]). Such tools have been proved to be useful in studying various bounds on the value of reputation.

In this article, by adopting and generalizing recent results from stochastic control theory, we provide a new approach and establish refined results on reputation games. Before stating our contributions and the problem setup more explicitly, we provide a brief overview of the related literature in the following subsection.

### A. Related Literature

Kreps *et al.* [31], [32] and Milgrom and Roberts [39] introduced the adverse selection approach to study reputations in finitely repeated games. Fudenberg and Levine [19], [20] extended this approach to infinitely repeated games and showed that a patient long-lived player facing infinitely many short-lived players can guarantee himself a payoff close to his Stackelberg payoff when there is a slight probability that the long-lived player is a commitment type who always plays the stage game Stackelberg action. When compared to the *folk theorem* [22], [23], their results imply an intuitive expectation: the equilibria with relatively high payoffs are more likely to arise due to reputation effects. Even though the results of Fudenberg and Levine [19], [20] hold for both perfect and imperfect public monitoring, Cripps *et al.* [10] showed that reputation effects are not sustainable in the long run when there is imperfect public monitoring. In other words, under imperfect public monitoring, it is impossible to maintain a permanent reputation for playing a strategy that does not play an equilibrium of the complete information game. There has been further literature, which studies the possibility/impossibility of maintaining permanent reputations (we refer the reader to [2]–[4], [14]–[17], [27], [34], and [40]).

Sorin [43] unified and improved some of the results in reputation literature by using tools from Bayesian learning and merging due to Kalai and Lehrer [29], [30]. Gossner [24] utilized relative entropy (that is, information divergence or Kullback–Leibler

Manuscript received November 5, 2018; revised June 10, 2019; accepted December 10, 2019. Date of publication January 23, 2020; date of current version October 21, 2020. This work was supported in part by the Scientific and Technological Research Council of Turkey and in part by the Natural Sciences and Engineering Research Council of Canada. This article was presented in part at the 2nd Occasional Workshop in Economic Theory at University of Graz, the 69th European Meeting of the Econometric Society, Geneva, Switzerland, and the 11th World Congress of the Econometric Society, Montreal, QC, Canada. Recommended by Associate Editor U. V. Shanbhag. *(Corresponding author: Serdar Yüksel.)*

N. A. Dalkıran is with the Department of Economics, Bilkent University, Ankara 06800, Turkey (e-mail: dalkiran@bilkent.edu.tr).

S. Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: yuksel@mast.queensu.ca).

Digital Object Identifier 10.1109/TAC.2020.2968861

divergence) to obtain bounds on the value of reputations; these bounds coincide in the limit (as the strategic long-lived player becomes arbitrarily patient) with the bounds provided by Fudenberg and Levine [19], [20].

Recently, there have been a number of related results in the information theory and control literature on real-time signaling, which provide powerful structural, topological, and operational results that are, in principle, similar to the reputation models analyzed in the game theory literature, despite the simplifications that come about due to the fact that in these fields, the players typically have a common utility function. Furthermore, such studies typically assume finitely repeated setups, whereas we also consider here infinitely repeated setups, which require nontrivial generalizations (see, e.g., [8], [33], [36], [44]–[47], and [48] for various contexts, but note that all of these studies except [8], [33], and [47] have focused on finite horizon problems).

Using such tools from stochastic control theory and zero-delay source coding, we provide new techniques to study reputations. These techniques not only result in a number of conclusions reaffirming certain results documented in the reputation literature, but also provide new results and interpretations as we briefly discuss in the following.

*Contributions of this article:* Our findings contribute to the reputation literature by obtaining structural and computational results on the equilibrium behavior in finite-horizon, infinite-horizon, and undiscounted settings in sequential reputation games, as well as refined upper and lower bounds on the value of reputations. We analyze reputation games, where a strategic long-lived player acts in a repeated sequential-move game against a collection of short-lived players each of whom plays the stage game only once but observes signals correlated with interactions of the previous short-lived players. The key assumption in our model is that the information of the short-lived players is nested in that of the long-lived player in a causal fashion. This nested information structure is obtained through an appropriate monitoring structure. Under this monitoring structure, we obtain stronger results than what currently exists in the literature in a number of directions.

1) Given mild assumptions, we show that the set of perfect Bayesian equilibrium payoffs coincides with the set of Markov perfect equilibrium payoffs.
2) A dynamic programming formulation is obtained for the computation of equilibrium strategies of the strategic long-lived player in the discounted setup.
3) In the undiscounted setup, under further technical conditions, we obtain an optimal strategy for the strategic long-lived player. In particular, we provide new techniques to investigate the optimality of mimicking a Stackelberg commitment type in the undiscounted setup.
4) The optimal strategy we obtain in the undiscounted setup also lets us obtain, through an Abelian inequality, an upper payoff bound for the arbitrarily patient long-lived player—in the discounted setup. We show that this achievable upper bound is identified with a stage game Stackelberg equilibrium payoff.
5) By using measure concentration techniques, we obtain a refined lower payoff bound on the value of reputation for a fixed discount factor. This lower bound coincides with the lower bounds identified by Fudenberg and Levine [20] and Gossner [24] as the long-lived player becomes arbitrarily patient, i.e., as the discount factor tends to 1.
6) Finally, we establish conditions for the continuity of equilibrium payoffs in the priors.

In the next section, we present preliminaries of our model as well as two motivating examples. Section III provides our structural results leading to the equivalence of perfect Bayesian equilibrium payoffs and Markov perfect equilibrium payoffs in the discounted setup. Section IV provides results characterizing the optimal behavior of the long-lived player for the undiscounted setup, which lead us to an upper bound for the equilibrium payoffs in the discounted setup when the long-lived player becomes arbitrarily patient. Section V studies the continuity problem in the priors. Section VI provides, through an explicit measure concentration analysis, a refined lower bound for the equilibrium payoffs of the strategic long-lived player in the discounted setup.

## II. Model

A long-lived player (Player 1) plays a repeated stage game with a sequence of different short-lived players (each of whom is referred to as Player 2).

*The stage game:* The stage game is a *sequential-move* game: Player 1 moves first; when action $a^1$ is chosen by Player 1 in the stage game; a *public* signal $s^2 \in \mathbb{S}^2$ is observed by Player 2, which is drawn according to the probability distribution $\rho^2(.|a^1) \in \Delta(\mathbb{S}^2)$. Player 2, observing this public signal (and all preceding public signals), moves second. At the end of the stage game, Player 1 observes a *private* signal $s^1 \in \mathbb{S}^1$, which depends on actions of both players in the stage game and is drawn according to the probability distribution $\rho^1(.|(a^1, a^2))$. That is, the stage game can be considered as a Stackelberg game with imperfect monitoring, where Player 1 is the leader and Player 2 is the follower. Action sets of Players 1 and 2 in the stage game are assumed to be finite and denoted by $\mathbb{A}^1$ and $\mathbb{A}^2$, respectively. We also assume that the set of Player 1's all possible private signals, denoted by $\mathbb{S}^1$, and the set of (Player 2s') all possible public signals, denoted by $\mathbb{S}^2$, are finite.

*The information structure:* There is incomplete information regarding the type of the long-lived Player 1. Player 1 can either be a strategic type (or normal type), denoted by $\omega^n$, or one of finitely many simple commitment types. Each of these commitment types is committed to simply playing the same action $\hat{\omega} \in \Delta(\mathbb{A}^1)$ at every stage of the repeated game—independent of the history of the play.[1] The set of all possible commitment types of Player 1 is given by $\hat{\Omega}$. Therefore, the set of all possible types of Player 1 can be denoted as $\Omega = \{\omega^n\} \cup \hat{\Omega}$. The type of Player 1 is determined once and for all at the beginning of the game according to a *common knowledge* and *full-support*

[1] $\Delta(\mathbb{A}^i)$ denotes the set of all probability measures on $\mathbb{A}^i$ for both $i = 1, 2$. That is, the commitment types can be committed to playing mixed stage-game actions as well. We would like to also note here that simple commitment type assumption is a standard assumption in reputation games.

prior $\mu_0 \in \Delta(\Omega)$. Only Player 1 is informed of his type, i.e., Player 1's type is his private information.

We note that there is a *nested information structure* in the repeated game in the following sense: The signals observed by Player 2 s are public and hence available to all subsequent players, whereas Player 1's signals are his private information. Therefore, the information of Player 2 at time $t-1$ is a subset of the information of Player 1 at time $t$. Formally, a generic history for Player 2 at time $t-1$ and a generic history for Player 1 at time $t$ are given as follows:

$$h_{t-1}^2 = (s_0^2, s_1^2, \ldots, s_{t-1}^2) \in H_{t-1}^2 \tag{1}$$

$$h_t^1 = (a_0^1, s_0^1, s_0^2, \ldots, a_{t-1}^1, s_{t-1}^1, s_{t-1}^2) \in H_t^1 \tag{2}$$

where $H_{t-1}^2 := (\mathbb{S}^2)^t$ and $H_t^1 := (\mathbb{A}^1 \times \mathbb{S}^1 \times \mathbb{S}^2)^t$.

That is, each Player 2 observes, before he acts, a finite sequence of public signals, which are correlated with Player 1's action in each of his interaction with preceding Player 2 s. On the other hand, Player 1 observes not only these public signals, but also a sequence of private signals for each particular interaction that happened in the past and his actions in the previous periods—but not necessarily the actions of preceding Player 2 s.[2]

We note also that having such a monitoring structure is not a strong assumption. In particular, it is weaker than the information structure in [20], where it is assumed that only the same sequence of public signals is observable by the long-lived and short-lived players, i.e., there is only public monitoring. Yet, it is stronger than the information structure in [24], which allows private monitoring for both the long-lived and the short-lived players.

The stage game payoff function of the strategic (or normal) type long-lived Player 1 is given by $u^1$, and each short-lived Player 2's payoff function is given by $u^2$, where $u^i : \mathbb{A}^1 \times \mathbb{A}^2 \to \mathbb{R}$. The set of all possible histories for Player 2 of stage $t$ is $H_t^2 = H_{t-1}^2 \times \mathbb{S}^2$, where $H_{t-1}^2 = (\mathbb{S}^2)^t$. On the other hand, the set of all possible histories observable by the long-lived Player 1 prior to stage $t$ is $H_t^1 = (\mathbb{A}^1 \times \mathbb{S}^1 \times \mathbb{S}^2)^t$. It is assumed that $H_0^1 := \emptyset$ and $H_0^2 := \emptyset$, which is the usual convention. Let $\mathcal{H}^1 = \bigcup_{t\geq 0} H_t^1$ be the set of all possible histories of the long-lived Player 1.

A (behavioral) strategy for Player 1 is a map

$$\sigma^1 : \Omega \times \mathcal{H}^1 \to \Delta(\mathbb{A}^1)$$

which satisfies $\sigma^1(\hat{\omega}, h_{t-1}^1) = \hat{\omega}$ for any $\hat{\omega} \in \hat{\Omega}$ and for every $h_{t-1}^1 \in H_{t-1}^1$, since commitment types are required to play the corresponding (fixed) action of the stage game independent of the history. The set of all strategies for Player 1 is denoted by $\Sigma^1$, i.e., $\Sigma^1$ is the set of all functions from $\Omega \times \mathcal{H}^1$ to $\Delta(\mathbb{A}^1)$.

A strategy for Player 2 of stage $t$ is a map

$$\sigma_t^2 : H_{t-1}^2 \times \mathbb{S}^2 \to \Delta(\mathbb{A}^2).$$

We let $\Sigma_t^2$ be the set of all such strategies and let $\Sigma^2 = \Pi_{t\geq 0}\Sigma_t^2$ denote the set of all sequences of all such strategies. A history (or path) $h_t$ of length $t$ is an element of $\Omega \times (\mathbb{A}^1 \times \mathbb{A}^2 \times \mathbb{S}^1 \times \mathbb{S}^2)^t$ describing Player 1's type, actions, and signals realized up to stage $t$. By standard arguments (e.g., Ionescu–Tulcea theorem [25]), a strategy profile $\sigma = (\sigma^1, \sigma^2) \in \Sigma^1 \times \Sigma^2$ induces a unique probability distribution $P_\sigma$ over the set of all paths of play $H^\infty = \Omega \times (\mathbb{A}^1 \times \mathbb{A}^2 \times \mathbb{S}^1 \times \mathbb{S}^2)^{\mathbb{Z}_+}$ endowed with the product $\sigma$-algebra. We let $a_t = (a_t^1, a_t^2)$ represent the action profile realized at stage $t$ and let $s_t = (s_t^1, s_t^2)$ denote the signal profile realized at stage $t$. Given $\omega \in \Omega$, $P_{\omega,\sigma}(.) = P_\sigma(.|\omega)$ represents the probability distribution over all paths of play conditional on Player 1 being type $\omega$. Player 1's discount factor is assumed to be $\delta \in (0,1)$, and hence, the expected discounted average payoff to the strategic (normal type) long-lived Player 1 is given by

$$\pi_1(\sigma) = \mathbb{E}_{P_{\omega^n,\sigma}}(1-\delta)\sum_{t\geq 0}\delta^t u^1(a_t).$$

In all of our results except Lemma III.1, we will assume that Player 2 s are Bayesian rational.[3] Hence, we will restrict attention to perfect Bayesian equilibrium: In any such equilibrium, the strategic Player 1 maximizes his expected discounted average payoff given that the short-lived players play a best response to their expectations according to their updated beliefs (This will be appropriately modified when we consider the undiscounted setup). Each Player 2, playing the stage game only once, will be best-responding to his expectation according to his beliefs, which are updated according to the Bayes' rule.

A strategy of Player 2 s, $\sigma^2$, is a best response to $\sigma^1$ if, for all $t$,

$$\mathbb{E}_{P_\sigma}\left[u^2(a_t^1, a_t^2)|s_{[0,t]}^2\right] \geq \mathbb{E}_{P_\sigma}\left[u^2(a_t^1, a^2)|s_{[0,t]}^2\right] \quad \text{for all } a^2 \in A^2 \ \ (P_\sigma - \text{a.s.})$$

where $s_{[0,t]}^2 = (s_0^2, s_1^2, \ldots, s_t^2)$ denotes the information available to Player 2 at time $t$.

### A. Motivating Example I: The Product Choice Game

Our first example is a simple product choice game, which describes how a strategic player can build up reputation: There is a (long-lived) firm (Player 1) who faces an infinite sequence of different consumers (Player 2 s) with identical preferences. There are two actions available to the firm: $A_1 = \{H, L\}$, where $H$ and $L$ denote exerting *high effort* and *low effort* in the production of its output, respectively. Each consumer also has two possible actions: buying a *high-priced* product, $(h)$, or a *low-priced* product, $(l)$, i.e., $A_2 = \{h, l\}$. Each consumer prefers a high-priced product if the firm exerted high effort and a low-priced product if the firm exerted low effort. The firm is willing to commit to high effort only if the consumers purchase the high-priced product, i.e., the firm's (pure) Stackelberg action—in the stage game—is exerting high level of effort. Therefore, if the level of effort of the firm were observable, each consumer would best reply to the Stackelberg action by buying a high-priced product. However, the choice of effort level of

[2]Note that Player 1 gets to observe the realizations of his earlier possibly mixed actions.

[3]A Bayesian rational Player 2 tries to maximize his expected payoff after updating his beliefs according to the Bayes' rule whenever possible. We also note that Lemma III.1 does not require Bayesian rationality and holds for non-Bayesian Player 2 s, who might underreact or overreact to new (or recent) information as in [13] as well.

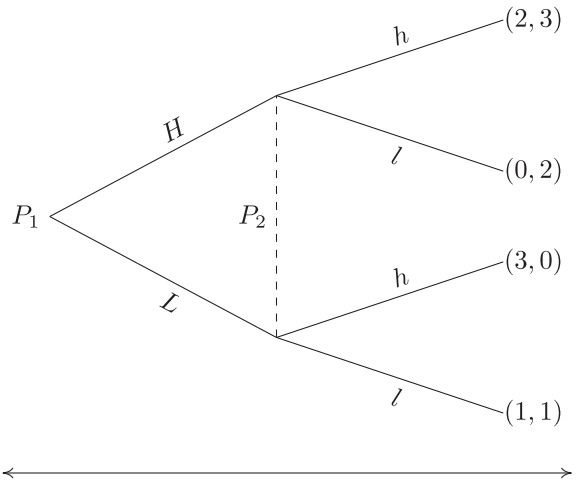


Fig. 1. Illustration of the stage game.

the firm is not observable before consumers choose the product. Furthermore, exerting high effort is costly, and hence, for each type of product, the firm prefers to exert low effort rather than high effort. That is, there is a moral hazard problem.

The corresponding stage game and the preferences regarding the stage game can be illustrated as in Figure 1:

| | $h$ | $l$ |
|---|---|---|
| $H$ | $2,3$ | $0,2$ |
| $L$ | $3,0$ | $1,1$ |

Note that since the stage game is a sequential-move game, where actions are not observable, it is strategically equivalent to a simultaneous-move game represented by the corresponding payoff matrix, which is given above. Furthermore, there is a *unique* Nash equilibrium of this stage game, and in this equilibrium, the firm (the row player) plays $L$ (exerts *low effort*) and the consumer (the column player) plays $l$ (buying a *low-priced* product).

Suppose that there is a small but positive probability $p_0 > 0$ that the firm is an honorable firm who always exerts *high effort*. That is, with $p_0 > 0$ probability, Player 1 is a *commitment type* who plays $H$ at every period of the repeated game—independent of the history. Suppose further that each consumer can observe all the outcomes of the previous play. Yet, before he acts, the consumer cannot observe the effort level of the firm in his own period of play.

Consider now a strategic (noncommitment or normal type) firm who has a discount factor $\delta < 1$: Can the firm build up a reputation that he is (or acts as if he is) an honorable firm? The answer to this question is "Yes"—when he is patient enough.

To see this, observe that a rational consumer (Player 2) would play $h$ only if he anticipates that the firm (Player 1) plays $H$ with a probability of at least $\frac{1}{2}$. Let $p_t$ be the posterior belief that Player 1 is a commitment type after observing some public history $h_t$. Suppose Player 2 of period $t+1$ observes $(H, l)$ as the outcome of the preceding period $t$. This means the probability that Player 2 of period $t$ anticipated for $H$ was less than (or equal to) $\frac{1}{2}$. This probability is $p_t + (1-p_t)\sigma^1(\omega^n, h_t)(H)$, where $\sigma^1(\omega^n, h_t)(H)$ is the probability that the strategic (or normal) type Player 1 assigns to playing $H$ at period $t$ after observing $h_t$. Therefore, we have $p_t + (1-p_t)\sigma^1(\omega^n, h_t)(H) \leq \frac{1}{2}$. But, this implies that the posterior belief of Player 2 of period $t+1$ that Player 1 is a commitment type—after observing $(H, l)$—will be $p_{t+1} = \frac{p_t}{p_t + (1-p_t)\sigma^1(\omega^n, h_t)(H)} \geq 2p_t$. This means every time the strategic player plays $H$, he doubles his reputation, i.e., the belief that he is a commitment type doubles. Therefore, mimicking the commitment type finitely many rounds, the firm can increase the belief that he is an honorable firm (a commitment type) with more than probability $\frac{1}{2}$. In such a case, the short-lived consumers (Player 2 s) will start best replying by buying high-priced products. If the firm is patient enough—when $\delta$ is high—payoffs from those finitely many periods will be negligible. Furthermore, as $\delta \to 1$, one can show that the strategic Player 1 can guarantee himself a discounted average payoff arbitrarily close to 2—which is his payoff under his (pure) Stackelberg action.

### B. Motivating Example II: A Consultant With Reputational Concerns Under Moral Hazard

Our second example presents finer details regarding the nested information structure: A consultant is to advise different firms in different projects. In each of these projects, a supervisor from the particular firm is to inspect the consultant regarding his effort during the particular project. The consultant can either exert a (H)igh level of effort or a (L)ow level of effort while working on the project.

The effort of the consultant is not directly observable to the supervisor. Yet, after the consultant chooses his effort level, the supervisor gets to observe a public signal $s^2 \in \{h, l\}$, which is correlated with the effort level of the consultant according to the probability distribution $\rho^2(h|H) = \rho^2(l|L) = p > \frac{1}{2}$.

Observing this public signal, the supervisor recommends to the upper administration to give the consultant a (B)onus or (N)ot.

The supervisor prefers to recommend a (B)onus when the consultant works hard (exerts (H)igh effort) and (N)ot to recommend a bonus when the consultant shirks (exerts (L)ow effort). For the consultant, exerting a high level of effort is costly. Therefore, the stage game and the preferences regarding the stage game can be illustrated as in Fig. 2. and the following payoff matrix:[4]

| | $B$ | $N$ |
|---|---|---|
| $H$ | $1,1$ | $-1,-1$ |
| $L$ | $2,-2$ | $0,0$ |

It is commonly known that there is a positive probability $p_0 > 0$, with which the consultant is an honorable consultant who always exerts (H)igh level of effort. That is, with $p_0 > 0$ probability, the consultant is a *commitment type* who plays $H$ at every period of the repeated game independent of the history.

Consider the incentives of a strategic (noncommitment or normal type) consultant: Does such a consultant have an incentive to build a reputation by exerting high level of effort, if the game is repeated only finitely many times? What kind of equilibrium behavior would one expect from such a consultant if the game

[4]Note that the stage game is a sequential-move game; the payoffs are summarized in a payoff matrix just for illustrative purposes.

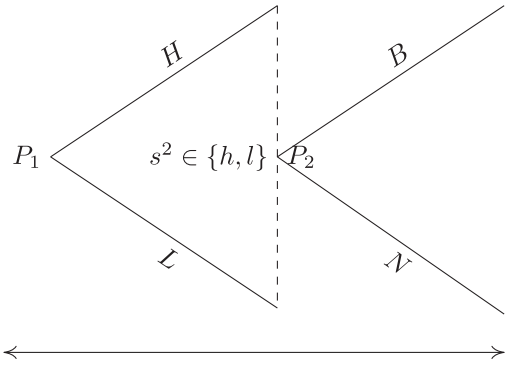


Fig. 2. Illustration of the stage game.

is repeated infinitely many times with discounting for *a fixed discount factor*? For example, if he is building a reputation, how often does he shirk (exert (L)ow level of effort)? Does there exist reputation cycles, i.e., does the consultant build a reputation by exerting high effort for a while and then exploit it by exerting low effort until his reputation level falls under a particular threshold? What happens when the consultant becomes arbitrarily patient, i.e., his discount factor tends to 1? What can we say about the consultant's optimal reputation building strategy when he does not discount the future but rather cares about his undiscounted average payoff?

The aim of this article is to provide tractable techniques to answer similar questions in settings, where agents have reputational concerns in repeated game setups described in our model.

## III. Optimal Strategies and Equilibrium Behavior

Our first set of results will be regarding the optimal strategies of the strategic long-lived Player 1.

Briefly, since each Player 2 plays the stage game only once, we show that when the information of Player 2 is nested in that of Player 1, under a plausible assumption to be noted, the strategic long-lived Player 1 can, without any loss in payoff performance, formulate his strategy as a controlled Markovian system optimization, and thus through dynamic programming. The discounted nature of the optimization problem then leads to the existence of a stationary solution. This implies that for any perfect Bayesian equilibrium, there exists a payoff-equivalent stationary Markov perfect equilibrium. Hence, we conclude that the perfect Bayesian equilibrium payoff set and the Markov perfect equilibrium payoff set of the strategic long-lived Player 1 coincide with each other.

In the following, we provide three results on optimal strategies following steps parallel to [49], which builds on [44]–[46] and [48]. These structural results on optimal strategies will be the key for the following Markov chain construction as well as Theorems III.1 and III.2.

### A. Optimal Strategies: Finite Horizon

We first consider the finitely repeated game setup, where the stage game is to be repeated $T \in \mathbb{N}$ times. In such a case, the strategic long-lived Player 1 is to maximize $\pi_1(\sigma)$ given by

$$\pi_1(\sigma) = \mathbb{E}_{P_{\omega^n,\sigma}}(1-\delta)\sum_{t=0}^{T-1}\delta^t u^1(a_t).$$

Our first result, Lemma III.1, shows that, given any fixed sequence of strategies of the short-lived Player 2 s, any optimal strategy of the strategic long-lived Player 1 can be replaced, without any loss in payoff performance, by another optimal strategy, which only depends on the (public) information of Player 2 s. More specifically, we show that for any private strategy of the long-lived Player 1 against an arbitrary sequence of strategies of Player 2 s, there exists a public strategy of the long-lived Player 1 against the very same sequence of strategies of Player 2 s which gives the strategic long-lived player a better payoff.[5]

To the best of our knowledge, this is a new result in the repeated game literature. What is different here from similar results in the repeated game literature is that this is true even when Player 2 s strategies are non-Bayesian.[6]

Before we state Lemma III.1, we note here that the signal $s_t^2$ that will be available to short-lived Player 2 s after round $t$ only depends on the action of the long-lived Player 1 at round $t$ and that the following holds for all $t \geq 1$:

$$P_\sigma(s_t^2|a_t^1; a_{t'}^1, a_{t'}^2, t' \leq t-1) = P_\sigma(s_t^2|a_t^1). \tag{3}$$

Observation (3) plays an important role in the proof of our first result:

*Lemma III.1* In the finitely repeated setup, given any sequence of strategies of short-lived Player 2 s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $\{s_0^2, s_1^2, \ldots, s_{t-1}^2\}$, which yields the strategic long-lived Player 1 a better payoff against the given sequence of strategies of Player 2 s.

*Proof:* See Appendix A. ■

A brief word of caution is in order. The structural results of the type Lemma III.1, while extremely useful in team theory and zero-delay source coding [49], do not always apply to generic games unless one further restricts the setup. In particular, a generic (Nash) equilibrium may be lost once one alters the strategy structure of one of the players, while keeping the other one fixed (in team problems, the parties can agree to have a better performing team policy even if it is not a strict equilibrium). However, we consider the perfect Bayesian equilibrium concept here, which is of a leader–follower type (i.e., *Stackelberg in the policy space*): Perfect Bayesian equilibrium requires sequential rationality and hence eliminates noncredible threats. That is, Player 2 s respond in a Bayesian fashion to Player 1 who, in turn, is aware of Player 2 s commitment to this policy. This subtle difference is crucial also in signaling games; the features that distinguish Nash equilibria (as in the classical setup

[5] A public strategy is a strategy that uses only public information that is available to all the players. On the other hand, a strategy that is based on private information of a player is referred to as a private strategy. In particular, any strategy of Player 1 that is based on $s_t^1$ for some $t$ is a private strategy.

[6] A relevant result appears in [21], which shows that sequential equilibrium payoffs and perfect public equilibrium payoffs coincide (See [21, Appendix B]) in a similar infinitely repeated game setup.

studied in [9]) from Stackelberg equilibria in signaling games are discussed in detail in [42, Sec. 2].

Lemma III.1 implies that any private information of Player 1 is statistically irrelevant for optimal strategies: for any private strategy of the long-lived Player 1, there exists a public strategy, which performs at least as good as the original one against a given sequence of strategies of Player 2 s. That is, in the finitely repeated setup, the long-lived Player 1 can make his strategy depend only on the public information and his type without any loss in payoff performance. We would like to note here once again that Lemma III.1 above holds for any sequence of strategies of Player 2 s, even non-Bayesian ones.

On the other hand, when Player 2 s are Bayesian rational, as is the norm in repeated games, we obtain a more refined structural result, which we state below as Lemma III.2. As mentioned before, in a perfect Bayesian equilibrium, the short-lived Player 2 at time $t$, playing the stage game only once, seeks to maximize $\sum_{a^1} P_\sigma(a_t^1 = a^1|s^2_{[0,t]})u^2(a^1,a^2)$. However, it may be that his best response set, i.e., the maximizing action set $\arg\max(\sum_{a^1} P_\sigma(a_t^1 = a^1|s^2_{[0,t]})u^2(a^1,a^2))$, may not be unique.

To avoid such set-valued correspondence dynamics, we consider the following assumption, which requires that the best response of each Player 2 is essentially unique: Note that any strategy for Player 2 of time $t$ who chooses

$$\arg\max\left(\sum_{a^1} P_\sigma\left(a_t^1 = a^1|s^2_{[0,t]}\right) u^2\left(a^1,a^2\right)\right)$$

in a measurable fashion does not have to be continuous in the conditional probability $\kappa(\cdot) = P_\sigma(a_t^1 = \cdot|s^2_{[0,t]})$, since such a strategy partitions (or quantizes) the set of probability measures on $\mathbb{A}^1$. The set of $\kappa$, which borders these partitions, is a subset of the set of probability measures $\mathcal{B}_e = \cup_{k,m\in\mathbb{A}^2}\mathcal{B}^{k,m}$, where for any pair $k,m \in \mathbb{A}^2$, the belief set $\mathcal{B}^{k,m}$ is defined as

$$\mathcal{B}^{k,m} = \Bigg\{\kappa \in \Delta(\mathbb{A}^1) : \sum_{a^1\in\mathbb{A}^1} \kappa(a^1)u^2(a^1,k) = \sum_{a^1\in\mathbb{A}^1} \kappa(a^1)u^2(a^1,m)\Bigg\}. \quad (4)$$

These are the sets of probability measures, where Player 2 is indifferent between multiple actions.[7]

*Assumption III.1:* Either of the following holds.

i) The prior measure and the probability space is so that $P_\sigma(P_\sigma(a_t^1 = \cdot|s^2_{[0,t]}) \in \mathcal{B}_e) = 0$ for all $t \geq 0$. In particular, Player 2 s have a unique best response so that the set of discontinuity, $\mathcal{B}_e$, is never visited (with probability 1).

ii) Whenever Player 2 s are indifferent between multiple actions, they choose the action that is better for Player 1.

The following remarks are on Assumption III.1.

*Remark III.1:*

i) In the classical reputation literature, a standard result is that under mild conditions, Bayesian rational short-lived players can be *surprised* at most finitely many times, e.g., [20, Th. 4.1], [43, Lemma 2.4], implying that the jumps in the corresponding belief dynamics of Player 2s will be bounded away from zero in a transient phase until the optimal responses of Player 2s converge to a fixed action. In such cases, the payoff structure can be designed so that the set of discontinuity, $\mathcal{B}_e$, is visited with 0 probability, and hence, Assumption III.1(i) holds.

ii) Assumption III.1(ii) is a standard assumption in the contract theory literature. In a principal-agent model, whenever an agent is indifferent between two actions, he chooses the action that is better for the principal, e.g., when an incentive compatibility condition binds so that the agent is indifferent between exerting a high level of effort and exerting a low-level effort, then the agent chooses to exert the high level of effort (see [5] for further details). Assumption III.1(ii) trivially holds also when the stage game payoff functions are identical for both players (as in *team* setups) or are aligned (as in a *potential game*).

*Lemma III.2:* In the finitely repeated setup, under Assumption III.1, given any arbitrary sequence of strategies of Bayesian rational short-lived Player 2s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $P_\sigma(\omega|s^2_{[0,t-1]}) \in \Delta(\Omega)$ and $t$, which yields the strategic long-lived Player 1 a better payoff against the given sequence of strategies of Player 2s.

*Proof:* See Appendix B. ■

## B. Controlled Markov Chain Construction

The proof of Lemma III.2 reveals the construction of a controlled Markov chain. Building on this proof, we will explicitly construct the dynamic programming problem as a controlled Markov chain optimization problem (that is, a *Markov decision process*). Under Assumption III.1, given any sequence of strategies of Bayesian rational Player 2s, the solution to this optimization problem characterizes the equilibrium behavior of the strategic long-lived player in an associated Markov perfect equilibrium. The state space, the action set, the transition kernel, and the per-stage reward function of the controlled Markov chain mentioned above are given as follows.

1) *The state space* is $\Delta(\Omega)$; $\mu_t \in \Delta(\Omega)$ is often called the *belief* state. We endow this space with the weak convergence topology, and we note that since $\Omega$ is finite, the set of probability measures on $\Omega$ is a compact space.

2) *The action set* is the set of all maps $\Gamma^1 := \{\gamma^1 : \Omega \to \mathbb{A}^1\}$. We note that since the commitment type policies are given *a priori*, one could also regard the action set to be the set $\mathbb{A}^1$ itself.[8]

[7] In particular, in both of our motivating examples, the set $B_e$ is the singleton probability measure $\{(\frac{1}{2},\frac{1}{2})\}$. To see this, it is enough to consider the corresponding payoff matrix for each of the motivating examples. One can verify that in both of the motivating examples, Player 2 becomes indifferent only when Player 1 randomizes between H and L with $\frac{1}{2}$ probability.

[8] We note that randomized strategies may also be considered by adding a randomization variable.

3) *The transition kernel* is given by $P:\Delta(\Omega)\times\Gamma^1\to\mathcal{B}(\Delta(\Omega))$[9] so that for all $B\in\mathcal{B}(\Delta(\Omega))$ as (5) shown at the bottom of this page.
In the above derivation, we use the fact that the term $P_\sigma(a^1_{t-1}|\omega,s^2_{[0,t-2]})$ is uniquely identified by $P_\sigma(\omega|s^2_{[0,t-2]})$ and $\gamma^1_{t-1}$. Here, $\gamma^1_{t-1}$ is the *control action*.
4) *The per-stage reward function*, given $\gamma^2_t$, is $U(\mu_t,\gamma^1):\Delta(\Omega)\times\Gamma^1\to\mathbb{R}$, which is defined as follows:

$$U(\mu_t,\gamma^1):=\sum_{\omega}P_\sigma\left(\omega|s^2_{[0,t-1]}\right)\sum_{\mathbb{A}^1}\left(1_{\{a^1_t=\gamma^1(\omega)\}}u^1(a^1_t,\gamma^2_t(P_\sigma(a^1_t|s^2_{[0,t-1]}),s^2_t))\right) \quad (6)$$

where $\mu_t=P_\sigma(\omega|s^2_{[0,t-1]})$. Here, $\gamma^2_t$ is a given measurable function of the posterior $P_\sigma(a^1_t|s^2_{[0,t]})$. We note again that for each Bayesian rational short-lived Player 2, we have

$$\gamma^2_t(P_\sigma(a^1_t|s^2_{[0,t-1]}),s^2_t))\in\arg\max\left(\sum_{a^1}P_\sigma(a^1_t|s^2_{[0,t]})u^2(a^1,a^2)\right).$$

Lemma III.2 implies that in the finitely repeated setup, under Assumption III.1, when Player 2s are Bayesian rational, the long-lived strategic Player 1 can depend his strategy only on Player 2s' posterior belief and time without any loss in payoff performance.

Consider now any perfect Bayesian equilibrium, where the strategic long-lived Player 1 plays a private strategy; since the strategic long-lived Player 1 cannot have a profitable deviation, the public strategy identified in Lemma III.2 must also give him the same payoff against the given sequence of strategies of Player 2s. Hence, in the finitely repeated setup, under Assumption III.1, any perfect Bayesian equilibrium payoff of the normal type Player 1 is also a *perfect public equilibrium* payoff.[10] Therefore, given our Markov chain construction, we have the following.

*Theorem III.1:* In the finitely repeated game, under Assumption III.1, the set of perfect Bayesian equilibrium payoffs of the strategic long-lived Player 1 is equal to the set of Markov perfect equilibrium payoffs.

[9] $\mathcal{B}(\Delta(\Omega))$ is the set of all Borel sets on $\Delta(\Omega)$.

[10] A perfect public equilibrium is a perfect Bayesian equilibrium, where each player uses a public strategy, i.e., a strategy that only depends on the information which is available to both players.

*Proof:* Markov perfect equilibrium payoff set is a subset of perfect Bayesian equilibrium payoff set. Hence, it is enough to show that for each perfect Bayesian equilibrium, there exists a properly defined Markov perfect equilibrium, which is payoff equivalent for the strategic long-lived Player 1. This follows from Lemma III.2 and our Markov chain construction. ■

Lemmas III.1 and III.2 above have a coding theoretic flavor: The classic works by Witsenhausen [46] and Walrand and Varaiya [45] are of particular relevance; Teneketzis [44] extended these approaches to the more general setting of non-feedback communication, and Yüksel and T. Başar [48], [49] extended these results to more general state spaces (including $\mathbb{R}^d$). Extensions to infinite horizon stages have been studied in [33]. In particular, Lemma III.1 can be viewed as a generalization of [46]. On the other hand, Lemma III.2 can be viewed as a generalization of [33] and [45]. The proofs build on [48]. However, these results are different from the above contributions due to the fact that the utility functions do not depend explicitly on the type of Player 1, but depend explicitly on the actions $a^1_t$, and that these actions are not available to Player 2 unlike the setup in [48]. Next, we consider the infinitely repeated setup in the following.

### C. Infinite Horizon and Equilibrium Strategies

We proceed with Lemma III.3, which is the extension of Lemma III.2 to the infinitely repeated setup. Lemma III.3 will be the key result that gives us a similar controlled Markov chain construction for the infinitely repeated game, hence a payoff-equivalent stationary Markov perfect equilibrium for each perfect Bayesian equilibrium.

*Lemma III.3:* In the infinitely repeated game, under Assumption III.1, given any arbitrary sequence of strategies of Bayesian rational short-lived Player 2s, for any (private) strategy of the strategic long-lived Player 1, there exists a (public) strategy that only conditions on $P_\sigma(\omega|s^2_{[0,t-1]})\in\Delta(\Omega)$ and $t$, which yields the strategic long-lived Player 1 a better payoff against the given sequence of strategies of Player 2s.

Furthermore, the strategic long-lived Player 1's optimal stationary strategy against this given sequence of strategies of Player 2s can be characterized by solving an infinite horizon discounted dynamic programming problem.

*Proof:* See Appendix C. ■

Therefore, in the infinitely repeated setup as well, under Assumption III.1, any private strategy of the normal type Player 1 can be replaced, without any loss in payoff performance, with a public strategy, which only depends on $P_\sigma(\omega|s^2_{[0,t-1]})$ and

$$\begin{aligned}
&P\left(P_\sigma(\omega|s^2_{[0,t-1]})\in B\middle|P_\sigma(\omega|s^2_{[0,t'-1]}),\gamma^1_{t'},t'\le t-1\right)\\
&=P\left(\left\{\frac{\sum_{a^1_{t-1}}P_\sigma(s^2_{t-1}|a^1_{t-1})P_\sigma(a^1_{t-1}|\omega,s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}{\sum_{a^1_{t-1},\omega}P_\sigma(s^2_{t-1}|a^1_{t-1})P_\sigma(a^1_{t-1}|\omega,s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}\right\}\in B\middle|P_\sigma(\omega|s^2_{[0,t'-1]}),\gamma^1_{t'},t'\le t-1\right)\\
&=P\left(\left\{\frac{\sum_{a^1_{t-1}}P_\sigma(s^2_{t-1}|a^1_{t-1})P_\sigma(a^1_{t-1}|\omega,s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}{\sum_{a^1_{t-1},\omega}P_\sigma(s^2_{t-1}|a^1_{t-1})P_\sigma(a^1_{t-1}|\omega,s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}\right\}\in B\middle|P_\sigma(\omega|s^2_{[0,t-2]}),\gamma^1_{t-1}\right) \quad (5)
\end{aligned}$$

$t$. Hence, for any perfect Bayesian equilibrium, there exists a *perfect public equilibrium*, which is payoff equivalent for the strategic long-lived Player 1 in the infinitely repeated game as well.

Furthermore, since there is a stationary optimal public strategy for the strategic long-lived Player 1 against any given sequence of strategies of Bayesian rational Player 2s, any payoff the strategic long-lived Player 1 obtains in a perfect Bayesian equilibrium, he can also obtain in a *Markov perfect equilibrium*.[11]

*Theorem III.2:* In the infinitely repeated game, under Assumption III.1, the set of perfect Bayesian equilibrium payoffs of the strategic long-lived Player 1 is equal to the set of Markov perfect equilibrium payoffs.

*Proof:* The proof follows from Lemma III.3 and our Markov chain construction as in the proof of Theorem III.1. ■

Observe that $\{\mu_t(\bar{\omega}) = \mathbb{E}[1_{\omega=\bar{\omega}}|s^2_{[0,t]}]\}$, for every fixed $\bar{\omega}$, is a bounded martingale sequence adapted to the information at Player 2, and as a result, as $t \to \infty$, by the submartingale convergence theorem [6], there exists (a random) $\bar{\mu}$ such that $\mu_t \to \bar{\mu}$ almost surely. Let $\bar{\mu}$ be an *invariant posterior*, that is, a (sample-path) limit of the $\mu_t$ process. Equation (15) leads to the following fixed point equation:[12]

$$V^1(\omega, \bar{\mu}) = \max_{a^1=\gamma^1_t(\mu,\omega)} (\mathbb{E}[u^1(a^1_t, \gamma^2(\mu)) + \delta \mathbb{E}[V^1[(\omega, \bar{\mu})]).$$

Therefore, we have

$$V^1(\omega, \bar{\mu}) = \frac{1}{1-\delta} \max_{\gamma^1_t} \mathbb{E}[u^1(a^1_t, a^2_t(\bar{\mu}))] \tag{7}$$

and since the solution is asymptotically stationary, the optimal strategy of the strategic long-lived Player 1 when $\mu_0 = \bar{\mu}$ has to be a Stackelberg solution for a Bayesian game with prior $\bar{\mu}$; thus, a perfect Bayesian equilibrium strategy for the strategic long-lived Player 1 has to be mimicking the stage game Stackelberg type forever. This insight will be useful in the following section with further refinements.

## IV. Undiscounted Average Payoff Case and an Upper Payoff Bound for the Arbitrarily Patient Long-Lived Player

We next analyze the setup, where the strategic long-lived Player 1 were to maximize his *undiscounted* average payoff instead of his discounted average payoff. Not only we identify an optimal strategy for the strategic long-lived Player 1 in this setup, but also we establish an *upper payoff bound* for the arbitrarily patient strategic long-lived Player 1 in the standard *discounted* average payoff case—through an Abelian inequality.[13]

[11] A Markov perfect equilibrium is a perfect Bayesian equilibrium, where there is a payoff-relevant state space, and both players are playing Markov strategies that only depend on the state variable.

[12] Equation (15) appears in the proof of Lemma III.3 in Appendix C.

[13] Even though there is a large literature on repeated games with incomplete information in the undiscounted setup, the only papers that we know of that study the reputation games explicitly in the this setup are [11] and [43]. As opposed to our model, [11] analyzes a two-person reputation game, where both of the players are long-lived. On the other hand, [43] unifies results from merging of probabilities, reputation, and repeated games with incomplete information in both discounted and undiscounted setups.

The only difference from our original setup is that the strategic long-lived Player 1 now wishes to maximize

$$\liminf_{N\to\infty} \frac{1}{N} \mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2} \left[ \sum_{t=0}^{N-1} u^1(a^1_t, a^2_t) \right].$$

Therefore, in any perfect Bayesian equilibrium, same as before, the short-lived (Bayesian rational) Player 2s will continue to be best replying to their updated beliefs. On the other hand, the strategic long-lived Player 1 will be playing a strategy, which maximizes his undiscounted average payoff given that each Player 2 will be best replying to their updated beliefs.

The main problem in analyzing the undiscounted setup is that most of the structural coding/signaling results that we have for finite horizon or infinite horizon discounted optimal control problems do not generalize for the undiscounted case, since the construction of controlled Markov chains (which is almost given apriori in stochastic control problems) is based on backwards induction arguments leading to structural results that are applicable only for finite horizon problems.

Let us revisit the discounted setup: Let $\bar{\mu}$ be an *invariant posterior*, that is, a (sample-path) limit of the $\mu_t$ process, which exists by the discussion with regard to the submartingale convergence theorem. Equation (7) is applicable for every $\delta \in (0,1)$ so that

$$(1-\delta)V^1(\omega, \bar{\mu}) = \max_{\gamma^1_t} \mathbb{E}[u^1(a^1_t, a^2_t(\bar{\mu}))] \tag{8}$$

and the optimal strategy of the strategic long-lived Player 1 when $\mu_0 = \bar{\mu}$ is a Stackelberg solution for a Bayesian game with prior $\bar{\mu}$; thus, a perfect Bayesian equilibrium strategy for the strategic long-lived Player 1 has to be mimicking the stage game Stackelberg type forever. In the following, we will identify conditions when the limit $\bar{\mu}$ will turn out to be a dirac delta distribution at the normal type, that is, $\bar{\mu} = \delta_w$ (basically, as in the complete information case). Furthermore, the above discussion implies the following observation: By a direct application of the Abelian inequality (see 17), we have that when $\mu_0 = \bar{\mu}$, we have

$$\begin{aligned} &\sup_{\sigma^1,\sigma^2} \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_{\sigma^1,\sigma^2} \left[ \sum_{m=0}^{N-1} u^1(a^1_m, a^2_m) \right] \\ &\quad \le \limsup_{\delta\to 1} \sup_{\sigma^1,\sigma^2} \mathbb{E}_{\sigma^1,\sigma^2} (1-\delta) \left[ \sum_{m=0}^{\infty} \delta^m u^1(a^1_m, a^2_m) \right] \\ &\quad = \max_{\gamma^1_t} \mathbb{E}[u^1(a^1_t, a^2_t(\bar{\mu}))] \end{aligned} \tag{9}$$

where the last equality follows from (8). In the following, we will elaborate further on these observations and arrive at more refined results. We state the following identifiability assumption.

*Assumption IV.1:* Uniformly over all stationary and optimal (for sufficiently large discount parameters $\delta$) strategies $\tilde{\sigma}^1, \tilde{\sigma}^2$, we have

$$\begin{aligned} \lim_{\delta\to 1} \sup_{\tilde{\sigma}^1,\tilde{\sigma}^2} \Bigg| &\mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2} (1-\delta) \left[ \sum_{t=0}^{\infty} \delta^t u^1(a^1_t, a^2_t) \right] \\ &- \limsup_{N\to\infty} \frac{1}{N} \mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2} \left[ \sum_{t=0}^{N-1} u^1(a^1_t, a^2_t) \right] \Bigg| = 0. \end{aligned} \tag{10}$$

A sufficient condition for Assumption IV.1 is the following.

*Assumption IV.2:* Whenever the strategic long-lived Player 1 adopts a stationary strategy, for any initial commitment prior, there exists a stopping time $\tau$ such that for $t \geq \tau$, Player 2s' posterior beliefs become so that his best response does not change (that is, his best response to his beliefs leads to a constant action). Furthermore, $\mathbb{E}[\tau] < \infty$ uniformly over any stationary strategy $\sigma^1$.

Furthermore, Proposition IV.1 shows that Assumption IV.1 is indeed implied by one of the most standard identifiability assumptions in the repeated games literature.

*Assumption IV.3:* Consider the matrix $A$ whose rows consist of the vectors

$$[P_\sigma(s_t^2 = k|a_t^1 = 1) \quad P_\sigma(s_t^2 = k|a_t^1 = 2) \quad \cdots \quad P_\sigma(s_t^2 = k|a_t^1 = |\mathbb{A}^1|)]$$

where $k \in \{1, 2, \ldots, |\mathbb{S}^2|\}$. We have that $rank(A) = |\mathbb{A}^1|$

*Proposition IV.1:* Under Assumption IV.3, we have

$$\|P_\sigma(a_t^1 \in \cdot|h_t^2) - P_\sigma(a_t^1 \in \cdot|h_t^2, \omega)\|_{TV} \to 0$$

for every $\sigma$. Furthermore, under Assumption IV.3, Assumption IV.1 holds.

*Proof:* See Appendix D. ■

The sufficient condition described in Proposition IV.1 is a standard identifiability assumption, sometimes referred to as the full-rank monitoring assumption in the reputation literature (see, e.g., [10, Assumption 2]). Under Assumption IV.1, we establish that mimicking a Stackelberg commitment type forever is an optimal strategy for the strategic long-lived Player 1 in the undiscounted setup.

*Theorem IV.1:* In the undiscounted setup, under Assumption IV.3, an optimal strategy for the strategic long-lived Player 1 in the infinitely repeated game is the stationary strategy *mimicking the Stackelberg commitment type forever.*

*Proof:* See Appendix E. ■

*Remark IV.1:*

i) We note that we cannot directly use the arguments in [33] with regard to the optimality of Markovian strategies (those given in Lemma III.2) for average-cost/average-payoff problems, since a crucial argument in that article is to establish a nearly optimal coding scheme, which uses the fact that more information cannot hurt both the encoder and the decoder; in our case here, we have a game and the value of (or the lack of) information can be positive or negative in the absence of a further analysis.

ii) Under the conditions noted, it follows that Player 1 cannot *abuse* his reputation in the undiscounted setup: An optimal policy is an *honest* stagewise Stackelberg policy. Abusing (through *exploiting*) the reputation is inherently a discounted optimality phenomenon.

As an implication of Theorem IV.1, we next state the aforementioned upper bound for perfect Bayesian equilibrium payoffs of the arbitrarily patient strategic long-lived Player 1 in the discounted setup as Theorem IV.2.

*Theorem IV.2:* Under the assumptions of Theorem IV.1, we have

$$\limsup_{\delta \to 1} V_\delta^1(\omega, \mu^0) \leq \max_{\alpha_1 \in \Delta(A_1), \alpha_2 \in BR(\alpha_1)} u_1(\alpha_1, \alpha_2).$$

That is, an upper bound for the value of the reputation for an arbitrarily patient strategic long-lived Player 1 in any perfect Bayesian equilibrium of the discounted setup is his stage game Stackelberg equilibrium payoff.

Theorem IV.2 provides an upper bound on the value of reputation for the strategic long-lived Player 1 in the discounted setup. That is, in the discounted setup, an arbitrarily patient strategic long-lived Player 1 cannot do any better than his best Stackelberg payoff under reputational concerns as well. This upper bound coincides with those provided before by Fudenberg and Levine [20] and Gossner [24].

## V. Continuity of Payoff Values

Next, we consider the continuity of the payoff values of the strategic long-lived Player 1 in the prior beliefs of Player 2s for any Markov perfect equilibrium obtained through the aforementioned dynamic programming. In this section, we assume the following.

*Assumption V.1:* Either Assumption III.1(i) holds or the stage game payoff functions are identical for both players.

*Lemma V.1:* The transition kernel of the aforementioned Markov chain is weakly continuous in the (belief) state and action.

*Proof:* See Appendix F. ■

We note that, as in [33], if the game is an identical interest game, the continuity results would follow. By Assumption V.1, the per-stage reward function, $U(\mu_t, \gamma^1)$, is continuous in $\mu_t$. The continuity of the transition kernel and per-stage reward function together with the compactness of the action space leads to the following continuity result.

*Theorem V.1:* Under Assumption V.1, the value function $V_t^1$ of the dynamic program given in (15) is continuous in $\mu_t$ for all $t \geq 0$.[14]

*Proof of Theorem V.1:* Given Lemma V.1 and Assumption III.1(i), the proof follows from an inductive argument and the measurable selection hypothesis. In this case, the discounted optimality operator becomes a contraction mapping from the Banach space of continuous functions on $\Delta(\Omega)$ to itself, leading to a fixed point in this space. ■

Theorem V.1 implies that any Markov perfect equilibrium payoff of the strategic long-lived Player 1 obtained through the dynamic program in (15) is robust to small perturbations in the prior beliefs of Player 2s under Assumption III.1. This further implies that the following conjecture made by Cripps *et al.* [10] is indeed true in our setup: There exists a particular equilibrium in the complete information game and a bound such that for *any* commitment type prior (of Player 2s) less than this bound, there exists an equilibrium of the incomplete information game, where the strategic long-lived Player 1's payoff is arbitrarily close

[14]The dynamic program (15) appears in the proof of Lemma III.3 in Appendix C.

to his payoff from the particular equilibrium in the complete information game.[15] This is also in line with the findings of [12], which uses the methods of [1] to show a similar upper semi continuity result.

For the undiscounted setup, however, in Section IV, we were able to achieve a much stronger continuity result, without requiring Assumption V.1 but instead Assumption IV.3, in addition to the assumptions stated at the beginning of this article. We formally state this result next.

*Theorem V.2:* Under the conditions of Theorem IV.1, the undiscounted average value function does not depend on the prior $\mu_0$.

## VI. Lower Payoff Bound on Reputation Through Measure Concentration

We next identify a lower payoff bound for the value of reputation through an explicit measure concentration analysis. As mentioned before, it was Fudenberg and Levine [19], [20] who provided such a lower payoff bound for the first time, to our knowledge. They constructed a lower bound for any equilibrium payoff of the strategic long-lived player by showing that Bayesian rational short-lived players can be surprised at most finitely many times when a strategic long-lived player mimics a commitment type forever. Using the chain rule property of the concept of relative entropy, Gossner [24] obtained a lower bound for any equilibrium payoff of the strategic long-lived player by showing that any equilibrium payoff of the strategic long-lived player is bounded from below (and above) by a function of the average discounted divergence between the prediction of the short-lived players conditional on the long-lived player's type and its marginal.

Our analysis below provides a sharper lower payoff bound for the value of reputation through a refined measure concentration analysis. To obtain this lower bound, as in [20] as well as [24], we let the strategic long-lived Player 1 mimic (forever) a commitment type, $\hat{\omega}=m$, to investigate the best responses of the short-lived Player 2s. In any perfect Bayesian equilibrium, such a deviation, i.e., deviating to mimicking a particular commitment type forever, is always possible for the strategic long-lived Player 1.

Let $|\Omega|=M$ be the number of all possible types of the long-lived Player 1. We will assume for simplicity that all the types are deterministic, as opposed to the more general mixed types considered earlier in this article. With $m$ being the type mimicked forever by Player 1, we will identify a function $f$ below such that for any $\hat{\omega}\in\hat{\Omega}$ when criterion (11) holds

$$\frac{P_\sigma(\omega=m|s^2_{[0,t]})}{P_\sigma(\omega=\hat{\omega}|s^2_{[0,t]})}\geq f(M) \tag{11}$$

Player 2 of time $t$ will act as if he knew that the type of the long-lived Player 1 is $m$. This will follow from the fact that $\max_{a^2}\sum P_\sigma(\hat{\omega}|s^2_{[0,t]})u^2(a^1,a^2)$ is continuous in $P_\sigma(\hat{\omega}|s^2_{[0,t]})$ and that $P_\sigma(\hat{\omega}|s^2_{[0,t]})$ concentrates around the true type under a mild informativeness condition on the observable variables. Let

$$\tau_m=\left\{t\geq 0:\max_{a^2}\sum_{a^1}P_\sigma(a^1|s^2_{[0,t]})u^2(a^1,a^2)\right.$$
$$\left.=\max_{a^2}\sum_{a^1}P_\sigma(a^1|\omega=m)u^2(a^1,a^2)\right\}.$$

Intuitively, $\tau_m$ is the (random) set of times that Players 2 behave as if the type of the long-lived Player 1 is $m$ as far as their optimal strategies are concerned.

*Lemma VI.1:* Let $\epsilon>0$ be such that for any $\bar{a}^1\in\mathbb{A}^1$ and $\tilde{a}^2,\hat{a}^2\in\mathbb{A}^2$, we have

$$|u^2(\bar{a}^1,\tilde{a}^2)-u^2(\bar{a}^1,\hat{a}^2)|\geq\frac{\epsilon}{1-\epsilon}\left(\max_{a^1,a^2}|u^2(a^1,a^2)|\right).$$

If (11) holds at time $t$ when $f(M)=\frac{(1-\epsilon)}{\epsilon}M$, then $t\in\tau_m$.

*Proof:* See Appendix G. ∎

Lemma VI.1 implies that when criterion (11) holds to be true for $f(M)=\frac{(1-\epsilon)}{\epsilon}M$, at time $t$, any Player 2 of time $t$ and onwards will be best responding to the commitment type $m$. This can be interpreted as the long-lived Player *having a reputation to behave like type* $m$ when criterion (11) is satisfied.

*Theorem VI.1:* Suppose that $0<\frac{P_\sigma(s^2|\omega=m)}{P_\sigma(s^2|\omega=\hat{\omega})}<\infty$ for all $\hat{\omega}\in\hat{\Omega}$ and $s^2\in\mathbb{S}^2$. For all $k\in\mathbb{N}$, $P_\sigma(k\notin\tau_m)\leq R\rho^k$ for some $\rho\in(0,1)$ and $R\in\mathbb{R}$.

*Proof:* See Appendix H. ∎

We are now ready to provide our lower bound for perfect Bayesian equilibrium payoffs of the strategic long-lived Player 1, for a fixed discount factor $\delta\in(0,1)$.

*Theorem VI.2:* A lower bound for the expected payoff of the strategic long-lived Player 1 in any perfect Bayesian equilibrium (in the discounted setup) is given by $\max_{m\in\hat{\Omega}}L(m)$, where

$$L(m)=\mathbb{E}_{\{\omega=m\}}\left[\sum_{k\notin\tau_m}\delta^k u^1(a^1_t,a^2_t)\right]$$
$$+\mathbb{E}_{\{\omega=m\}}\left[\sum_{k\in\tau_m}\delta^k\underline{u}^{1^*}(m)\right]$$

where $\underline{u}^{1^*}(m):=\min_{a^2\in BR^2(m)}u^1(m,a^2)$ and $BR^2(m):=\arg\max_{a^2\in\mathbb{A}^2}u^2(m,a^2)$.

*Proof:* By Theorem VI.1, the discounted average payoff can be lower bounded by the sum of the following two terms:

$$\mathbb{E}_{\{\omega=m\}}\left[\sum_{k\notin\tau_m}\delta^k u^1(a^1_t,a^2_t)\right]+\mathbb{E}_{\{\omega=m\}}\left[\sum_{k\in\tau_m}\delta^k\underline{u}^{1^*}(m)\right]$$

where $\underline{u}^{1^*}(m):=\min_{a^2\in BR^2(m)}u^1(m,a^2)$ and $BR^2(m):=\arg\max_{a^2\in\mathbb{A}^2}u^2(m,a^2)$. Since a deviation to mimicking any of the commitment types forever is available to the strategic long-lived Player 1 in any perfect Bayesian equilibrium, taking the maximum of the lower bound above for all commitment types gives the desired result. ∎

[15] This conjecture appears as a presumption of [10, Th. 3], where Cripps *et al.* write "We conjecture this hypothesis is redundant, given the other conditions of the theorem, but have not been able to prove it."

Observe that when $m$ is a Stackelberg type, i.e., a commitment type who is committed to play the stage game Stackelberg action $\arg\max_{\alpha_1\in\Delta(A_1)} u_1(\alpha_1, BR^2(\alpha^1))$ for which Player 2s have a unique best reply, then

$$\underline{u}^{1^*}(m) = \max_{\alpha_1\in\Delta(A_1),\alpha_2\in BR(\alpha_1)} u_1(\alpha_1,\alpha_2)$$

becomes the stage game Stackelberg payoff.

We next turn to the case of the arbitrarily patient strategic long-lived Player 1. That is what happens when $\delta\to 1$. To emphasize the dependence on $\delta$, we use a superscript in $L^\delta(m)$.

*Theorem VI.3:*

$$\lim_{\delta\to 1}(1-\delta)L^\delta(m) \geq \underline{u}^{1^*}(m).$$

*Proof:* The proof follows from Theorem VI.2 by taking the limit $\delta\to 1$. Since in $\tau_m$, we can bound the payoff to strategic long-lived Player 1 below by the worst possible payoff, and in $\tau_m$, the strategic long-lived Player 1 guarantees the associated Stackelberg payoff, we obtain the desired result by an application of the Abelian inequality. ■

Theorem VI.3 implies that the lower payoff bound that we provided in Theorem VI.2 coincides in the limit as $\delta\to 1$ with those of Fudenberg and Levine [20] and Gossner [24]. That is, if there exists a Stackelberg commitment type, an arbitrarily patient strategic long-lived Player 1 can guarantee himself a payoff arbitrarily close to the associated Stackelberg payoff in every perfect Bayesian equilibrium in the discounted setup.

## VII. Conclusion

In this article, we studied the reputation problem of an informed long-lived player, who controls his reputation against a sequence of uninformed short-lived players by employing tools from stochastic control theory. Our findings contribute to the reputation literature by obtaining new results on the structure of equilibrium behavior in finite-horizon, infinite-horizon, and undiscounted settings, as well as continuity results in the prior probabilities, and improved upper and lower bounds on the value of reputations. In particular, we exhibited that a control-theoretic formulation can be utilized to characterize the equilibrium behavior. Even though there are studies that employed dynamic programming methods to study reputation games in the literature, e.g., [28], these studies restrict themselves directly to Markov strategies—hence to the concept of Markov perfect equilibrium without mentioning its relation to the more general (and possibly more appropriate) concept of perfect Bayesian equilibrium. Under technical assumptions, we have identified that a nested information structure implies the equivalence of the set of Markov perfect equilibrium payoffs and the set of perfect Bayesian equilibrium payoffs. It is our hope that the machinery we provide in this article will open a new avenue for applied work studying reputations in different frameworks.

## Appendix A
## Proof of Lemma III.1

At time $t=T$, the payoff function can be written as follows, where $\gamma_t^2$ denotes a given fixed strategy for Player 2:

$$\mathbb{E}[u^1(a_t^1,\gamma_t^2(s^2_{[0,t]}))|s^2_{[0,t-1]}] = \mathbb{E}[F(a_t^1, s^2_{[0,t-1]}, s_t^2)|s^2_{[0,t-1]}]$$

where, $F(a_t^1, s^2_{[0,t-1]}, s_t^2) = u^1(a_t^1,\gamma_t^2(s^2_{[0,t]}))$.

Now, by a stochastic realization argument (see [7]), we can write $s_t^2 = R(a_t^1, v_t)$ for some independent noise process $v_t$. As a result, the expected payoff conditioned on $s^2_{[0,t-1]}$ is equal to, by the smoothing property of conditional expectation, the following:

$$\mathbb{E}[\mathbb{E}[G(a_t^1, s^2_{[0,t-1]}, v_t)|\omega, a_t^1, s^2_{[0,t-1]}]|s^2_{[0,t-1]}]$$

for some $G$. Since $v_t$ is independent of all the other variables at times $t'\leq t$, it follows that there exists $H$ so that $\mathbb{E}[G(a_t^1, s^2_{[0,t-1]}, v_t)|\omega, a_t^1, s^2_{[0,t-1]}] =: H(\omega, a_t^1, s^2_{[0,t-1]})$. Note that when $\omega$ is a commitment type, $a_t^1$ is fixed quantity or a fixed random variable.

Now, we will apply Witsenhausen's two-stage lemma [46] to show that we can obtain a lower bound for the double expectation by picking $a_t^1$ as a result of a measurable function of $\omega, s^2_{[0,t-1]}$. Thus, we will find a strategy, which only uses $(\omega, s^2_{[0,t-1]})$, which performs as well as one which uses the entire memory available at Player 1. To make this precise, let us fix $\gamma_t^2$ and define for every $k\in\mathbb{A}^1$

$$\beta_k := \{\omega, s^2_{[0,t-1]} : G(\omega, s^2_{[0,t-1]}, k) \leq G(\omega, s^2_{[0,t-1]}, q), \forall q\neq k\}.$$

Such a construction covers the domain set consisting of $(x_t, q_{[0,t-1]})$ but possibly with overlaps. It covers the elements in $\Omega\times\prod_{t=0}^{T-1}\mathbb{S}^2$, since for every element in this product set, there is a maximizing $k\in\mathbb{A}^1$. To avoid the overlap, define a function $\gamma_t^{*,1}$ as

$$q_t = \gamma_t^{*,1}(\omega, s^2_{[0,t-1]}) = k, \text{if}(\omega, \mathrm{s}^2_{[0,t-1]})\in\beta_{\mathrm{k}}\backslash\cup_{\mathrm{i}=1}^{\mathrm{k}-1}\beta_{\mathrm{i}}$$

with $\beta_0=\emptyset$. The new strategy performs at least as well as the original strategy even though it has a restricted structure.

The same discussion applies for earlier time stages as we discuss below. We iteratively proceed to study the other time stages. For a three-stage problem, the payoff at time $t=2$ can be written as

$$\mathbb{E}[u^1(a_2^1,\gamma_2^2(s_1^2,s_2^2)) + \mathbb{E}[u^1(\gamma_3^{*,1}(\omega, s^2_{[1,2]}), \gamma_3^2(s_1^2, s_2^2, R(\gamma_3^{*,1}(\omega, s^2_{[1,2]}), v_3))|\omega, s_1^2, s_2^2]|s_1^2].$$

The expression inside the expectation is equal to for some measurable $F_2$, $F_2(\omega, a_2^1, s_1^2, s_2^2)$. Now, once again expressing $s_2^2 = R(a_2^1, v_2)$, by a similar argument as above, a strategy at time 2, which uses $\omega$ and $s_2^1$ and which performs at least as good as the original strategy, can be constructed. By similar arguments, a strategy that at time $t$, $1\leq t\leq T$, only uses $(\omega, s^2_{[1,t-1]})$ can be constructed. The strategy at time $t=0$ uses $\omega$. ■

## APPENDIX B
## PROOF OF LEMMA III.2

The proof follows from a similar argument as that for Lemma III.1, except that the information at Player 2 is replaced by the sufficient statistic that Player 2 uses: his posterior information. At time $t = T-1$, an optimal Player 2 will use $P_\sigma(a_t^1|s^2_{[0,t]})$ as a sufficient statistic for an optimal decision. Let us fix a strategy for Player 2 at time $t$, $\gamma_t^2$, which only uses the posterior $P_\sigma(a_t^1|s^2_{[0,t]})$ as its sufficient statistic. Let us further note that

$$P_\sigma(a_t^1|s^2_{[0,t]}) = \frac{P_\sigma(s_t^2, a_t^1|s^2_{[0,t-1]})}{\sum_{a_t^1} P_\sigma(s_t^2, a_t^1|s^2_{[0,t-1]})}$$
$$= \frac{\sum_\omega P_\sigma(s_t^2|a_t^1)P_\sigma(a_t^1|\omega, s^2_{[0,t-1]})P_\sigma(\omega|s^2_{[0,t-1]})}{\sum_\omega \sum_{a_t^1} P_\sigma(s_t^2|a_t^1)P_\sigma(a_t^1|\omega, s^2_{[0,t-1]})P_\sigma(\omega|s^2_{[0,t-1]})}. \tag{12}$$

The term $P_\sigma(a_t^1|\omega, s^2_{[0,t-1]})$ is determined by the strategy of Player 1 (this follows from Lemma III.1), $\gamma_t^1$.

As in [49], this implies that the payoff at the last stage conditioned on $s^2_{[0,t-1]}$ is given by

$$\mathbb{E}[u^1(a_t^1, \gamma_t^2(P_\sigma(a_t^1 = \cdot|s^2_{[0,t]})))|s^2_{[0,t-1]}]$$
$$= \mathbb{E}[F(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot|s^2_{[0,t-1]}))|s^2_{[0,t-1]}]$$

where, as earlier, we use the fact that $s_t^2$ is conditionally independent of all the other variables at times $t' \le t$ given $a_t^1$. Let $\gamma_t^{1,s^2_{[0,t-1]}}$ denote the strategy of Player 1. The above state is then equivalent to, by the smoothing property of conditional expectation, the following:

$$\mathbb{E}[\mathbb{E}[F(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot|s^2_{[0,t-1]}))|\omega, \gamma_t^{1,s^2_{[0,t-1]}},$$
$$P_\sigma(\omega = \cdot|s^2_{[0,t-1]}), s^2_{[0,t-1]}]|s^2_{[0,t-1]}]$$
$$= \mathbb{E}[\mathbb{E}[F(a_t^1, \gamma_t^1, P_\sigma(\omega = \cdot|s^2_{[0,t-1]}))|\omega, \gamma^{1,s^2_{[0,t-1]}},$$
$$P_\sigma(\omega = \cdot|s^2_{[0,t-1]})]|s^2_{[0,t-1]}]. \tag{13}$$

The second line follows since once one picks the strategy $\gamma^{1,s^2_{[0,t-1]}}$, the dependence on $s^2_{[0,t-1]}$ is redundant given $P_\sigma(\omega = \cdot|s^2_{[0,t-1]})$.

Now, one can construct an equivalence class among the past $s^2_{[0,t-1]}$ sequences, which induce the same $\mu_t(\cdot) = P_\sigma(\omega \in \cdot|s^2_{[0,t-1]})$, and can replace the strategy in this class with one, which induces a higher payoff among the finitely many elements in each class for the final time stage. An optimal output thus may be generated using $\mu_t$ and $\omega$ and $t$, by extending Witsenhausen's argument used earlier in the proof of Lemma III.1 for the terminal time stage. Since there are only finitely many past sequences and finitely many $\mu_t$, this leads to a (Borel measurable) selection of $\omega$ for every $\mu_t$, leading to a measurable strategy in $\mu_t, \omega$. Hence, the final stage payoff can be expressed as $F_t(\mu_t)$ for some $F_t$, without any performance loss.

The same argument applies for all time stages. To show this, we will apply induction as in [48]. At time $t = T-1$, the sufficient statistic both for the immediate payoff and the *continuation payoff* is $P_\sigma(\omega|s^2_{[0,t-1]})$, and thus for the payoff impacting the time stage $t = T$, as a result of the optimality result for $\gamma_T^1$. To show that the separation result generalizes to all time stages, it suffices to prove that $\{(\mu_t, \gamma_t^1)\}$ has a controlled Markov chain form, if the players use the structure above.

Now, for $t \ge 1$, for all $B \in \mathcal{B}(\Delta(\Omega))$ (14), shown at the bottom of this page, holds.

In the above derivation, we use the fact that the term $P_\sigma(a_{t-1}^1|\omega, s^2_{[0,t-2]})$ is uniquely identified by $P_\sigma(\omega|s^2_{[0,t-2]})$ and $\gamma_{t-1}^1$.

## APPENDIX C
## PROOF OF LEMMA III.3

First, going from a finite horizon to an infinite horizon follows from a change of order of limit and infimum, as we discuss in the following. Observe that for any strategy $\{\gamma_t^1\}$ and any $T \in \mathbb{N}$, we have

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right] \ge \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right]$$

and thus

$$\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right] \ge \limsup_{T\to\infty} \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right].$$

Since the above holds for an arbitrary strategy, it then follows that:

$$\inf_{\{\gamma_t^1\}} \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right]$$
$$\ge \limsup_{T\to\infty} \inf_{\{\gamma_t^1\}} \mathbb{E}\left[\sum_{t=0}^{T-1} \delta^t u^1(a_t^1, a_t^2)\right].$$

On the other hand, due to the discounted nature of the problem, the right-hand side can be studied through the dynamic programming (Bellman) iteration algorithms: The following dynamic

$$P(P_\sigma(\omega|s^2_{[0,t-1]}) \in B|P_\sigma(\omega|s^2_{[0,t'-1]}), \gamma_{t'}^1, t' \le t-1)$$
$$= P\left(\left\{\frac{\sum_{a_{t-1}^1} P_\sigma(s_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}{\sum_{a_{t-1}^1,\omega} P_\sigma(s_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}\right\} \in B|P_\sigma(\omega|s^2_{[0,t'-1]}), \gamma_{t'}^1, t' \le t-1\right)$$
$$= P\left(\left\{\frac{\sum_{a_{t-1}^1} P_\sigma(s_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}{\sum_{a_{t-1}^1,\omega} P_\sigma(s_{t-1}^2|a_{t-1}^1)P_\sigma(a_{t-1}^1|\omega, s^2_{[0,t-2]})P_\sigma(\omega|s^2_{[0,t-2]})}\right\} \in B|P_\sigma(\omega|s^2_{[0,t'-1]}), \gamma_{t'}^1, t' = t-1\right) \tag{14}$$

program holds: Let $\mu_t(w) = P_\sigma(\omega = w|s^2_{[0,t-1]})$

$$V^1(\omega,\mu_t) = \mathbb{T}(V^1)(\omega,\mu_t)$$
$$:= \max_{\gamma^1_t}(\mathbb{E}[u^1(a^1_t,a^2_t) + \delta\mathbb{E}[V^1(\omega,\mu_{t+1})|\mu_t,\gamma^1_t]) \quad (15)$$

where $\mathbb{T}$ is an operator defined by

$$\mathbb{T}(f)(\omega,\mu_t) = \max_{\gamma^1_t}(\mathbb{E}[u^1(a^1_t,a^2_t) + \delta\mathbb{E}[f(\omega,\mu_{t+1})|\mu_t,\gamma^1_t])$$

A value iteration sequence with $V^1_0 = 0$ and $V_{t+1} = \mathbb{T}(V_t)$, which is well defined by the measurable selection conditions noted in [25] due to the finiteness of our action set (and hence continuity of the iterations in the actions), leads to a stationary solution. This is an infinite horizon discounted payoff optimal dynamic programming equation with finite action spaces (where the strategy is now the *action* $\gamma^1_t$). Since the action set is finite in our formulation, it follows that there is a stationary solution as $t \to \infty$. Thus, the sequence of maximizations $\sup_{\gamma^1}\mathbb{E}[\sum_{t=0}^{T-1}\delta^t u^1(a^1_t,a^2_t)]$ leads to a stationary solution as $T \to \infty$, and this sequence of policies admits the structure given in the statement of the theorem. ■

## Appendix D
## Proof of Proposition IV.1

Recall that the chain rule of relative entropy implies the following: For joint measures $P, Q$ on random variables $X, Y$ with finite relative entropy, we have $D(P(X,Y)\|Q(X,Y)) = D(P(X)\|Q(X)) + D(P(Y|X)\|Q(Y|X))$. Let $X = \omega$ and $Y := s^2_{[0,\infty)}$, $P := P_{\sigma,\omega=w}((\omega, s^2_{[0,\infty)}) \in \cdot)$ (i.e., with the true distribution given the type of the long-run player) and $Q := P_\sigma((\omega, s^2_{[0,\infty)}) \in \cdot)$ (this is the distribution seen by Players 2). Then (following [24], see also [38, Sec. 8]), the conditional relative entropies are summable with the bound $D(\delta_\omega|\mu_0) < \infty$, which also implies that

$$\mathbb{E}[D(P_\sigma(s^2_t \in \cdot|h^2_t,\omega)||P_\sigma(s^2_t \in \cdot|h^2_t))] \to 0.$$

From Pinsker's inequality noting that convergence in total variation is implied by convergence in relative entropy, we have

$$\mathbb{E}[||P_\sigma(s^2_t \in \cdot|h^2_t) - P_\sigma(s^2_t \in \cdot|h^2_t,\omega)||^2_{TV}] \to 0 \quad (16)$$

where the expectation is with respect to the true distribution (given the type of the long-run player). But

$$P_\sigma(s^2_t = s|h^2_t) = \sum_{a^1} P_\sigma(s^2_t = s|a^1_t = a^1)P_\sigma(a^1_t = a^1|h^2_t).$$

Thus, all we need to ensure is that Player 2's belief $P_\sigma(a^1_t \in \cdot|h^2_t)$ is sufficiently close to a terminal value. Suppose that the conditions of the theorem hold, but $|P_\sigma(a^1_t|h^2_t) - P_\sigma(a^1_t|h^2_t,\omega)| > \delta$ for some subsequence of time values. If the rank of $A$ is $|\mathbb{A}^1|$, then $|P_\sigma(a^1_t|h^2_t) - P_\sigma(a^1_t|h^2_t,\omega)| > \delta$ would imply that $|P_\sigma(s^2_t|h^2_t) - P_\sigma(s^2_t|h^2_t,\omega)| > \epsilon$ for some positive $\epsilon$, which would be a contradiction (to see this, observe that the vector $P_\sigma(a^1_t \in \cdot|h^2_t) - P_\sigma(a^1_t \in \cdot|h^2_t,\omega)$ cannot be orthogonal to each of the rows of $A$, due to the rank condition). In particular, (16) implies the convergence of $P_\sigma(a^1_t \in \cdot|h^2_t) - P_\sigma(a^1_t \in \cdot|h^2_t,\omega)$ to zero: the summability (bounded from above) of the conditional relative entropies implies that the expected number of instances where the error between the conditional probabilities is above any specified amount will be finite (uniform over all policies).

Now, using uniform continuity of the per-stage utility in the posterior of player 2 (e.g., through a related result from [24]), we can uniformly bound the error in the per-stage from the setup when the posterior seen by Player 2 is exactly $P_\sigma(a^1_t \in \cdot|h^2_t,\omega)$ (where crucially the error is uniform over all posteriors, regardless of the strategy of Player 1). In particular, the payoff into the future would be so that it would be within the payoff for the setup when the posterior of player 2 would correspond to having the prior $\delta_w$ on the normal type, as in the complete information case, for any considered normal policy (which in the statement of Assumption IV.1 is a stationary policy). On the other hand, we know, by the analysis in (8), that any optimal stationary policy with prior $\delta_w$ will be a stagewise Stackelberg policy, and the average payoff [the right-hand side of (20)] in this case will correspond exactly to

$$(1-\delta)V^1(\omega,\delta_w) = \max_{\gamma^1_t}\mathbb{E}[u^1(a^1_t,a^2_t(\delta_w))].$$

Together with the uniformity (over strategies) of the relative entropy bound, we conclude that Assumption IV.1 holds. ■

## Appendix E
## Proof of Theorem IV.1

Note the following *Abelian* inequalities (see, e.g., [25, Lemma 5.3.1]): Let $a_n$ be a sequence of nonnegative numbers and $\beta \in (0,1)$. Then, we have

$$\liminf_{N\to\infty}\frac{1}{N}\sum_{m=0}^{N-1}a_m \le \liminf_{\beta\uparrow 1}(1-\beta)\sum_{m=0}^{\infty}\beta^m a_m$$
$$\le \limsup_{\beta\uparrow 1}(1-\beta)\sum_{m=0}^{\infty}\beta^m a_m \le \limsup_{N\to\infty}\frac{1}{N}\sum_{m=0}^{N-1}a_m. \quad (17)$$

Thus, for every strategy pair $\sigma^1, \sigma^2$, and $\epsilon > 0$, there exists $\delta_\epsilon$ (depending possibly on the strategies) so that

$$\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}(1-\delta_\epsilon)\left[\sum_{m=0}^{\infty}\beta^m_\epsilon u^1(a^1_m,a^2_m)\right] + \epsilon$$
$$\ge \liminf_{N\to\infty}\frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\left[\sum_{m=0}^{N-1}u^1(a^1_m,a^2_m)\right].$$

Now, let $\sigma^1_n$ and $\sigma^2_n$ be a sequence of strategies, which converge to the supremum for the average payoff. Let $\tilde{\sigma}^1_n$ and $\tilde{\sigma}^2_n$ be one, which comes within $\epsilon/2$ of the supremum so that

$$\sup_{\sigma^1,\sigma^2}\liminf_{N\to\infty}\frac{1}{N}\mathbb{E}_{\sigma^1,\sigma^2}\left[\sum_{m=0}^{N-1}u^1(a^1_m,a^2_m)\right]$$
$$\le \liminf_{N\to\infty}\frac{1}{N}\mathbb{E}_{\tilde{\sigma}^1_n,\tilde{\sigma}^2_n}\left[\sum_{m=0}^{N-1}u^1(a^1_m,a^2_m)\right] + \epsilon/2.$$

Now, let $\delta_\epsilon$ close to 1 be a discount factor, whose optimal payoff comes within $\epsilon/2$ of the limit when $\delta = 1$. For this parameter, under $\tilde{\sigma}^1_n$ and $\tilde{\sigma}^2_n$, one obtains an upper bound on this payoff, which can be further upper bounded by optimizing over all possible strategies for this $\delta_\epsilon$ value. This leads to a stationary strategy. Thus, we have

$$\begin{aligned}
&\sup_{\sigma^1,\sigma^2} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}_{\sigma^1,\sigma^2}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right] - \epsilon/2 \\
&\leq \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}_{\tilde{\sigma}^1_n,\tilde{\sigma}^2_n}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right]\Bigg] \\
&\leq \mathbb{E}_{\tilde{\sigma}^1_n,\tilde{\sigma}^2_n}(1-\delta_\epsilon)\left[\sum_{m=0}^{\infty} \delta^m_\epsilon u^1(a^1_m,a^2_m)\right] + \epsilon/2 \\
&\leq \mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2}(1-\delta_\epsilon)\left[\sum_{m=0}^{\infty} \delta^m_\epsilon u^1(a^1_m,a^2_m)\right] + \epsilon/2 \\
&\leq \limsup_{N\to\infty} \frac{1}{N}\mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right] + \epsilon/2 + \epsilon' \quad (18) \\
&= \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right] + \epsilon/2 + \epsilon' \quad (19)
\end{aligned}$$

where $\epsilon'$ in (18) is a consequence of the following analysis. Under any stationary optimal strategy $\tilde{\sigma}^1, \tilde{\sigma}^2$ for a discounted problem,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2}(1-\delta_\epsilon)\left[\sum_{m=0}^{\infty} \delta^m_\epsilon u^1(a^1_m,a^2_m)\right] \\
&- \limsup_{N\to\infty} \frac{1}{N}\mathbb{E}_{\tilde{\sigma}^1,\tilde{\sigma}^2}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right] \quad (20)
\end{aligned}$$

is uniformly bounded over all stationary policies under Assumption IV.1. Note finally that since $\tilde{\sigma}^1$ is stationary, limit infimum in (19) and limit supremum in (18) are identical by an application of the dominated convergence theorem (since the actual limit exists as $N \to \infty$). Thus, one can select $\epsilon'$ and then $\epsilon$ arbitrarily small so that the result holds in the following fashion: First pick $\epsilon' > 0$ and find a corresponding $\delta_{\epsilon'}$ with the understanding that for all $\delta_\epsilon \in [\delta_{\epsilon'}, 1)$, (18) holds. Now, select $\delta_\epsilon \geq \delta_{\epsilon'}$ to satisfy the second inequality; such a $\delta_\epsilon$ is guaranteed to exist since there are infinitely many such $\delta$ values up to 1 that satisfies this inequality. Here, the uniformity of the convergence in (20) over all stationary policies is crucial.

In the above analysis, $\tilde{\sigma}^1$ and $\tilde{\sigma}^2$ are stationary, and with this stationary strategy, we have

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu^1,\mu^2}_{\mu_0}\left[\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m)\right] \to \int \nu^*(d\mu,\gamma)G(\mu,\gamma)$$

by the convergence of the expected empirical occupation measures, where $\nu^*$ is some invariant probability measure induced by some optimal stationary strategy. Observe also that such an optimal stationary strategy places a dirac delta measure on the normal type given the stated observability assumptions under its invariant probability measure (which, in turn, is a stagewise commitment policy). This leads to the following result, which says that the supremum over all strategies is equal to the supremum over stationary strategies, which satisfy the structure given in Lemma III.3; let us call such strategies $\Sigma_M$ :

$$\begin{aligned}
&\sup_{\sigma^1} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \\
&= \sup_{\sigma^1\in\Sigma_M} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0=\nu^*}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m). \quad (21)
\end{aligned}$$

Accordingly by Assumption IV.3, the invariant measure on $\mu_t$ will place a full mass on this type, and by (9), we conclude that an optimal strategy exists for Player 1, which will be of commitment type. Finally, we establish that this payoff is attainable for an arbitrary initial prior satisfying the stated assumptions

$$\begin{aligned}
&\sup_{\sigma^1} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \\
&= \sup_{\sigma^1\in\Sigma_M} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m). \quad (22)
\end{aligned}$$

This follows from the fact that:

$$\begin{aligned}
&\sup_{\sigma^1} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \\
&\geq \sup_{\sigma^1\in\Sigma_M} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0=\mu^*}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \quad (23)
\end{aligned}$$

and that by the identifiability condition Assumption IV.3, the same expected payoff (induced by the Stackelberg mimicking commitment strategy) is incurred for every initial prior (satisfying the aforementioned absolute continuity condition, that is, the full-support prior condition)

$$\begin{aligned}
&\inf_{\sigma^1} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \\
&- \inf_{\sigma^1\in\Sigma_M} \liminf_{N\to\infty} \frac{1}{N}\mathbb{E}^{\mu_0=\mu^*}_{\sigma^1,\sigma^2}\sum_{m=0}^{N-1} u^1(a^1_m,a^2_m) \\
&= 0 \quad (24)
\end{aligned}$$

Thus, any optimal strategy will need to be infinite repetition of a stage game Stackelberg action. ■

## APPENDIX F
## PROOF OF LEMMA V.1

From (14), we observe the following. Let $f$ be a continuous function on $\Delta(\Omega)$. Then, $E[f(\mu_{t+1})|\mu_t, \gamma^1_t]$ is continuous in $(\mu_t, \gamma^1_t)$ if

$$\sum_{s^2_t} f(H(\mu_t, s^2_t, \gamma^1_t))P_\sigma(s^2_t|\gamma^1_t)$$

is continuous in $\mu_t, \gamma_t^1$, where $\mu_{t+1} = H(\mu_t, s_t^2, \gamma_t^1)$ defined by (14) with the variables

$$1_{\{\gamma_t^1(\omega, s^2_{[0,t-1]})=a_t^1\}} = P_\sigma(a_t^1|\omega, s^2_{[0,t-1]}), \mu_t(\omega)$$
$$= P_\sigma(\omega|s^2_{[0,t-1]}).$$

Instead of considering continuous functions on $\Delta(\Omega)$, we can also consider continuity of $\mu_{t+1}(\omega)$ for every $\omega$, since pointwise convergence implies convergence in total variation by Scheffé's theorem, which, in turn, implies weak convergence. Now, for every fixed $s_t^2 = s$, $\mu_{t+1}(\omega)$ is continuous in $\mu_t$ for every $\omega$, and hence, $H(\mu_t, s_t^2, \gamma_t^1)$ is continuous in total variation, since pointwise convergence implies convergence in total variation. Furthermore, $P_\sigma(s_t^2|\gamma_t^1, \mu_t)$ is continuous in $\mu_t$ for a given $\gamma_t^1$; thus, weak continuity follows. ■

## Appendix G
## Proof of Lemma VI.1

Suppose that $\max_x u^2(a^1, x) = u^2(a^1, x^*)$. Let $P_\sigma(a^1|s^2_{[0,t]}) \geq 1 - \epsilon$. Let the maximum of

$$P_\sigma(a^1|s^2_{[0,t]})u^2(a^1, x) + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1|s^2_{[0,t]})u^2(\bar{a}_j^1, x)$$

be achieved by $x^*$ so that

$$P_\sigma(a^1|s^2_{[0,t]})u^2(a^1, x') + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1|s^2_{[0,t]})u^2(\bar{a}_j^1, x')$$
$$\leq P_\sigma(a^1|s^2_{[0,t]})u^2(a^1, x^*) + \sum_{\bar{a}_j^1 \neq a^1} P_\sigma(\bar{a}_j^1|s^2_{[0,t]})u^2(\bar{a}_j^1, x^*)$$

for any $x'$. For this to hold, it suffices that

$$P_\sigma(a^1|s^2_{[0,t]})(u^2(a^1, x^*) - u^2(a^1, x')) \geq \max_{s,t} \epsilon u^2(s,t)$$

and since $P_\sigma(a^1|s^2_{[0,t]}) \geq 1 - \epsilon$,

$$(u^2(a^1, x^*) - u^2(a^1, x')) \geq \frac{\max_{s,t} \epsilon u^2(s,t)}{1-\epsilon}.$$

Thus, if $P_\sigma(a^1|s^2_{[0,t]}) > \epsilon$, then the optimal response is to $a^1$. In particular, with $P_\sigma(a^1|s^2_{[0,t]}) \geq 1 - \epsilon$ and for all $\bar{a}_j^1 \neq a^1$, we have $P_\sigma(\bar{a}_j^1|s^2_{[0,t]}) \leq \epsilon/M$; (11) holds. ■

## Appendix H
## Proof of Theorem VI.1

Equation (11) is equivalent to, by Bayes' rule, the following:

$$\frac{P_\sigma(s^2_{[0,t]}|\hat{\omega} = m)}{P_\sigma(s^2_{[0,t]}|\hat{\omega} = k)} \geq \frac{P_\sigma(\hat{\omega} = k)f(M)}{P_\sigma(\hat{\omega} = m)}$$

and

$$\sum_{j=0}^{n} \log \frac{(P_\sigma(s_j^2|\hat{\omega} = m)}{(P_\sigma(s_j^2|\hat{\omega} = k)} \geq \log\left(\frac{P_\sigma(\hat{\omega} = k)f(M)}{P_\sigma(\hat{\omega} = m)}\right).$$

Note now that (11) implies that $t \subset \tau_m$. Thus, we can now apply a measure concentration result through McDiarmid's inequality (see [41]) to deduce that

$$\begin{aligned}
&P_\sigma(t \notin \tau_m) \\
&\leq P\left(\sum_{j=0}^{t} \log\left(\frac{P_\sigma(s_j^2|\hat{\omega} = m)}{P_\sigma(s_j^2|\hat{\omega} = k)}\right)\right. \\
&\qquad \left. \leq \log\left(\frac{P_\sigma(\hat{\omega} = k)f(M)}{P_\sigma(\hat{\omega} = m)}\right)\right) \\
&\leq P\left(\frac{1}{t+1}\sum_{j=0}^{t} \log\left(\frac{P_\sigma(s_j^2|\hat{\omega} = m)}{P_\sigma(s_j^2|\hat{\omega} = k)}\right)\right. \\
&\qquad - \mathbb{E}\left[\frac{\log(P_\sigma(s_j^2|\hat{\omega} = m)}{(P_\sigma(s_j^2|\hat{\omega} = k)}\right] \\
&\qquad \leq \frac{1}{t+1}\log\left(\frac{P_\sigma(\hat{\omega} = k)f(M)}{P_\sigma(\hat{\omega} = m)}\right) \\
&\qquad \left. - \mathbb{E}\left[\frac{\log(P_\sigma(s_j^2|\hat{\omega} = m)}{(P_\sigma(s_j^2|\hat{\omega} = k)}\right]\right) \\
&\leq P\left(\left|\frac{1}{t+1}\sum_{j=0}^{t} \log\left(\frac{P_\sigma(s_j^2|\hat{\omega} = m)}{P_\sigma(s_j^2|\hat{\omega} = k)}\right)\right.\right. \\
&\qquad - \mathbb{E}\left[\frac{\log\left(P_\sigma(s_j^2|\hat{\omega} = m)\right.}{(P_\sigma(s_j^2|\hat{\omega} = k)}\right]\Bigg| \\
&\qquad \geq \left|\mathbb{E}\left[\frac{\log\left(P_\sigma(s_j^2|\hat{\omega} = m)\right.}{(P_\sigma(s_j^2|\hat{\omega} = k)}\right] - \frac{1}{t+1}\log\right. \\
&\qquad \left.\left. \times \left(\frac{P_\sigma(\hat{\omega} = k)f(M)}{P_\sigma(\hat{\omega} = m)}\right)\right|\right) \\
&\leq 2e^{-t\left(\mathbb{E}\left[\log\frac{P_\sigma\left(s_j^2|\hat{\omega}=m\right)}{P_\sigma\left(s_j^2|\hat{\omega}=k\right)}\right] - \frac{1}{t+1}\log\left(\frac{P_\sigma(\hat{\omega}=k)f(M)}{P_\sigma(\hat{\omega}=m)}\right)\right)^2 \Big/ (b-a)}
\end{aligned} \tag{25}$$

where $a \leq \mathbb{S}^j \leq b$ with $\mathbb{S}^j = \frac{P_\sigma(s_j^2|\hat{\omega}=m)}{P_\sigma(s_j^2|\hat{\omega}=k)}$. This implies that the probability of $t \notin \tau_m$ is upper bounded asymptotically by a geometric random variable, that is, there exists $R < \infty$ and $\rho \in (0,1)$ so that for all $t \in \mathbb{N}$, $P_\sigma(t \notin \tau_m) \leq R\rho^t$. ■

## References


[1] D. Abreu, D. Pearce, and E. Stacchetti, "Toward a theory of discounted repeated games with imperfect monitoring," *Econometrica*, vol. 58, no. 5, pp. 1041–1063, 1990.

[2] A. Atakan and M. Ekmekci, "Reputation in long-run relationships," *Rev. Econ. Stud.*, vol. 79, no. 2, pp. 451–480, 2012.

[3] A. Atakan and M. Ekmekci, "A two-sided reputation result with long-run players," *J. Econ. Theory*, vol. 148, no. 1, pp. 376–392, 2013.

[4] A. Atakan and M. Ekmekci, "Reputation in the long-run with imperfect monitoring," *J. Econ. Theory*, vol. 157, pp. 553–605, 2015.

[5] P. Bolton and M. Dewatripont, *Contract Theory*. Cambridge, MA, USA: MIT Press, 2005.

[6] V. S. Borkar, *Probability Theory: An Advanced Course*. Berlin, Germany: Springer, 2012.
[7] V. S. Borkar, "White-noise representations in stochastic realization theory," *SIAM J. Control Optim.*, vol. 31, pp. 1093–1102, 1993.
[8] V. S. Borkar, S. K. Mitter, and S. Tatikonda, "Optimal sequential vector quantization of Markov sources," *SIAM J. Control Optim.*, vol. 40, pp. 135–148, 2001.
[9] V. P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, vol. 50, pp. 1431–1451, 1982.
[10] M. W. Cripps, G. J. Mailath, and L. Samuelson, "Imperfect monitoring and impermanent reputations," *Econometrica*, vol. 72, no. 2, pp. 407–432, 2004.
[11] M. W. Cripps and J. P. Thomas, "Reputation and commitment in two-person repeated games without discounting," *Econometrica*, vol. 63, no. 6), pp. 1401–1419, 1995.
[12] N. A. Dalkiran, "Order of limits in reputations," *Theory Decis.*, vol. 81, no. 3, pp. 393–411, 2016.
[13] L. Epstein, J. Noor, and A. Sandroni, "Non-Bayesian learning," *B. E. J. Theor. Econ.*, vol. 10, no. 1, 2010, Art. no. 3.
[14] M. Ekmekci, "Sustainable reputations with rating systems," *J. Econ. Theory*, vol. 146, no. 2, pp. 479–503, 2011.
[15] M. Ekmekci, O. Gossner, and A. Wilson, "Impermanent types and permanent reputation," *J. Econ. Theory*, vol. 147, no. 1, pp. 162–178, 2012.
[16] E. Faingold, "Reputation and the flow of information in repeated games," *Econometrica*, 2020, to be published.
[17] E. Faingold and Y. Sannikov, "Reputation in continuous-time games," *Econometrica*, vol. 79, no. 3, pp. 773–876, 2011.
[18] D. Fudenberg, D. M. Kreps, and E. Maskin, "Repeated games with long-run and short-run players," *Rev. Econ. Stud.*, vol. 57, pp. 555–573, 1990.
[19] D. Fudenberg and D. K. Levine, "Reputation and equilibrium selection in games with a patient player," *Econometrica*, vol. 57, no. 4, pp. 759–778, 1989.
[20] D. Fudenberg and D. K. Levine, "Maintaining a reputation when strategies are imperfectly observed," *Rev. Econ. Stud.*, vol. 59, no. 3, pp. 561–579, 1992.
[21] D. Fudenberg and D. K. Levine, "Efficiency and observability with long-run and short-run players," *J. Econ. Theory*, vol. 62, no. 1, pp. 103–135, 1994.
[22] D. Fudenberg, D. K. Levine, and E. Maskin, "The folk theorem with imperfect public information," *Econometrica*, vol. 62, no. 5, pp. 997–1039, 1994.
[23] D. Fudenberg and E. Maskin, "The folk theorem in repeated games with discounting or with incomplete information," *Econometrica*, vol. 54, no. 3, pp. 533–554, 1986.
[24] O. Gossner, "Simple bounds on the value of a reputation," *Econometrica*, vol. 79, pp. 1627–1641, 2011.
[25] O. Hernandez-Lerma and J. Lasserre, *Discrete-Time Markov Control Processes*. Berlin, Germany: Springer, 1996.
[26] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*. Berlin, Germany: Springer, 1999.
[27] J. Hörner and S. Lovo, "Belief-free equilibria in games with incomplete information," *Econometrica*, vol. 77, no. 2, pp. 453–487, 2009.
[28] B. Jullien and I.-U. Park, "New, like new, or very good? Reputation and credibility," *Rev. Econ. Stud.*, vol. 81, no. 4, pp. 1543–1574, 2014.
[29] E. Kalai and E. Lehrer, "Rational learning leads to Nash equilibrium," *Econometrica*, vol. 61, no. 5, pp. 1019–1045, 1993.
[30] E. Kalai and E. Lehrer, "Weak and strong merging of opinions," *J. Math. Econ.*, vol. 23, no. 1, pp. 73–86, 1994.
[31] D. M. Kreps, P. Milgrom, D. Roberts, and R. Wilson, "Rational cooperation in the finitely repeated prisoners' dilemma," *J. Econ. Theory*, vol. 27, no. 2, pp. 245–252, 1982.
[32] D. M. Kreps and R. Wilson, "Reputation and imperfect information," *J. Econ. Theory*, vol. 27, no. 2, pp. 253–279, 1982.
[33] T. Linder and S. Yüksel, "On optimal zero-delay quantization of vector Markov sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 2975–5991, Oct. 2014.
[34] Q. Liu, "Information acquisition and reputation dynamics," *Rev. Econ. Stud.*, vol. 78, no. 4, pp. 1400–1425, 2011.
[35] Q. Liu and A. Skrzypacz, "Limited records and reputation bubbles," *J. Econ. Theory*, vol. 151, pp. 2–29, 2014.
[36] A. Mahajan and D. Teneketzis, "On the design of globally optimal communication strategies for real-time noisy communication with noisy feedback," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 580–595, May 2008.
[37] G. J. Mailath and L. Samuelson, *Repeated Games and Reputations*. London, U.K.: Oxford Univ. Press, 2006.
[38] C. McDonald and S. Yüksel, "Stability of non-linear filters and observability of stochastic dynamical systems," *arXiv:1812.01772*.
[39] P. Milgrom and D. Roberts, "Predation, reputation and entry deterrence," *J. Econ. Theory*, vol. 27, no. 2, pp. 280–312, 1982.
[40] A. Özdoğan, "Disappearance of reputations in two-sided incomplete-information games," *Games Econ. Behav.*, vol. 88, pp. 211–220, 2014.
[41] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications and coding," in *Foundations and Trends in Communications and Information Theory*. Delft, The Netherlands: NOW Publishers, 2013.
[42] S. Saritas, S. Yüksel, and S. Gezici, "Quadratic multi-dimensional signaling games and affine equilibria," *IEEE Trans. Autom. Control*, vol. 62, no. 2, pp. 605–619, Feb. 2017.
[43] S. Sorin, "Merging, reputation, and repeated games with incomplete information," *Games Econ. Behav.*, vol. 29, pp. 274–308, 1999.
[44] D. Teneketzis, "On the structure of optimal real-time encoders and decoders in noisy communication," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4017–4035, Sep. 2006.
[45] J. C. Walrand and P. Varaiya, "Optimal causal coding-decoding problems," *IEEE Trans. Inf. Theory*, vol. IT-29, no. 6, pp. 814–820, Nov. 1983.
[46] H. S. Witsenhausen, "On the structure of real-time source coders," *Bell Syst. Tech. J*, vol. 58, pp. 1437–1451, Jul./Aug. 1979.
[47] R. G. Wood, T. Linder, and S. Yüksel, "Optimal zero delay coding of Markov sources: Stationary and finite memory codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5968–5980, Apr. 2017.
[48] S. Yüksel, "On optimal causal coding of partially observed Markov sources in single and multi-terminal settings," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 424–437, Jan. 2013.
[49] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization Under Information Constraints*. Boston, MA, USA: Birkhäuser, 2013.

**Nuh Aygün Dalkıran** received the B.Sc. degree in economics and mathematics from Middle East Technical University, Ankara, Turkey, in 2004, the M.A. degree in economics from Sabancı University, Istanbul, Turkey, in 2006, and the Ph.D. degree in managerial economics and strategy from the Kellogg School of Management, Northwestern University, Evanston, IL, USA, in 2012.

In 2012, he joined the Department of Economics, Bilkent University, Ankara, as an Assistant Professor. His research interests include game theory, repeated games and reputations, economics of information, and decision theory.

**Serdar Yüksel** (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2003 and 2006, respectively.

He was a Postdoctoral Researcher with Yale University, New Haven, CT, USA, before joining the Department of Mathematics and Statistics, Queen's University, Kingston, ON, Canada. His research interests include stochastic control, decentralized control, information theory, and probability.

Dr. Yüksel is an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, *Automatica*, and *Systems and Control Letters*.