Nash Equilibria in Partial-Information Games on Markov Chains[†]

Technical Report

João P. Hespanha[‡] Maria Prandini[*]
hespanha@usc.edu prandini@ing.unibs.it

[‡] *Dept. of Electrical Engineering–Systems, Univ. of Southern California*
3740 McClintock Ave., Room 318, MC 2563, Los Angeles, CA 90089-2563, USA
phone: +1 (213) 740-9137, fax: +1 (213) 821-1109

[*] *Dept. of Electrical Engineering for Automation, University of Brescia*
Via Branze, 38, 25123 Brescia, Italy
phone: +39 (030) 3715-433, fax: +39 (030) 380014

March 6, 2001

Abstract

In this paper we consider two-player partial-information games on Markov chains. These are games in which two players are able to influence the state transitions in a Markov chain by taking appropriate actions. Each player attempts to minimize its own cost that is additive over time with the incremental costs depending on the state of the Markov chain and the actions taken by the players. We consider here finite games played over a finite time horizon and with possibly nonzero-sum costs.

We show that, although the game may not have a Nash equilibrium in deterministic behavioral policies, it has a solution in stochastic behavioral policies. These are policies in which, at each instant of time, both players choose actions randomly according to given probability distributions. These policies are called behavioral because the distributions are a function of the observations and the past actions of the players.

The technique used to prove that a Nash equilibrium exists in stochastic behavioral policies is constructive but has severe limitations because it involves solving an extremely large matrix game. To alleviate this problem we derive a dynamic-programming-like condition that is necessary and sufficient for a pair of stochastic behavioral policies to be a Nash equilibrium. This condition automatically gives Nash equilibria in stochastic behavioral policies when a pair of "cost-to-go" functions can be found that satisfy two given inequalities.

[†]This research was supported by the Office of Naval Research, the Defense Advanced Research Projects Agency, and the Ministero dell'Università e della Ricerca Scientifica e Tecnologica.

[‡]Corresponding author.

1 Introduction

Competitive games are usually classified as either having full or partial-information. In *full-information* games both players know the whole state of the game when they have to make decisions. By state, we mean all information that is needed to completely describe the future evolution of the game, when the decision rules used by both players are known. Examples of full-information games include Chess, Checkers, and Go. *Partial-information* games differ from these in that at least one of the players does not know the whole state of the game. Poker, Bridge, and Hearts are examples of such games. In full-information games, as a player is planning its next move, it only needs to hypothesize over its and the opponent's future moves to predict the possible outcomes of the game. This is key to using dynamic programming [1] to solve full-information games. Partial-information games are especially challenging because this reasoning may fail. In many partial-information games, to predict the possible outcomes of the game, a player must hypothesize not only on the future moves of both players, but also on the past moves of the opponent. This often leads to a tremendous increase in the complexity of the games. In general, partial-information stochastic games are poorly understood and the literature is relatively sparse. Notable exceptions are games with lack of information for one of the players [2, 3], single-stage games [4, 5], or games with particular structures such as the Duel game [6], the Rabbit and Hunter game [7], and the Searchlight game [8, 9]. Linear, quadratic, Gaussian differential games have also been studied but the results available are mostly restricted to cases where the information available to one of the players is a subset of the information available to the other player [10, 11] or when both players share the same observations (although they may not know the controls used by the other player) [12].

This paper addresses fairly general partial-information games on Markov chains. These are games in which two players are able to influence the state transitions in a Markov chain by taking appropriate actions [13]. Each player attempts to minimize its own cost that is additive over time with the incremental costs depending on the state of the Markov chain and the actions taken by the players. We deviate from most of the literature on Markov games in that we do not assume full-information. In fact, each player only has available stochastic measurements that, in general, do not allow it to determine the current state of the Markov chain. We consider here finite games played over a finite time horizon and with possibly nonzero-sum costs.

This paper contains two main contributions. First we show in Section 3 that for the general class of two-player partial-information finite games on Markov chains there always exists a Nash equilibrium in an appropriately defined set of "stochastic behavioral policies". These are policies inspired by the behavioral policies for extensive games (considered, e.g., in [14]) for which, at each instant of time, the players choose actions randomly according to given probability distributions. Here, we extend this class of policies to the setting of partial-information Markov games. Moreover, we enlarge such a class by allowing the distributions over the available actions to also be a function of the actions applied in the past, and not only of the collected observations. With this enlarged type of behavioral policies, we are able to guarantee that a Nash equilibrium always exists for the games considered here. In general, this is not the case for the usual behavioral policies in multi-act extensive games [14, p. 53].

The second contribution of this paper is the derivation in Section 4 of a dynamic programming-like condition that, when satisfied by a pair of "cost-to-go" functions, provides a Nash equilibrium in stochastic behavioral policies. We show that this condition is non-conservative because it is also necessary for the existence of a Nash equilibrium.

Notation: We denote by $(\Omega, \mathcal{F})$ the relevant *measurable space*. Consider a probability measure $\mathrm{P} : \mathcal{F} \to [0,1]$. Given two events $A, B \in \mathcal{F}$ with $\mathrm{P}(B) \neq 0$, we write $\mathrm{P}(A|B)$ for the *conditional probability of A given B*, i.e., $\mathrm{P}(A|B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$. Bold face symbols are used to denote random variables. Following the usual abuse of notation, given a multidimensional random variable $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n)$, where $\boldsymbol{\xi}_i : \Omega \to \mathbb{R}$, $i = 1, 2, \ldots, n$, and some $C = (C_1, C_2, \ldots, C_n)$, where $C_i \subset \mathbb{R}$, $i = 1, 2, \ldots, n$, we write $\mathrm{P}(\boldsymbol{\xi} \in C)$ for $\mathrm{P}(\{\omega \in \Omega : \boldsymbol{\xi}_i(\omega) \in C_i, i = 1, 2, \ldots, n\})$. A similar notation is used for conditional probabilities. Moreover, we write $\mathrm{E}[\boldsymbol{\xi}]$ for the expected value of $\boldsymbol{\xi}$ and $\mathrm{E}[\boldsymbol{\xi}|A]$ for the expected value of $\boldsymbol{\xi}$ conditioned to the event $A \in \mathcal{F}$.

2 Two-player Markov Game

We consider a game between two players: U and D. The game takes place on a controlled Markov chain with state in a finite *state space* $\mathcal{S}$. We denote by $\mathbf{s}_t : \Omega \to \mathcal{S}$ the random variable representing the *state* of the chain at time $t \in \mathcal{T} := \{0, 1, 2, \ldots, T\}$, $T \leq +\infty$. Each player can affect the evolution of the game by applying some control action, and it generally chooses it so as to achieve a certain goal. The control actions are selected based on observations or measurements. The description of the game requires the introduction of the notions of transition and observation probabilities together with those of policies and cost functions.

Transition probability The probability of transition from a given state $s \in \mathcal{S}$ at time t to another state $s' \in \mathcal{S}$ at time $t+1$ is only a function of the *control actions* $\mathbf{u}_t$ and $\mathbf{d}_t$ taken by players U and D, respectively, at time t. By this we mean that the random variable $\mathbf{s}_{t+1}$ is conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{s}_t$, $\mathbf{u}_t$, and $\mathbf{d}_t$. Here, we assume a stationary transition probability, i.e.,

$$\mathrm{P}\left(\mathbf{s}_{t+1} = s' \mid \mathbf{s}_t = s, \mathbf{u}_t = u, \mathbf{d}_t = d\right) = p(s, s', u, d), \qquad s, s' \in \mathcal{S}, u \in \mathcal{U}, d \in \mathcal{D}, t \in \mathcal{T}, \tag{1}$$

where $p : \mathcal{S} \times \mathcal{S} \times \mathcal{U} \times \mathcal{D} \to [0,1]$ is the *transition probability function*. For clarity of notation we shall write $p(s \xrightarrow{ud} s')$ for $p(s, s', u, d)$. We assume here that the control actions $\mathbf{u}_t$ and $\mathbf{d}_t$, $t \in \mathcal{T}$, take values in finite *action spaces* $\mathcal{U}$ and $\mathcal{D}$, respectively. The initial state $\mathbf{s}(0)$ is assumed to be independent of all the other random variables involved in the game at time $t = 0$, and it has probability distribution $\mathrm{P}(\boldsymbol{s}_0 = s)$, $s \in \mathcal{S}$.

Observation probability To choose their actions, *measurements* $\mathbf{y}_t$ and $\mathbf{z}_t$ are available to players U and D, respectively, at time $t \in \mathcal{T}$. The random variable $\mathbf{y}_t$ is assumed to be conditionally independent of all the other random variables at times smaller or equal to t, given $\mathbf{s}_t$. Similarly for $\mathbf{z}_t$. Moreover, the conditional distributions of $\mathbf{y}_t$ and $\mathbf{z}_t$, given the current value of the state $\mathbf{s}_t$, are assumed to be stationary, i.e.,

$$\mathrm{P}\left(\mathbf{y}_t = y \mid \mathbf{s}_t = s\right) = p_Y(y, s), \qquad \mathrm{P}\left(\mathbf{z}_t = z \mid \mathbf{s}_t = s\right) = p_Z(z, s), \qquad s \in \mathcal{S}, y \in \mathcal{Y}, z \in \mathcal{Z}, t \in \mathcal{T},$$

where $p_Y : \mathcal{Y} \times \mathcal{S} \to [0,1]$ and $p_Z : \mathcal{Z} \times \mathcal{S} \to [0,1]$ are the *observation probability functions* for players U and D, respectively. We assume here that the measurements $\mathbf{y}_t$ and $\mathbf{z}_t$, $t \in \mathcal{T}$, take values in finite *measurement spaces* $\mathcal{Y}$ and $\mathcal{Z}$, respectively. To decide which action to take at time $t \in \mathcal{T}$, the information available to player U is then given by the following sequence of measurements and past controls

$$\mathbf{Y}_t := \{\mathbf{y}_0, \mathbf{u}_0, \mathbf{y}_1, \mathbf{u}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{u}_{t-1}, \mathbf{y}_t\},$$

and the information available to player D is the sequence

$$\mathbf{Z}_t := \{\mathbf{z}_0, \mathbf{d}_0, \mathbf{z}_1, \mathbf{d}_1, \ldots, \mathbf{z}_{t-1}, \mathbf{d}_{t-1}, \mathbf{z}_t\}.$$

These sequences are said to be of *length* t. The set of all possible outcomes for $\mathbf{Y}_t$ and $\mathbf{Z}_t$, $t \in \mathcal{T}$, are denoted by $\mathcal{Y}^*$ and $\mathcal{Z}^*$, respectively. In the sequel, we denote the length of the sequences $Y \in \mathcal{Y}^*$ and $Z \in \mathcal{Z}^*$ by $\ell(Y)$ and $\ell(Z)$, respectively. These sequences then have $2\ell(Y)+1$ and $2\ell(Z)+1$ elements, respectively.

When the measurements available to the players are sufficient to let them know at each time instant $t \in \mathcal{T}$ which is the current value realized by the state $\boldsymbol{s}_t$ with probability one, the game is said to be of *full-information*. Formally, this happens when, for every $t \in \mathcal{T}$, the σ-algebras generated by $\mathbf{y}_t$ and $\mathbf{z}_t$ contain the σ-algebra generated by $\mathbf{s}_t$. Games for which this does not happen are said to be of *partial-information*.

Stochastic Policies Informally, a "policy" for one of the players is the rule the player uses to select which actions to take over the time horizon $\mathcal{T}$. We consider here policies that are *stochastic* in that, at every time $t \in \mathcal{T}$, each player selects an action over the action set according to some probability distribution. The policies considered are also *behavioral* in that the specific probability distribution depends on the information collected up to time t. Specifically, a *stochastic behavioral policy* μ of player U is a function $\mu : \mathcal{Y}^* \to [0,1]^{\mathcal{U}}$, where $[0,1]^{\mathcal{U}}$ denotes the set (simplex) of probability distributions over $\mathcal{U}$. We denote by Π_{U} the set of all such policies. For each $Y \in \mathcal{Y}^*$, $\mu(Y)$ is called a *stochastic action* of player U. Similarly, a *stochastic behavioral policy* δ of player D is a function $\delta : \mathcal{Z}^* \to [0,1]^{\mathcal{D}}$, where $[0,1]^{\mathcal{D}}$ denotes the set (simplex) of distributions over $\mathcal{D}$. We denote by Π_{D} the set of all such policies and we call $\delta(Z)$, $Z \in \mathcal{Z}^*$, a *stochastic action* of player D. Stochastic behavioral policies are a generalization to partial-information Markov games of the behavioral policies for multi-act extensive games considered, e.g., in [14]. However, because $\mathbf{Y}_t$, $t \in \mathcal{T}$ contains the past actions of player U, the stochastic actions for this player are now allowed to depend on its past actions. Similarly for player D. The added flexibility provided by these policies is key to the results presented here.

We say that $\mu \in \Pi_{U}$ is a *pure behavioral policy* for player U if, for every $Y \in \mathcal{Y}^*$, the entries of the vector $\mu(Y)$ are in the set $\{0,1\}$. Pure policies are those policies for which the player deterministically selects an action as a function of its past observations. The set of all pure policies for player U is denoted by $\bar{\Pi}_{U}$. A pure behavioral policy for player D is defined similarly, and we denote by $\bar{\Pi}_{D}$ the set of pure policies for player D. Since the domains and images of each pure policy are finite sets, there is only a finite number of pure policies for each player, i.e., the sets $\bar{\Pi}_{U}$ and $\bar{\Pi}_{D}$ are finite.

Opting for specific stochastic policies corresponds to selecting a particular probability measure that is consistent with the distributions chosen for the control actions. In the following we use the subscript $_{\mu\delta}$ in the probability measure P to denote the probability measure associated with $\mu \in \Pi_{U}$ and $\delta \in \Pi_{D}$. When an assertion holds true with respect to $\mathrm{P}_{\mu\delta}$ independently of $\mu \in \Pi_{U}$, or of $\delta \in \Pi_{D}$, or of both $\mu \in \Pi_{U}$ and $\delta \in \Pi_{D}$, we use the notation P_{δ}, P_{μ}, or P, respectively. Similarly for the expected value operator E. According to this notation, the transition and observation probabilities, and the initial state distribution introduced earlier are independent of μ and δ.

When player U selects a stochastic behavioral policy $\mu \in \Pi_{U}$ and player D a policy $\delta \in \Pi_{D}$, the two players are, in fact, jointly selecting a probability measure in the family $\{\mathrm{P}_{\mu\delta} : \mu \in \Pi_{U}, \delta \in \Pi_{D}\}$. This family of probability measures has the property that

$$\mathrm{P}_{\mu}(\mathbf{u}_t = u \mid \mathbf{Y}_t = Y) = \mu_u(Y), \qquad t := \ell(Y), \qquad u \in \mathcal{U},\ Y \in \mathcal{Y}^*, \tag{2}$$

where each $\mu_u(Y)$ denotes the probability in the distribution $\mu(Y)$ over $\mathcal{U}$ that corresponds to the action u. Moreover, $\mathbf{u}_t$ is conditionally independent of all other random variables for times smaller or equal to t, given $\mathbf{Y}_t$. Similarly,

$$\mathrm{P}_{\delta}(\mathbf{d}_t = d \mid \mathbf{Z}_t = Z) = \delta_d(Z), \qquad t := \ell(Z), \qquad d \in \mathcal{D}, Z \in \mathcal{Z}^*, \tag{3}$$

with $\mathbf{d}_t$ conditionally independent of all other random variables for times smaller or equal to t, given $\mathbf{Z}_t$.

Cost Structure In this paper we consider non-cooperative games over a finite horizon $T < \infty$, in which players U and D choose the actions so as to respectively minimize the costs

$$J^U_{\mu\delta} := \mathrm{E}_{\mu\delta}\Big[\sum_{\tau=0}^{T} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau)\Big], \qquad J^D_{\mu\delta} := \mathrm{E}_{\mu\delta}\Big[\sum_{\tau=0}^{T} l_D(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau)\Big]. \tag{4}$$

Typically, $l_U, l_D : \mathcal{S} \times \mathcal{U} \times \mathcal{D} \times \mathcal{T} \to \mathbb{R}$ take the form

$$l_U(s,u,d,\tau) = \begin{cases} c_U(s,u,d), & \tau < T \\ r_U(s), & \tau = T \end{cases}, \qquad l_D(s,u,d,\tau) = \begin{cases} c_D(s,u,d), & \tau < T \\ r_D(s), & \tau = T \end{cases},$$

where the function $r_U : \mathcal{S} \to \mathbb{R}$ assigns to each state s the cost for player U of finishing the game at s, whereas the function $c_U : \mathcal{S} \times \mathcal{U} \times \mathcal{D} \to \mathbb{R}$ assigns to each state s and pair of actions u, d the marginal cost for player U for continuing the game at state s with action u, assuming that player D executes action d. The functions $r_D : \mathcal{S} \to \mathbb{R}$ and $c_D : \mathcal{S} \times \mathcal{U} \times \mathcal{D} \to \mathbb{R}$ play similar roles for player D. We say that the game is zero-sum if $l_D = -l_U$, since under this condition $J^U_{\mu\delta} + J^D_{\mu\delta} = 0$. The coupling between the players' cost functions and the fact that the cost incurred by each player depends on the other player's choices model situations where they are sharing a common environment and competing for the same resources.

We suppose that each player tries to best counteract the other player's action so as to achieve a certain performance level irrespectively of the other player's choice. This is formalized by the concept of *Nash equilibria*. Specifically, a Nash equilibrium occurs when the players select policies for which any unilateral deviation from the equilibrium causes a degradation of performance for the deviating player. In the game of interest, this translates into the players selecting a pair of stochastic behavioral policies $(\mu^*, \delta^*) \in \Pi_U \times \Pi_D$ for which

$$J^U_{\mu^*\delta^*} \le J^U_{\mu\delta^*}, \qquad \forall \mu \in \Pi_U, \tag{5}$$

$$J^D_{\mu^*\delta^*} \le J^D_{\mu^*\delta}, \qquad \forall \delta \in \Pi_D. \tag{6}$$

The policies (μ^*, δ^*) satisfying these conditions are said to constitute a *Nash equilibrium (in stochastic behavioral policies)*.

3 Existence of Nash equilibria

It turns out that, if we restrict our attention to pure policies, a Nash equilibrium in the sense of (5)–(6) may not exist. In fact, we consider stochastic policies precisely because the set of pure policies is too small to always permit Nash equilibria. Before proving that Nash equilibria always exist in stochastic behavioral policies we introduce another type of policies—called mixed policies—for which one can use standard arguments to show that an equilibrium always exists. Mixed policies can be regarded as another method to enlarge the set of pure policies for games that are played repeatedly.

Suppose that both players restrict their attention to pure behavioral policies but independently extract at random which policy to use according to some probability distribution over the sets of pure policies. This extraction is done before the game starts and the resulting game is therefore known as a *prior commitment game*. Denoting by $\rho := \{\rho_\mu : \mu \in \bar{\Pi}_U\}$ and $\sigma := \{\sigma_\delta : \delta \in \bar{\Pi}_D\}$ the distributions used by players U and D, respectively, to choose among their pure policies, the expected costs for players U and D are then equal to

$$\bar{J}^U_{\rho\sigma} := \sum_{\mu\in\bar{\Pi}_U,\ \delta\in\bar{\Pi}_D} \rho_\mu \sigma_\delta J^U_{\mu\delta}, \qquad \bar{J}^D_{\rho\sigma} := \sum_{\mu\in\bar{\Pi}_U, \delta\in\bar{\Pi}_D} \rho_\mu \sigma_\delta J^D_{\mu\delta},$$

respectively. These costs can be thought of as the asymptotic values for the empirical averages of $J^U_{\mu\delta}$ and $J^D_{\mu\delta}$ when the game is played repeatedly, each time extracting at random and independently the policies μ and δ from the sets $\bar{\Pi}_U$ and $\bar{\Pi}_D$, according to the distributions ρ and σ. The extractions are assumed to be independent from game to game, and each game to have the same distribution for the initial state.

The distributions ρ and σ are called *mixed behavioral policies* for players U and D, respectively [15]. The sets of all mixed policies for players U and D (i.e., the simplexes consisting of all distributions over $\bar{\Pi}_U$ and $\bar{\Pi}_D$) are denoted by $[0,1]^{\bar{\Pi}_U}$ and $[0,1]^{\bar{\Pi}_D}$, respectively. The costs $\bar{J}^U_{\rho\sigma}$ and $\bar{J}^D_{\rho\sigma}$ can also be expressed in matrix form as

$$\bar{J}^U_{\rho\sigma} = \rho' A^U \sigma, \qquad\qquad \bar{J}^D_{\rho\sigma} = \rho' A^D \sigma,$$

where A^U and A^D are $|\bar{\Pi}_U| \times |\bar{\Pi}_D|$ matrices defined by

$$[A^U]_{(\mu,\delta)\in\bar{\Pi}_U\times\bar{\Pi}_D} := J^U_{\mu\delta}, \qquad\qquad [A^D]_{(\mu,\delta)\in\bar{\Pi}_U\times\bar{\Pi}_D} := J^D_{\mu\delta},$$

with one row corresponding to each pure policy for player U and one column corresponding to each pure policy for player D, and ρ and μ are interpreted as column vectors. Here, we denoted by $|\bar{\Pi}_U|$ and $|\bar{\Pi}_D|$ the number of elements in the sets $\bar{\Pi}_U$ and $\bar{\Pi}_D$, respectively.

It is well know that at least one Nash equilibrium always exists in mixed policies [15]. In particular, that there always exists a pair of mixed policies $(\rho^*, \sigma^*) \in [0,1]^{\bar{\Pi}_U} \times [0,1]^{\bar{\Pi}_D}$ for which

$$\rho^{*\prime} A^U \sigma^* \le \rho' A^U \sigma^*, \qquad \forall \rho \in [0,1]^{\bar{\Pi}_U}, \tag{7}$$

$$\rho^{*\prime} A^D \sigma^* \le \rho^{*\prime} A^D \sigma, \qquad \forall \sigma \in [0,1]^{\bar{\Pi}_D}. \tag{8}$$

Remark 1. The number of pure policies for each of the players is often extremely large. In the game defined above the number of elements in the domain $\mathcal{Y}^*$ of a pure policy is equal to

$$\sum_{t\in\mathcal{T}} n_y^{t+1} n_u^t = \frac{n_y^{T+2} n_u^{T+1} - n_y}{n_y n_u - 1}$$

where n_y and n_u are the number of elements in $\mathcal{Y}$ and $\mathcal{U}$, respectively. Since the number of possible values for a pure policy is equal to n_u, the total number of pure policies is equal to[1]

$$n_u^{\frac{n_y^{T+2} n_u^{T+1} - n_y}{n_y n_u - 1}}$$

We proceed now to prove the existence of Nash equilibria in stochastic behavioral policies. To achieve this, we need the following result:

[1] Actually, in pure policies the elements u_τ in each Y_t are deterministically determined by the previous y_s, $s \le \tau$, so the number of relevant elements in $\mathcal{Y}^*$ that need to be considered for computing the total number of pure policies is only

$$\sum_{t\in\mathcal{T}} n_y^{t+1} = \frac{n_y^{T+2} - n_y}{n_y - 1},$$

and the corresponding number of pure policies is

$$n_u^{\frac{n_y^{T+2} - n_y}{n_y - 1}}$$

However, this does not alleviate the exponential growth in the number of pure policies.

Lemma 1. *There exist surjective functions $L^U : [0,1]^{\bar{\Pi}_U} \to \Pi_U$ and $L^D : [0,1]^{\bar{\Pi}_D} \to \Pi_D$ such that, for every pair of mixed policies $(\rho, \sigma) \in [0,1]^{\bar{\Pi}_U} \times [0,1]^{\bar{\Pi}_D}$,*

$$\bar{J}^U_{\rho\sigma} = J^U_{\mu\delta}, \qquad \bar{J}^D_{\rho\sigma} = J^D_{\mu\delta}, \tag{9}$$

where $(\mu, \delta) \in \Pi_U \times \Pi_D$ are stochastic policies given by $\mu := L^U(\rho)$, $\delta := L^D(\sigma)$.

To prove Lemma 1 we need the following notation: Given a sequence of measurements $Y \in \mathcal{Y}^*$ and an action $u \in \mathcal{U}$ for player U, we denote by $\bar{\Pi}_{U|\{Y,u\}}$ the set of pure policies that are *compatible* with Y and u, i.e., the set of polices μ for which

$$\mu_u\big(Y\big) = 1, \qquad \mu_{u_\tau}\big(Y_\tau\big) = 1, \quad \forall \tau < \ell(Y),$$

where Y_τ denotes the truncation to length τ of Y and u_τ denotes the action at time τ in the sequence Y. Similarly, given a sequence of measurements $Z \in \mathcal{Z}^*$ and an action $d \in \mathcal{D}$ for player D, we denote by $\bar{\Pi}_{D|\{Z,d\}}$ the set of pure policies that are *compatible* with Z and d. The following Lemma (proved in the Appendix) is also needed to prove Lemma 1.

Lemma 2. *Given any $t \in \mathcal{T}$, $s_0, s_1, \ldots, s_t \in \mathcal{S}$, $u_t \in \mathcal{U}$, $d_t \in \mathcal{D}$, and $Y_t \in \mathcal{Y}^*$, $Z_t \in \mathcal{Z}^*$, with $\ell(Y_t) = \ell(Z_t) = t$, for every pair of stochastic policies $(\mu, \delta) \in \Pi_U \times \Pi_D$,*

$$\begin{aligned} &\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0, \mathbf{s}_1 = s_1, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_t = u_t, \mathbf{d}_t = d_t, \mathbf{Y}_t = Y_t, \mathbf{Z}_t = Z_t) \\ &\qquad = k_t(s_0, s_1, \ldots, s_t, u_t, d_t, Y_t, Z_t) \prod_{\tau=0}^{t} \mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau), \end{aligned} \tag{10}$$

where k_t is a function that does not depends on (μ, δ). When (μ, δ) are pure policies, i.e., $(\mu, \delta) \in \bar{\Pi}_U \times \bar{\Pi}_D$, (10) reduces to

$$\begin{aligned} &\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0, \mathbf{s}_1 = s_1, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_t = u_t, \mathbf{d}_t = d_t, \mathbf{Y}_t = Y_t, \mathbf{Z}_t = Z_t) \\ &\qquad = \begin{cases} k_t(s_0, s_1, \ldots, s_t, u_t, d_t, Y_t, Z_t), & \mu \in \bar{\Pi}_{U|\{Y_t,u_t\}},\ \delta \in \bar{\Pi}_{D|\{Z_t,d_t\}} \\ 0, & \textit{otherwise.} \end{cases} \end{aligned} \tag{11}$$

In the equations above, Y_τ and Z_τ denote the truncations to length τ of Y and Z, respectively, and u_τ and d_τ the actions at time τ in the sequences Y_t and Z_t, respectively.

Proof of Lemma 1. The functions L^U and L^D can be defined as follows: for a given $\rho \in [0,1]^{\bar{\Pi}_U}$, $\sigma \in [0,1]^{\bar{\Pi}_D}$, $L^U(\rho) = \mu$ and $L^D(\sigma) = \delta$, with

$$\mu_u(Y) := \frac{\sum_{\bar{\mu} \in \bar{\Pi}_{U|\{Y,u\}}} \rho_{\bar{\mu}}}{\sum_{\hat{u} \in \mathcal{U}} \sum_{\bar{\mu} \in \bar{\Pi}_{U|\{Y,\hat{u}\}}} \rho_{\bar{\mu}}}, \quad \delta_d(Z) := \frac{\sum_{\bar{\delta} \in \bar{\Pi}_{D|\{Z,d\}}} \sigma_{\bar{\delta}}}{\sum_{\hat{d} \in \mathcal{D}} \sum_{\bar{\delta} \in \bar{\Pi}_{D|\{Z,\hat{d}\}}} \sigma_{\bar{\delta}}}, \quad u \in \mathcal{U}, d \in \mathcal{D}, Y \in \mathcal{Y}^*, Z \in \mathcal{Z}^*. \tag{12}$$

To verify that these functions are surjective it suffices to show that they have right-inverses. We show next that the functions $\bar{L}^U : \Pi_U \to [0,1]^{\bar{\Pi}_U}$ and $\bar{L}^D : \Pi_D \to [0,1]^{\bar{\Pi}_D}$, defined by $\bar{L}^U(\mu) := \rho$ and $\bar{L}^D(\delta) := \sigma$, with

$$\rho_{\bar{\mu}} := \prod_{Y \in \mathcal{Y}^*} \sum_{u \in \mathcal{U}} \bar{\mu}_u(Y)\mu_u(Y), \quad \bar{\mu} \in \bar{\Pi}_U, \qquad \rho_{\bar{\delta}} := \prod_{Z \in \mathcal{Z}^*} \sum_{d \in \mathcal{D}} \bar{\delta}_d(Z)\delta_d(Z), \quad \bar{\delta} \in \bar{\Pi}_D,$$

are right-inverses of L^U and L^D, respectively. To verify that this is true, let $\tilde{\mu} := L^U\big(\bar{L}^U(\mu)\big)$ for some $\mu \in \Pi_U$. Fix an arbitrary $Y \in \mathcal{Y}^*$. From the definitions of L^U and $\bar{L}^U$, we have that

$$\tilde{\mu}_u(Y) = \frac{\sum_{\bar{\mu} \in \bar{\Pi}_{U|\{Y,u\}}} \prod_{\bar{Y}} \sum_{\bar{u}} \bar{\mu}_{\bar{u}}(\bar{Y})\mu_{\bar{u}}(\bar{Y})}{\sum_{\hat{u} \in \mathcal{U}} \sum_{\hat{\mu} \in \bar{\Pi}_{U|\{Y,\hat{u}\}}} \prod_{\bar{Y}} \sum_{\bar{u}} \hat{\mu}_{\bar{u}}(\bar{Y})\mu_{\bar{u}}(\bar{Y})}$$

$$= \frac{\sum_{\bar\mu\in\bar\Pi_{U|\{Y,u\}}} \mu_u(Y)\prod_{\bar Y\neq Y}\sum_{\bar u}\bar\mu_{\bar u}(\bar Y)\mu_{\bar u}(\bar Y)}{\sum_{\hat u\in\mathcal{U}}\sum_{\hat\mu\in\bar\Pi_{U|\{Y,\hat u\}}}\mu_{\hat u\in\mathcal{U}}(Y)\prod_{\bar Y\neq Y}\sum_{\bar u}\hat\mu_{\bar u}(\bar Y)\mu_{\bar u}(\bar Y)} = \frac{\mu_u(Y)}{\sum_{\hat u\in\mathcal{U}}\mu_{\hat u\in\mathcal{U}}(Y)} = \mu_u(Y) \tag{13}$$

Here, we used the fact that

$$\sum_{\bar\mu\in\bar\Pi_{U|\{Y,u\}}}\prod_{\bar Y\neq Y}\sum_{\bar u}\bar\mu_{\bar u}(\bar Y)\mu_{\bar u}(\bar Y) = \sum_{\hat\mu\in\bar\Pi_{U|\{Y,\hat u\}}}\prod_{\bar Y\neq Y}\sum_{\bar u}\hat\mu_{\bar u}(\bar Y)\mu_{\bar u}(\bar Y), \qquad \forall\hat u\in\mathcal{U}. \tag{14}$$

This equality holds because for each $\bar\mu\in\bar\Pi_{U|\{Y,u\}}$ there is exactly one $\hat\mu\in\bar\Pi_{U|\{Y,\hat u\}}$ such that $\hat\mu(\bar Y)=\bar\mu(\bar Y)$, $\bar Y\neq Y$. This means that each term in the summation on the right-hand-side of (14) equals exactly one term in the summation in the left-hand-side of the same equation (and vice-versa). Equation (13) proves that $\bar L^U$ is a right-inverse of L^U. A proof that $\bar L^D$ is a right-inverse of L^D can be constructed in a similar way.

We are now ready to prove that (9) holds. To accomplish this let $\mu := L^U(\rho)$, $\rho\in[0,1]^{\bar\Pi_U}$, and $\delta := L^D(\sigma)$, $\sigma\in[0,1]^{\bar\Pi_D}$. By definition,

$$\begin{aligned} J^U_{\mu\delta} &:= \sum_{\substack{s_0,\ldots,s_T,\\ u_T,d_T,Y_T,Z_T}}\sum_{t=0}^{T} l_U(s_t,u_t,d_t,t)\,\mathrm{P}_{\mu\delta}(\mathbf{s}_0=s_0,\ldots,\mathbf{s}_T=s_T,\mathbf{u}_T=u_T,\mathbf{d}_T=d_T,\mathbf{Y}_T=Y_T,\mathbf{Z}_T=Z_T)\\ &= \sum_{\substack{s_0,s_1,\ldots,s_T,\\ u_T,d_T,Y_T,Z_T}}\sum_{t=0}^{T} l_U(s_t,u_t,d_t,t)\,k_T(s_0,\ldots,s_T,u_T,d_T,Y_T,Z_T)\prod_{\tau=0}^{T}\mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau) \end{aligned} \tag{15}$$

where $Y_T := \{y_0,u_0,y_1,u_1,\ldots,u_{T-1},y_T\}$, $Z_T := \{z_0,d_0,z_1,d_1,\ldots,d_{T-1},z_T\}$, and Y_τ, Z_τ denote the truncations to length τ of Y_T and Z_T, respectively. Here, we used (10) in Lemma 2. We show next that

$$\prod_{\tau=0}^{t}\mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau) = \sum_{\bar\mu\in\bar\Pi_{U|\{Y_t,u_t\}}}\rho_{\bar\mu}\sum_{\bar\delta\in\bar\Pi_{D|\{Z_t,d_t\}}}\sigma_{\bar\delta}, \qquad t\in\mathcal{T}, \tag{16}$$

by induction on t. For $t=0$ this immediately follows from the definition of μ and δ in equation (12). To do the induction step, pick sequences $Y_t\in\mathcal{Y}^*$, $Z_t\in\mathcal{Z}^*$, with $t:=\ell(Y_t)=\ell(Z_t)>0$. Using the induction hypothesis,

$$\begin{aligned} \prod_{\tau=0}^{t}\mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau) &= \mu_{u_t}(Y_t)\delta_{d_t}(Z_t)\prod_{\tau=0}^{t-1}\mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau)\\ &= \mu_{u_t}(Y_t)\delta_{d_t}(Z_t)\sum_{\bar\mu\in\bar\Pi_{U|\{Y_{t-1},u_{t-1}\}}}\rho_{\bar\mu}\sum_{\bar\delta\in\bar\Pi_{D|\{Z_{t-1},d_{t-1}\}}}\sigma_{\bar\delta}. \end{aligned}$$

From this and (12) one concludes that

$$\prod_{\tau=0}^{t}\mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau) = \sum_{\bar\mu\in\bar\Pi_{U|\{Y_t,u_t\}}}\rho_{\bar\mu}\sum_{\bar\delta\in\bar\Pi_{D|\{Z_t,d_t\}}}\sigma_{\bar\delta}.$$

We can now use (16) in (15) to conclude that

$$J^U_{\mu\delta} = \sum_{\substack{s_0,s_1,\ldots,s_T,\\ u_T,d_T,Y_T,Z_T}}\sum_{t=0}^{T} l_U(s_t,u_t,d_t,t)\;\; k_T(s_0,\ldots,s_T,u_T,d_T,Y_T,Z_T)\sum_{\bar\mu\in\bar\Pi_{U|\{Y_T,u_T\}}}\rho_{\bar\mu}\sum_{\bar\delta\in\bar\Pi_{D|\{Z_T,d_T\}}}\sigma_{\bar\delta}. \tag{17}$$

On the other hand,

$$\bar J^U_{\rho\sigma} := \sum_{\bar\mu\in\bar\Pi_U}\sum_{\bar\delta\in\bar\Pi_D} J^U_{\bar\mu\bar\delta}\rho_{\bar\mu}\sigma_{\bar\delta}. \tag{18}$$

Using now (11) in Lemma 2 to specialize (15) for pure policies, we conclude that

$$J^U_{\bar\mu\bar\delta} = \begin{cases} \displaystyle\sum_{\substack{s_0,s_1,\ldots,s_T,\\ u_T,d_T,Y_T,Z_T}} \sum_{t=0}^{T} l_U(s_t,u_t,d_t,t)\;\; k_T(s_0,\ldots,s_T,u_T,d_T,Y_T,Z_T), \quad \bar\mu \in \bar\Pi_{U|\{Y_T,u_T\}},\ \bar\delta \in \bar\Pi_{D|\{Z_T,d_T\}} \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

This, together with (18), show that both $\bar J^U_{\rho\sigma}$ and $J^U_{\mu\delta}$ are equal to the right-hand side of (17). The proof that $\bar J^D_{\rho\sigma} = J^D_{\mu\delta}$ is perfectly analogous. ■

We are now ready to state and prove the main result of this section:

Theorem 1. *Let $(\rho^*,\sigma^*) \in [0,1]^{\bar\Pi_U} \times [0,1]^{\bar\Pi_D}$ be a Nash equilibrium in mixed policies, i.e., a pair of mixed policies for which (7)–(8) hold. Then $(\mu^*,\delta^*) \in \Pi_U \times \Pi_D$, with $\mu^* := L^U(\rho^*)$, $\delta^* := L^D(\sigma^*)$, is a Nash equilibrium in stochastic policies, i.e.,*

$$J^U_{\mu^*\delta^*} \le J^U_{\mu\delta^*}, \quad \forall \mu \in \Pi_U \qquad\qquad J^D_{\mu^*\delta^*} \le J^D_{\mu^*\delta}, \quad \forall \delta \in \Pi_D. \tag{19}$$

Before proving this theorem note that, since there always exists one Nash equilibrium in mixed policies, from Theorem 1 it follows that there always exists at least one Nash equilibrium in stochastic policies. Moreover, Theorem 1 gives a procedure to actually compute the corresponding stochastic policies.

Proof of Theorem 1. By contradiction assume there is a policy $\mu \in \Pi_U$ for which

$$J^U_{\mu^*\delta^*} > J^U_{\mu\delta^*}. \tag{20}$$

Since the map L^U is surjective, there must exist some $\rho \in [0,1]^{\bar\Pi_U}$ such that $\mu = L^U(\rho)$. From (20) and Lemma 1, one then concludes that $\bar J^U_{\rho^*\sigma^*} > \bar J^U_{\rho\sigma^*}$, which violates (7). The second inequality in (19) can be proved similarly. ■

4 Dynamic Programming Approach

In this section we look for a necessary and sufficient condition for a pair of stochastic behavioral policies $(\mu^*,\delta^*) \in \Pi_U \times \Pi_D$ to be Nash equilibrium, based on a dynamic programming approach. With this in mind, we first treat the case where one player computes its optimal policy assuming that its opponent is following a known policy. In particular, in Section 4.1 we consider the following problem: Given $\delta \in \Pi_D$, determine $\mu^* \in \Pi_U$ such that

$$J^U_{\mu^*\delta} = \inf_{\mu\in\Pi_U} J^U_{\mu\delta}. \tag{21}$$

This can be viewed as determining the optimal policy for the follower in a Stackelberg equilibrium (with player D being the leader and player U the follower) [16]. Results analogous to the ones proven in Section 4.1 are valid for the case when the roles are inverted, i.e., player U acts as the leader and player D as the follower.

4.1 Solution to the optimization problem

For given policies $\mu \in \Pi_U, \delta \in \Pi_D$, we define $V^U_{\mu\delta}(Y)$, $Y \in \mathcal{Y}^*$ to be *player U's cost-to-go from Y* associated with the policies μ and δ, after having collected a sequence Y of length $t := \ell(Y) \in \mathcal{T}$ of observations and controls, i.e.,

$$\mathrm{E}_{\mu\delta}\Big[\sum_{\tau=t}^{T} l_U(\mathbf{s}_\tau,\mathbf{u}_\tau,\mathbf{d}_\tau,\tau)\;\Big|\;\mathbf{Y}_t = Y\Big].$$

The expected value above is only well defined when $\mathrm{P}_{\mu\delta}(\mathbf{Y}_t = Y) \neq 0$ but it is actually convenient to define cost-to-go for any $Y \in \mathcal{Y}^*$ such that there is some policy $\hat{\mu}_Y$ for which $\mathrm{P}_{\hat{\mu}_Y\delta}(\mathbf{Y}_t = Y) \neq 0$. To do this we formally define

$$V^U_{\mu\delta}(Y) := \mathrm{E}_{\tilde{\mu}_Y\delta}\Big[\sum_{\tau=t}^{T} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau) \;\Big|\; \mathbf{Y}_t = Y\Big], \tag{22}$$

where the policy $\tilde{\mu}_Y$ is given by

$$\tilde{\mu}_Y(\bar{Y}) := \begin{cases} \mu(\bar{Y}), & \ell(\bar{Y}) \geq \ell(Y) \\ \hat{\mu}_Y(\bar{Y}), & \ell(\bar{Y}) < \ell(Y) \end{cases}. \tag{23}$$

In particular, when $\mathrm{P}_{\mu\delta}(\mathbf{Y}_t = Y) \neq 0$, we can simply choose $\hat{\mu}_Y = \mu$, so that $\tilde{\mu}_Y = \mu$. The cost $V^U_{\mu\delta}$ is always well defined because, for all $Y \in \mathcal{Y}^*$,

1. $\mathrm{P}_{\tilde{\mu}_Y\delta}(\mathbf{Y}_t = Y) = \mathrm{P}_{\hat{\mu}_Y\delta}(\mathbf{Y}_t = Y) \neq 0$ (cf. (v) in Lemma 4 in the Appendix), and
2. the value of $V^U_{\mu\delta}(Y)$ is independent of the policy $\hat{\mu}_Y$ chosen to define $\tilde{\mu}_Y$ (cf. (iii) in Lemma 4).

The intuition behind this is that once $\mathbf{Y}_t = Y$, it does not really matter what was the value of the policy before time t. So we might as well set the value of the cost-to-go from Y, for a policy $\mu \in \Pi_U$ for which the event $\mathbf{Y}_t = Y$ has zero probability, to be identical to that of any policy taking the same stochastic actions as μ from t on.

The cost $J^U_{\mu\delta}$ associated with a pair of policies $\mu \in \Pi_U, \delta \in \Pi_D$ can be easily computed from player U's cost-to-go $V^U_{\mu\delta}$. Indeed, using Lemma 3 in the Appendix and the fact that the probability distribution of $\mathbf{y}_0$ is independent of the policies μ and δ, we conclude that

$$J^U_{\mu\delta} = \mathrm{E}_{\mu\delta}\left[V^U_{\mu\delta}\big(\{\mathbf{y}_0\}\big)\right] = \mathrm{E}\left[V^U_{\mu\delta}\big(\{\mathbf{y}_0\}\big)\right]. \tag{24}$$

We shall see next that it is possible to compute player U's cost-to-go $V^U_{\mu\delta}$ using the operator $T^U_{\mu\delta}$ from the set of functionals $\mathcal{V}^U := \{V : \mathcal{Y}^* \to \mathbb{R}\}$ into itself, defined by

$$T^U_{\mu\delta}V(Y) := \mathrm{E}_{\tilde{\mu}_Y\delta}\left[l_U\big(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t\big) + V(\mathbf{Y}_{t+1}) \;\middle|\; \mathbf{Y}_t = Y\right], \qquad Y \in \mathcal{Y}^*,\ t := \ell(Y), \tag{25}$$

where $V(\mathbf{Y}_{T+1}) := 0$ and $\tilde{\mu}_Y$ is defined in (23). It is straightforward to check that $T^U_{\mu\delta}$ is monotone in the sense that

$$V_1 \leq_{\mathcal{V}} V_2 \qquad \Longrightarrow \qquad T^U_{\mu\delta}V_1 \leq_{\mathcal{V}} T^U_{\mu\delta}V_2, \qquad\qquad V_1, V_2 \in \mathcal{V}^U,$$

where the partial order $\leq_{\mathcal{V}}$ on $\mathcal{V}^U$ is defined so that $V_1 \leq_{\mathcal{V}} V_2$ whenever $V_1(Y) \leq V_2(Y)$, $\forall Y \in \mathcal{Y}^*$. The following Theorem summarizes the relation between $V^U_{\mu\delta}$ and $T^U_{\mu\delta}$. To state this theorem we need to introduce the following notation: Given policies $\mu \in \Pi_U$ and $\delta \in \Pi_D$, we use $\mathcal{Y}^*_{\mu\delta}$ to denote the set of values $Y \in \mathcal{Y}^*$ for which $\mathrm{P}_{\mu\delta}(\mathbf{Y}_\tau = Y) > 0$, $\tau := \ell(Y)$.

Theorem 2. *Given arbitrary policies $\mu \in \Pi_U$ and $\delta \in \Pi_D$,*

(i) For every functional $V \in \mathcal{V}^U$, $V^U_{\mu\delta}(Y) = (T^U_{\mu\delta})^{T-\ell(Y)+1}V(Y)$, $Y \in \mathcal{Y}^*$, where $(T^U_{\mu\delta})^k$ denotes the composition of $T^U_{\mu\delta}$ with itself k times.

(ii) $V^U_{\mu\delta} = T^U_{\mu\delta}V^U_{\mu\delta}$.

(iii) Any function $V \in \mathcal{V}^U$ satisfying $V = T^U_{\mu\delta} V$ on $\mathcal{Y}^*_{\mu\delta}$ is equal to $V^U_{\mu\delta}$ on $\mathcal{Y}^*_{\mu\delta}$.

Proof of Theorem 2. To prove (i), we show by induction on $k \geq 1$ that

$$(T^U_{\mu\delta})^k V(Y) = \mathrm{E}_{\tilde{\mu}_Y \delta}\Big[\sum_{\tau=t}^{t+k-1} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau) + V(\mathbf{Y}_{t+k}) \;\Big|\; \mathbf{Y}_t = Y \Big], \qquad t := \ell(Y), \tag{26}$$

for all $Y \in \mathcal{Y}^*$ such that $\ell(Y) \leq T + 1 - k$, where $V(\mathbf{Y}_{T+1}) := 0$. This proves (i) when we set $k = T + 1 - \ell(Y) = T + 1 - t$. Equation (26) holds true for $k = 1$ because of the definition (25) of $T^U_{\mu\delta}$. Assume now that (26) holds true for $k - 1$. Consider $Y \in \mathcal{Y}^*$ with $\ell(Y) \leq T + 1 - k$. Then,

$$\begin{aligned}
(T^U_{\mu\delta})^k V(Y) &= T^U_{\mu\delta}(T^U_{\mu\delta})^{k-1} V(Y) \\
&= \mathrm{E}_{\tilde{\mu}_Y \delta}\big[l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t) + (T^U_{\mu\delta})^{k-1} V(\mathbf{Y}_{t+1}) \;\big|\; \mathbf{Y}_t = Y \big] \\
&= \mathrm{E}_{\tilde{\mu}_Y \delta}\big[l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t) + V^{U,k}_{\mu\delta}(\mathbf{Y}_{t+1}) \;\big|\; \mathbf{Y}_t = Y \big],
\end{aligned} \tag{27}$$

where, by the induction hypothesis $V^{U,k}_{\mu\delta}(\bar{Y}) := \mathrm{E}_{\tilde{\mu}_{\bar{Y}} \delta}\Big[\sum_{\tau=t+1}^{t+k-1} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau) + V(\mathbf{Y}_{t+k}) \;\Big|\; \mathbf{Y}_{t+1} = \bar{Y} \Big]$, since $\ell(\bar{Y}) = t + 1 \leq T + 1 - (k-1)$. Each $\bar{Y}$ such that $\mathrm{P}_{\tilde{\mu}_Y \delta}(\mathbf{Y}_{t+1} = \bar{Y} \;|\; \mathbf{Y}_t = Y) > 0$ can be expanded as $\{Y, u, y\}$, $u \in \mathcal{U}$, $y \in \mathcal{Y}$ [2]. Moreover, for such $\bar{Y}$ we can choose $\tilde{\mu}_{\bar{Y}} = \tilde{\mu}_Y$, thus getting

$$\begin{aligned}
V^{U,k}_{\mu\delta}\big(\{Y, u, y\}\big) &= \mathrm{E}_{\tilde{\mu}_{\bar{Y}} \delta}\Big[\sum_{\tau=t+1}^{t+k-1} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau) + V(\mathbf{Y}_{t+k}) \;\Big|\; \mathbf{Y}_{t+1} = \{Y, u, y\} \Big] \\
&= \mathrm{E}_{\tilde{\mu}_Y \delta}\Big[\sum_{\tau=t+1}^{t+k-1} l_U(\mathbf{s}_\tau, \mathbf{u}_\tau, \mathbf{d}_\tau, \tau) + V(\mathbf{Y}_{t+k}) \;\Big|\; \mathbf{y}_{t+1} = y, \mathbf{u}_t = u, \mathbf{Y}_t = Y \Big].
\end{aligned}$$

Then, using this expression for $V^{U,k}_{\mu\delta}\big(\bar{Y}\big)$ in equation (27) and Lemma 3 in the Appendix, one finally concludes that (26) is satisfied for every $k \geq 1$.

To prove (ii), we start by noticing that

$$(T^U_{\mu\delta})^{T-\ell(Y)+1}\Big(T^U_{\mu\delta} V^U_{\mu\delta}(Y)\Big) = T^U_{\mu\delta}\Big((T^U_{\mu\delta})^{T-\ell(Y)+1} V^U_{\mu\delta}(Y)\Big), \qquad Y \in \mathcal{Y}^*. \tag{28}$$

Using (i) on left and right-hand-sides of (28) with $V := T^U_{\mu\delta} V^U_{\mu\delta}$ and $V := V^U_{\mu\delta}$, respectively, we conclude that indeed $V^U_{\mu\delta} = T^U_{\mu\delta} V^U_{\mu\delta}$.

To prove (iii), consider some $V \in \mathcal{V}^U$ satisfying $V = T^U_{\mu\delta} V$ on $\mathcal{Y}^*_{\mu\delta}$. We next show by induction on k that, for $k \geq 1$,

$$V(Y) = (T^U_{\mu\delta})^k V(Y), \qquad Y \in \mathcal{Y}^*_{\mu\delta}. \tag{29}$$

Then, (iii) follows from (i) and (29) with $k = T - \ell(Y) + 1$. For $k = 1$, (29) is trivially satisfied. Assume now that (29) holds true for $k - 1$. Pick $Y \in \mathcal{Y}^*_{\mu\delta}$. Then,

$$(T^U_{\mu\delta})^k V(Y) = T^U_{\mu\delta}(T^U_{\mu\delta})^{k-1} V(Y) = \mathrm{E}_{\mu\delta}\big[l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t) + (T^U_{\mu\delta})^{k-1} V(\mathbf{Y}_{t+1}) \;\big|\; \mathbf{Y}_t = Y \big],$$

where $t := \ell(Y)$. Since, in the expected value above, we only have to consider the values of $\mathbf{Y}_{t+1}$ with nonzero probability (i.e., those in $\mathcal{Y}^*_{\mu\delta}$), we can use the induction hypothesis to further conclude that

$$(T^U_{\mu\delta})^k V(Y) = \mathrm{E}_{\mu\delta}\big[l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t) + V(\mathbf{Y}_{t+1}) \;\big|\; \mathbf{Y}_t = Y \big] = T^U_{\mu\delta} V(Y) = V(Y),$$

thus concluding the proof of (29). ∎

[2] Given $Y \in \mathcal{Y}^*$ and $u \in \mathcal{U}$, $y \in \mathcal{Y}$, $\{Y, u, y\}$ denotes the sequence obtained by appending to Y the ordered set $\{u, y\}$.

We proceed by showing how to actually compute the function in $\mathcal{V}^U$ that results from applying $T^U_{\mu\delta}$ to some function $V \in \mathcal{V}^U$. To this effect let $\mathcal{P} := [0,1]^{\mathcal{U}}$ denote the set of probability distributions over $\mathcal{U}$. For each $p := \{p_u : u \in \mathcal{U}\} \in \mathcal{P}$, and $\delta \in \Pi_D$ we can define an operator $H^U_{p\delta}$ from $\mathcal{V}^U$ into itself by setting for each $Y \in \mathcal{Y}^*$:

$$H^U_{p\delta}V(Y) := \sum_u p_u \Big(\sum_{\substack{s,s',y \\ d,Z}} \Big(l_U(s,u,d,\tau) + V\big(\{Y,u,y\}\big)\Big) p_Y(y,s') p(s \xrightarrow{ud} s') \delta_d(Z) I_\delta(s,Z,Y) \Big),$$

where $\tau := \ell(Y)$ and $V\big(\{Y,u,y\}\big) = 0$ for $\tau = T$. The function $I_\delta : \mathcal{S} \times \mathcal{Z}^* \times \mathcal{Y}^* \to \mathbb{R}$ is defined as follows: For given $Y \in \mathcal{Y}^*$ and $Z \in \mathcal{Z}^*$, $I_\delta(s,Z,Y) = 0$, $s \in \mathcal{S}$ when $\ell(Y) \neq \ell(Z)$. Otherwise, I_δ is defined recursively by

$$I_\delta(s',\{z\},\{y\}) = \frac{p_Y(y,s') p_Z(z,s') \,\mathrm{P}(\mathbf{s}_0 = s')}{\sum_{\bar{s}} p_Y(y,\bar{s}) \,\mathrm{P}(\mathbf{s}_0 = \bar{s})}, \tag{30}$$

and

$$I_\delta(s',\{Z,d,z\},\{Y,u,y\}) = \frac{\sum_s p_Y(y,s') p_Z(z,s') p(s \xrightarrow{ud} s') \delta_d(Z) I_\delta(s,Z,Y)}{\sum_{\substack{\bar{s},\bar{z},\bar{d} \\ \bar{s}',\bar{Z}}} p_Y(y,\bar{s}') p_Z(\bar{z},\bar{s}') p(\bar{s} \xrightarrow{u\bar{d}} \bar{s}') \delta_{\bar{d}}(\bar{Z}) I_\delta(\bar{s},\bar{Z},Y)}, \tag{31}$$

$s' \in \mathcal{S}, u \in \mathcal{U}, d \in \mathcal{D}, y \in \mathcal{Y}, z \in \mathcal{Z}$. From the definition of $T^U_{\mu\delta}$, for each $Y \in \mathcal{Y}^*$,

$$T^U_{\mu\delta}V(Y) = \sum_{\substack{s,s',y,u \\ d,Z}} \Big(l_U(s,u,d,\tau) + V\big(\{Y,u,y\}\big)\Big) \\ \mathrm{P}_{\tilde{\mu}_Y\delta}\big(\mathbf{y}_{\tau+1} = y, \mathbf{s}_{\tau+1} = s', \mathbf{s}_\tau = s, \mathbf{u}_\tau = u, \mathbf{d}_\tau = d, \mathbf{Z}_\tau = Z \;\big|\; \mathbf{Y}_\tau = Y\big),$$

where $\tau := \ell(Y)$ and $V\big(\{Y,u,y\}\big) = 0$ for $\tau = T$. Then, by (ii) in Lemma 4 in the Appendix and the definition of $H^U_{p\delta}$, one concludes that

$$T^U_{\mu\delta}V(Y) = H^U_{\tilde{\mu}_Y(Y)\delta}V(Y), \qquad Y \in \mathcal{Y}^*. \tag{32}$$

This is actually the key equation for showing that the multi-step optimization problem (21) can be reduced to multiple single-step optimization problems. This is the subject of the developments that follow.

Optimal Cost-to-Go For a given policy $\delta \in \Pi_D$ we define *player U's optimal cost-to-go* function V^{U*}_δ associated with the policy δ as

$$V^{U*}_\delta(Y) := \inf_{\mu \in \Pi_U} V^U_{\mu\delta}(Y), \qquad Y \in \mathcal{Y}^*.$$

The optimal cost

$$J^{U*}_\delta := \inf_{\mu \in \Pi_U} J^U_{\mu\delta^*}$$

can be easily computed from player U's optimal cost-to-go V^{U*}_δ. Indeed, it is straightforward to show (cf. [18]) that

$$J^{U*}_\delta = \mathrm{E}\left[V^{U*}_\delta\big(\{\mathbf{y}_0\}\big)\right]. \tag{33}$$

It turns out that one can compute player U's optimal cost-to-go using the operator $T^U_\delta : \mathcal{V}^U \to \mathcal{V}^U$ defined by

$$T^U_\delta V(Y) = \inf_{p \in \mathcal{P}} H^U_{p\delta}V(Y), \qquad Y \in \mathcal{Y}^*. \tag{34}$$

Due to the linearity of the map $p \longmapsto H^U_{p\delta}V(Y)$ and the particular structure of $\mathcal{P}$, the infimum is actually a minimum and can be achieved at some vector in $\mathcal{P}$ with all entries in the set $\{0,1\}$. The following Theorem summarizes the relation between $V_\delta^{U^*}$ and T^U_δ:

Theorem 3. *Given an arbitrary policy $\delta \in \Pi_{_D}$, the following statements hold true:*

(i) $V_\delta^{U^*} = T^U_\delta V_\delta^{U^*}$.

(ii) For any policy $\mu \in \Pi_{_U}$ satisfying $V^U_{\mu\delta} = T^U_\delta V^U_{\mu\delta}$ on $\mathcal{Y}^*_{\mu\delta}$, we have that $V^U_{\mu\delta} = V_\delta^{U^*}$ on $\mathcal{Y}^*_{\mu\delta}$.

(iii) If $\mu^* \in \Pi_{_U}$ is an optimal policy, then $T^U_{\mu^*\delta}V_\delta^{U^*} = T^U_\delta V_\delta^{U^*}$ on $\mathcal{Y}^*_{\mu^*\delta}$.

Proof of Theorem 3. To prove (i) we start by picking an arbitrary policy $\mu \in \Pi_{_U}$. Because of (ii) in Theorem 2, the definition of $V_\delta^{U^*}$, the monotonicity of $T^U_{\mu\delta}$, and equation (32),

$$V^U_{\mu\delta}(Y) = T^U_{\mu\delta}V^U_{\mu\delta}(Y) \geq T^U_{\mu\delta}V_\delta^{U^*}(Y) = H^U_{\tilde{\mu}_Y(Y)\delta}V_\delta^{U^*}(Y) \geq \inf_{p\in\mathcal{P}} H^U_{p\delta}V_\delta^{U^*}(Y), \qquad Y \in \mathcal{Y}^*.$$

Taking the infimum over all policies $\mu \in \Pi_{_U}$, one concludes that

$$V_\delta^{U^*}(Y) = \inf_{\mu\in\Pi_U} V^U_{\mu\delta}(Y) \geq \inf_{p\in\mathcal{P}} H^U_{p\delta}V_\delta^{U^*}(Y), \qquad Y \in \mathcal{Y}^*.$$

We have thus proved that

$$T^U_\delta V_\delta^{U^*} \leq_{\mathcal{V}} V_\delta^{U^*}. \tag{35}$$

We show next that actually $V_\delta^{U^*} = T^U_\delta V_\delta^{U^*}$. As argued before, due to the linearity of the map $p \longmapsto H^U_{p\delta}V_\delta^{U^*}(Y)$ and the particular structure of $\mathcal{P}$, the infimum in $\inf_{p\in\mathcal{P}} H^U_{p\delta}V_\delta^{U^*}(Y)$ is actually a minimum and can be achieved at some vector in $\mathcal{P}$ with all entries in the set $\{0,1\}$. Let then $\bar{\mu} \in \Pi_{_U}$ be a policy for which $\bar{\mu}(Y)$, $Y \in \mathcal{Y}^*$, is a vector in $\mathcal{P}$ that minimizes $H^U_{p\delta}V_\delta^{U^*}(Y)$, with all entries in the set $\{0,1\}$. By construction such policy is deterministic and

$$T^U_\delta V_\delta^{U^*} = T^U_{\bar{\mu}\delta}V_\delta^{U^*}. \tag{36}$$

From (36) and (35) we conclude that

$$T^U_{\bar{\mu}\delta}V_\delta^{U^*} \leq_{\mathcal{V}} V_\delta^{U^*},$$

and therefore, using the monotonicity of $T^U_{\bar{\mu}\delta}$ and (36), we obtain

$$(T^U_{\bar{\mu}\delta})^k V_\delta^{U^*}(Y) \leq T^U_{\bar{\mu}\delta}V_\delta^{U^*}(Y) = T^U_\delta V_\delta^{U^*}(Y), \qquad Y \in \mathcal{Y}^*,\ k \geq 1. \tag{37}$$

For each $Y \in \mathcal{Y}^*$, we can then construct the following chain of inequalities:

$$V_\delta^{U^*}(Y) \leq V^U_{\bar{\mu}\delta}(Y) = (T^U_{\bar{\mu}\delta})^{T-\ell(Y)+1}V_\delta^{U^*}(Y) \leq T^U_\delta V_\delta^{U^*}(Y) \leq V_\delta^{U^*}(Y), \tag{38}$$

where the first inequality follows from the definition of $V_\delta^{U^*}$, the next equality follows from (i) in Theorem 2, the next inequality follows from (37), and the last inequality follows from (35). From (38) we conclude that

$$V_\delta^{U^*}(Y) = V^U_{\bar{\mu}\delta}(Y) = T^U_\delta V_\delta^{U^*}(Y), \qquad Y \in \mathcal{Y}^*, \tag{39}$$

which proves that $V_\delta^{U^*}$ is a fixed point of the operator T^U_δ.

To prove (ii) consider a policy $\mu \in \Pi_U$ satisfying $V^U_{\mu\delta} = T^U_\delta V^U_{\mu\delta}$ on $\mathcal{Y}^*_{\mu\delta}$. From this and (ii) in Theorem 2, we conclude that

$$V^U_{\mu\delta}(Y) = T^U_\delta V^U_{\mu\delta}(Y) = T^U_{\mu\delta} V^U_{\mu\delta}(Y), \qquad Y \in \mathcal{Y}^*_{\mu\delta}. \tag{40}$$

Based on this, we next prove that

$$V^U_{\mu\delta}(Y) = {V^U_\delta}^*(Y), \qquad Y \in \mathcal{Y}^*_{\mu\delta}, \tag{41}$$

by induction on the length of Y. We start by considering an arbitrary sequence $Y_T \in \mathcal{Y}^*_{\mu\delta}$ of length T. Then

$$V^U_{\mu\delta}(Y_T) = \inf_{p\in\mathcal{P}} H^U_{p\delta} V^U_{\mu\delta}(Y_T) = \inf_{p\in\mathcal{P}} H^U_{p\delta} {V^U_\delta}^*(Y_T) = T^{U^*}_\delta {V^U_\delta}^*(Y_T) = {V^U_\delta}^*(Y_T).$$

The first equality follows from (40), the second from the fact that the value of $H^U_{p\delta}V$ at sequences of length T actually does not depend on V, the third from the definition of $T^{U^*}_\delta$, and the last equality follows from (i) proved above. For the induction step we assume that (41) holds for all sequences $Y \in \mathcal{Y}^*_{\mu\delta}$ of length $t+1 \le T$, and pick one sequence $Y \in \mathcal{Y}^*_{\mu\delta}$ of length t. Because of (40),

$$\begin{aligned} V^U_{\mu\delta}(Y) &= \inf_{p\in\mathcal{P}} H^U_{p\delta} V^U_{\mu\delta}(Y) \\ &= \sum_u \mu_u(Y) \Bigg(\sum_{\substack{s,s',y \\ d,Z}} \Big(l_U(s,u,d,t) + V^U_{\mu\delta}\big(\{Y,u,y\}\big)\Big) p_Y(y,s') p(s \xrightarrow{ud} s') \delta_d(Z) I_\delta(s,Z,Y) \Bigg). \end{aligned}$$

Since only the $\{Y,u,y\}$ with nonzero probability (i.e., those in $\mathcal{Y}^*_{\mu\delta}$) need to be considered and $\{Y,u,y\}$ has length $t+1$, by the induction hypothesis we conclude that $V^U_{\mu\delta}(\{Y,u,y\}) = {V^U_\delta}^*(\{Y,u,y\})$ in the expression above, and therefore

$$V^U_{\mu\delta}(Y) = \inf_{p\in\mathcal{P}} H^U_{p\delta} {V^U_\delta}^*(Y) = {V^U_\delta}^*(Y),$$

where the last equality follows from (i).

To prove (iii) note that, because of the definition of T^U_δ, we always have that $T^U_\delta {V^U_\delta}^* \le_{\mathcal{V}} T^U_{\mu^*\delta} {V^U_\delta}^*$. Next, we prove by contradiction that the previous inequality cannot be strict. To this effect, suppose that there exists $\bar{Y} \in \mathcal{Y}^*$ such that

$$T^U_\delta {V^U_\delta}^*(\bar{Y}) = \inf_{p\in\mathcal{P}} H^U_{p\delta} {V^U_\delta}^*(\bar{Y}) < T^U_{\mu^*\delta} {V^U_\delta}^*(\bar{Y}), \qquad \bar{Y} \in \mathcal{Y}^*_{\mu^*\delta}.$$

Construct now a policy $\bar{\mu} \in \Pi_U$ satisfying the following properties:

1. $\bar{\mu}(Y) = \mu^*(Y)$, for all $Y \in \mathcal{Y}^*$ for which $\ell(Y) < \ell(\bar{Y})$,
2. $T^U_{\bar{\mu}\delta} {V^U_\delta}^*(Y) = T^U_\delta {V^U_\delta}^*(Y) = \inf_{p\in\mathcal{P}} H^U_{p\delta} {V^U_\delta}^*(Y)$, for all $Y \in \mathcal{Y}^*$ for which $\ell(Y) \ge \ell(\bar{Y})$.

Then, by the same procedure that lead to equation (39), one can prove that for every $Y \in \mathcal{Y}^*$, with $\ell(Y) \ge \ell(\bar{Y})$

$$V^U_{\bar{\mu}\delta}(Y) = {V^U_\delta}^*(Y) \le V^U_{\mu^*\delta}(Y). \tag{42}$$

As for $Y = \bar{Y}$, the inequality is strict, i.e.,

$$V^U_{\bar{\mu}\delta}(\bar{Y}) = {V^U_\delta}^*(\bar{Y}) < V^U_{\mu^*\delta}(\bar{Y}), \tag{43}$$

since

$$V^U_{\bar{\mu}\delta}(\bar{Y}) = V^{U*}_{\delta}(\bar{Y}) = T^U_{\delta} V^{U*}_{\delta}(\bar{Y}) < T^U_{\mu^*\delta} V^{U*}_{\delta}(\bar{Y}) \leq T^U_{\mu^*\delta} V^U_{\mu^*\delta}(\bar{Y}) = V^U_{\mu^*\delta}(\bar{Y}).$$

From the definition (22) of the cost-to-go and (42), we conclude that

$$\begin{aligned} J^U_{\bar{\mu}\delta} &= \mathrm{E}_{\bar{\mu}\delta}\Big[\sum_{t=0}^{T} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] = \mathrm{E}_{\bar{\mu}\delta}\left[V^U_{\bar{\mu}\delta}(\mathbf{Y}_\tau)\right] + \mathrm{E}_{\bar{\mu}\delta}\Big[\sum_{t=0}^{\tau-1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] \\ &= \mathrm{E}_{\bar{\mu}\delta}\left[V^{U*}_{\delta}(\mathbf{Y}_\tau)\right] + \mathrm{E}_{\bar{\mu}\delta}\Big[\sum_{t=0}^{\tau-1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big]. \end{aligned} \tag{44}$$

From (v) in Lemma 4, it follows that

$$\mathrm{E}_{\bar{\mu}\delta}\left[V^{U*}_{\delta}(\mathbf{Y}_\tau)\right] = \mathrm{E}_{\mu^*\delta}\left[V^{U*}_{\delta}(\mathbf{Y}_\tau)\right]. \tag{45}$$

As for the second term in (44) (which only exists if $\tau \geq 1$), we show by induction that

$$\mathrm{E}_{\bar{\mu}\delta}\Big[\sum_{t=0}^{k} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] = \mathrm{E}_{\mu^*\delta}\Big[\sum_{t=0}^{k} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big], \tag{46}$$

for any $k < \tau$. For $k = 0$, we have

$$\begin{aligned} \mathrm{E}_{\bar{\mu}\delta}\left[l_U(\mathbf{s}_0, \mathbf{u}_0, \mathbf{d}_0, 0)\right] &= \sum_{\substack{s,u,d \\ y,z}} l_U(\mathbf{s}_0 = s, \mathbf{u}_0 = u, \mathbf{d}_0 = d, 0)\bar{\mu}_u(y)\delta_d(z)p_Y(y,s)p_Z(z,s)\,\mathrm{P}(\mathbf{s}_0 = s) \\ &= \mathrm{E}_{\mu^*\delta}\left[l_U(\mathbf{s}_0, \mathbf{u}_0, \mathbf{d}_0, 0)\right], \end{aligned}$$

since $\bar{\mu}$ and μ^* are equal for sequences of zero length (because $\tau \geq 1$). Assume now that (46) holds true for $k \geq 0$ and therefore

$$\mathrm{E}_{\bar{\mu}\delta}\Big[\sum_{t=0}^{k+1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] = \mathrm{E}_{\mu^*\delta}\Big[\sum_{t=0}^{k} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] + \mathrm{E}_{\bar{\mu}\delta}\Big[l_U(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{d}_{k+1}, k+1)\Big]. \tag{47}$$

Because of (ii) and (v) in Lemma 4, we have that

$$\begin{aligned} \mathrm{E}_{\bar{\mu}\delta}\left[l_U(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{d}_{k+1}, k+1)\right] = \sum_{\substack{y,s,s', \\ d,u,Z,Y}} & l_U(\mathbf{s}_{k+1} = s, \mathbf{u}_{k+1} = u, \mathbf{d}_{k+1} = d, k+1)p_Y(y,s')p(s \xrightarrow{ud} s') \\ & \bar{\mu}_u(Y)\delta_d(Z)I_\delta(s,Z,Y)\,\mathrm{P}_{\mu^*\delta}\left(\mathbf{Y}_{k+1} = Y\right) \\ = \mathrm{E}_{\mu^*\delta}&\left[l_U(\mathbf{s}_{k+1}, \mathbf{u}_{k+1}, \mathbf{d}_{k+1}, k+1)\right], \end{aligned}$$

provided that $k+1 < \tau$ and therefore that $\bar{\mu}$ and μ^* match for all the $Y \in \mathcal{Y}^*$ with length $k+1$. Combining this with (47) concludes the proof by induction of (46).

Using (45) and (46), we conclude that (44) can be rewritten as

$$\begin{aligned} J^U_{\bar{\mu}\delta} &= \mathrm{E}_{\mu^*\delta}\left[V^{U*}_{\delta}(\mathbf{Y}_\tau)\right] + \mathrm{E}_{\mu^*\delta}\Big[\sum_{t=0}^{\tau-1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] \\ &= \sum_{Y} V^{U*}_{\delta}(Y)\,\mathrm{P}_{\mu^*\delta}(\mathbf{Y}_\tau = Y) + \mathrm{E}_{\mu^*\delta}\Big[\sum_{t=0}^{\tau-1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big]. \end{aligned}$$

Since $\mathrm{P}_{\mu^*\delta}(\mathbf{Y}_\tau = \bar{Y}) > 0$, from (42) and (43), we then obtain

$$J^U_{\bar{\mu}\delta} < \mathrm{E}_{\mu^*\delta}\left[V^U_{\mu^*\delta}(\mathbf{Y}_\tau)\right] + \mathrm{E}_{\mu^*\delta}\Big[\sum_{t=0}^{\tau-1} l_U(\mathbf{s}_t, \mathbf{u}_t, \mathbf{d}_t, t)\Big] = J^U_{\mu^*\delta}$$

which contradicts the fact that μ^* is optimal. ■

4.2 Characterization of Nash equilibria

Using the results derived in the previous section, we finally derive a necessary and sufficient condition for $(\mu^*, \delta^*) \in \Pi_U \times \Pi_D$ to be a Nash equilibrium. To accomplish this, we need to extend to player D the notation that was introduced above for player U. In the sequel, we shall use $\mathcal{V}^D$, $V^D_{\mu\delta}$, $H^D_{\mu q}$, V^{D*}_{μ}, T^D_{μ}, and $\mathcal{Q}$ to denote the duals of $\mathcal{V}^U$, $V^U_{\mu\delta}$, $H^U_{p\delta}$, V^{U*}_{δ}, T^U_{δ}, and $\mathcal{P}$, respectively.

Theorem 4. *$(\mu^*, \delta^*) \in \Pi_U \times \Pi_D$ constitute a Nash equilibrium if and only if there exists two functionals $V^U \in \mathcal{V}^U$ and $V^D \in \mathcal{V}^D$ satisfying the following conditions:*

$$H^U_{p\delta^*} V^U(Y) \geq H^U_{\mu^*(Y)\delta^*} V^U(Y) = V^U(Y), \qquad \forall p \in \mathcal{P},\ Y \in \mathcal{Y}^*_{\mu^*\delta^*} \tag{48}$$

$$H^D_{\mu^* q} V^D(Z) \geq H^D_{\mu^*\delta^*(Z)} V^D(Z) = V^D(Z), \qquad \forall q \in \mathcal{Q},\ Z \in \mathcal{Z}^*_{\mu^*\delta^*} \tag{49}$$

Before proving Theorem 4, note that we actually know from Theorem 1 that there always exists at least one Nash equilibrium in stochastic policies. This mean that there must always exist functionals $V^U \in \mathcal{V}^U$ and $V^D \in \mathcal{V}^D$ satisfying (48)–(49).

Proof of Theorem 4. Suppose that there exist functionals $V^U \in \mathcal{V}^U$, $V^D \in \mathcal{V}^D$ satisfying (48)–(49). We prove next that (μ^*, δ^*) is a Nash equilibrium because these policies satisfy equations (5)–(6).

Because of (32) and the definition of T^U_{δ}, (48) can be rewritten as

$$V^U(Y) = T^U_{\delta^*} V^U(Y) = T^U_{\mu^*\delta^*} V^U(Y), \qquad Y \in \mathcal{Y}^*_{\mu^*\delta^*}. \tag{50}$$

Then, by (iii) in Theorem 2, $V^U(Y) = V^U_{\mu^*\delta^*}(Y)$ on $\mathcal{Y}^*_{\mu^*\delta^*}$. Based on this, (50) becomes

$$V^U_{\mu^*\delta^*}(Y) = T^U_{\delta^*} V^U_{\mu^*\delta^*}(Y) = T^U_{\mu^*\delta^*} V^U_{\mu^*\delta^*}(Y), \qquad Y \in \mathcal{Y}^*_{\mu^*\delta^*},$$

which, by (ii) in Theorem 3, leads to

$$V^U(Y) = V^U_{\mu^*\delta^*}(Y) = V^{U*}_{\delta^*}(Y), \qquad Y \in \mathcal{Y}^*_{\mu^*\delta^*}.$$

This equation specializes to

$$V^{U*}_{\delta^*}(\{y\}) = V^U_{\mu^*\delta^*}(\{y\}), \qquad y \in \mathcal{Y}, \mathrm{P}_{\mu^*\delta^*}\left(\mathbf{y}_0 = y\right) > 0,$$

for sequences of length 0, from which $J^{U*}_{\delta^*} = J^U_{\mu^*\delta^*}$ follows because of (33) and (24). This concludes the proof of (5). Equation (6) can be proved similarly by using (49) and reversing the roles of the players. We then conclude that (μ^*, δ^*) is a Nash equilibrium.

To prove the converse statement, suppose that (μ^*, δ^*) is a Nash equilibrium. This means, in particular, that μ^* minimizes the cost $J^U_{\mu\delta^*}$. From (i) and (iii) in Theorem 3, it then follows that

$$V^{U*}_{\delta^*}(Y) = T^U_{\delta^*} V^{U*}_{\delta^*}(Y) = T^U_{\mu^*\delta^*} V^{U*}_{\delta^*}(Y), \qquad Y \in \mathcal{Y}^*_{\mu^*\delta^*}.$$

Similarly,

$$V^{D*}_{\mu^*}(Z) = T^D_{\mu^*} V^{D*}_{\mu^*}(Z) = T^D_{\mu^*\delta^*} V^{D*}_{\mu^*}(Z), \qquad Z \in \mathcal{Z}^*_{\mu^*\delta^*}.$$

Equations (48)–(49) are then satisfied by $V^U = V^{U*}_{\delta^*}$ and $V^D = V^{D*}_{\mu^*}$. ∎

5 Conclusions

In this paper we showed that, for a fairly general class of two-player partial-information finite games on Markov chains, Nash equilibria always exist in stochastic behavioral policies. These are policies in which, at each instant of time, both players choose actions randomly according to given probability distributions. These policies are called behavioral because the distributions are a function of the observations and the past actions of the players.

The technique used to prove that a Nash equilibrium exists in stochastic behavioral policies is constructive but has severe limitations because it involves solving an extremely large matrix game. To alleviate this problem we derives a dynamic-programming-like condition that is necessary and sufficient for a pair of stochastic behavioral policies to be a Nash equilibrium. This condition automatically gives Nash equilibria in stochastic behavioral policies when a pair of "cost-to-go" functions can be found that satisfy two given inequalities. This paper falls short of actually providing an efficient algorithm to determine the cost-to-go functions for a generic partial information game. This is the subject of our current research.

Appendix

Proof of Lemma 2. We show by induction that

$$\begin{aligned}\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0, \mathbf{s}_1 = s_1, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_t = u_t, \mathbf{d}_t = d_t, \mathbf{Y}_t = Y_t, \mathbf{Z}_t = Z_t) \\ = \prod_{\tau=0}^{t} \mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau)p_Y(y_\tau, s_\tau)p_Z(z_\tau, s_\tau)\,\mathrm{P}(\mathbf{s}_0 = s_0),\end{aligned}$$

for all $t \in \mathcal{T}$. The thesis then follows by setting

$$k_t(s_0, s_1, \ldots, s_t, u_t, d_t, Y_t, Z_t) = \prod_{\tau=0}^{t} p_Y(y_\tau, s_\tau)p_Z(z_\tau, s_\tau)\,\mathrm{P}(\mathbf{s}_0 = s_0).$$

For $t = 0$, we have

$$\begin{aligned}\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0, &\mathbf{u}_0 = u_0, \mathbf{d}_0 = d_0, \mathbf{Y}_0 = \{y_0\}, \mathbf{Z}_0 = \{z_0\}) \\ &= \mathrm{P}_{\mu\delta}(\mathbf{u}_0 = u_0 | \mathbf{Y}_0 = \{y_0\}, \cdots)\,\mathrm{P}_{\mu\delta}(\mathbf{d}_0 = d_0 | \mathbf{Z}_0 = \{z_0\}, \cdots) \\ &\quad \mathrm{P}_{\mu\delta}(\mathbf{y}_0 = y_0 | \mathbf{s}_0 = s_0, \cdots)\,\mathrm{P}_{\mu\delta}(\mathbf{z}_0 = z_0 | \mathbf{s}_0 = s_0)\,\mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0) \\ &= \mu_{u_0}(\{y_0\})\delta_{d_0}(\{z_0\})p_Y(y_0, s_0)p_Z(z_0, s_0)\,\mathrm{P}(\mathbf{s}_0 = s_0),\end{aligned}$$

where the $\cdots$ denote events which do not affect the probability, due to conditional independence. For $t > 1$, we have

$$\begin{aligned}\mathrm{P}_{\mu\delta}(&\mathbf{s}_0 = s_0, \mathbf{s}_1 = s_1, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_t = u_t, \mathbf{d}_t = d_t, \mathbf{Y}_t = Y_t, \mathbf{Z}_t = Z_t) \\ &= \mu_{u_t}(Y_t)\delta_{d_t}(Z_t)p_Y(y_t, s_t)p_Z(z_t, s_t) \\ &\quad \mathrm{P}_{\mu\delta}(\mathbf{s}_0 = s_0, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_{t-1} = u_{t-1}, \mathbf{d}_{t-1} = d_{t-1}, \mathbf{Y}_{t-1} = Y_{t-1}, \mathbf{Z}_{t-1} = Z_{t-1}).\end{aligned}$$

By applying the induction hypothesis, we then get

$$\begin{aligned}\mathrm{P}_{\mu\delta}(&\mathbf{s}_0 = s_0, \ldots, \mathbf{s}_t = s_t, \mathbf{u}_t = u_t, \mathbf{d}_t = d_t, \mathbf{Y}_t = Y_t, \mathbf{Z}_t = Z_t) \\ &= \mu_{u_t}(Y_t)\delta_{d_t}(Z_t)p_Y(y_t, s_t)p_Z(z_t, s_t) \prod_{\tau=0}^{t-1} \mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau)p_Y(y_\tau, s_\tau)p_Z(z_\tau, s_\tau)\,\mathrm{P}(\mathbf{s}_0 = s_0)\end{aligned}$$

$$= \prod_{\tau=0}^{t} \mu_{u_\tau}(Y_\tau)\delta_{d_\tau}(Z_\tau)p_Y(y_\tau, s_\tau)p_Z(z_\tau, s_\tau)\,\mathrm{P}(\mathbf{s}_0 = s_0),$$

which concludes the proof. ■

Lemma 3. *[19, p. 216] Given two random variables $\boldsymbol{\xi}$, $\boldsymbol{\eta}$, and an event $A \in \mathcal{F}$,*

$$\mathrm{E}[\boldsymbol{\xi}|A] = \mathrm{E}[f(\boldsymbol{\eta})|A], \qquad f(y) := \mathrm{E}[\boldsymbol{\xi}|\boldsymbol{\eta} = y, A].$$

The proof of the following Lemma was omitted due to space limitations. See [18] for details.

Lemma 4. *Fix a policy $\delta \in \Pi_D$, pick any sequence $Y \in \mathcal{Y}^*$, and set $\tau := \ell(Y)$. Then, the following holds true:*

(i) For each $\mu \in \Pi_U$ such that $\mathrm{P}_{\mu\delta}(\boldsymbol{Y}_\tau = Y) > 0$,

$$\mathrm{P}_{\mu\delta}(\boldsymbol{s}_\tau = s, \boldsymbol{Z}_\tau = Z \,\big|\, \boldsymbol{Y}_\tau = Y) = I_\delta(s, Z, Y), \qquad s \in \mathcal{S}, Z \in \mathcal{Z}^*,$$

where $I_\delta(s, Y, Z)$ is given by equation (31) initialized with (30).

(ii) For each $\mu \in \Pi_U$ such that $\mathrm{P}_{\mu\delta}(\boldsymbol{Y}_\tau = Y) > 0$,

$$\mathrm{P}_{\mu\delta}\left(\mathbf{y}_{\tau+1} = y, \mathbf{s}_{\tau+1} = s', \mathbf{s}_\tau = s, \mathbf{u}_\tau = u, \mathbf{d}_\tau = d, \mathbf{Z}_\tau = Z \,\middle|\, \mathbf{Y}_\tau = Y\right) = p_Y(y, s')p(s \xrightarrow{ud} s')\mu_u(Y)\delta_d(Z)I_\delta(s, Z, Y),$$

for any $y \in \mathcal{Y}, s, s' \in \mathcal{S}, u \in \mathcal{U}, d \in \mathcal{D}, Z \in \mathcal{Z}^*$, where $I_\delta(s, Y, Z)$ is given by equation (31) initialized with (30).

(iii) For $\mu, \bar{\mu} \in \Pi_U$ such that $\mathrm{P}_{\mu\delta}(\boldsymbol{Y}_\tau = Y) > 0$, $\mathrm{P}_{\bar{\mu}\delta}(\boldsymbol{Y}_\tau = Y) > 0$, and $\mu(\bar{Y}) = \bar{\mu}(\bar{Y})$, for every $\bar{Y} \in \mathcal{Y}^*$ with length $\ell(\bar{Y})$ no smaller than τ, whose first $2\tau + 1$ elements are equal to Y,

$$\mathrm{P}_{\bar{\mu}\delta}\left(\boldsymbol{u}_\tau = u_\tau, \boldsymbol{d}_\tau = d_\tau, \boldsymbol{s}_\tau = s_\tau, \ldots, \boldsymbol{u}_T = u_T, \boldsymbol{d}_T = d_T, \boldsymbol{s}_T = s_T \,\middle|\, \mathbf{Y}_\tau = Y\right) = \mathrm{P}_{\mu\delta}\left(\boldsymbol{u}_\tau = u_\tau, \boldsymbol{d}_\tau = d_\tau, \boldsymbol{s}_\tau = s_\tau, \ldots, \boldsymbol{u}_T = u_T, \boldsymbol{d}_T = d_T, \boldsymbol{s}_T = s_T \,\middle|\, \mathbf{Y}_\tau = Y\right).$$

(iv) For $\bar{\mu}, \mu \in \Pi_U$ such that $\mathrm{P}_{\mu\delta}(\boldsymbol{Y}_\tau = Y) > 0$, $\mathrm{P}_{\bar{\mu}\delta}(\boldsymbol{Y}_\tau = Y) > 0$, and $\bar{\mu}(\bar{Y}) = \mu(\bar{Y})$ for every $\bar{Y} \in \mathcal{Y}^*$ with length $\ell(\bar{Y})$ no smaller than τ, whose first $2\tau+1$ elements are equal to Y, then $V^U_{\bar{\mu}\delta}(Y) = V^U_{\mu\delta}(Y)$.

(v) For $\mu, \bar{\mu} \in \Pi_U$ such that $\mu(\bar{Y}) = \bar{\mu}(\bar{Y})$, $\bar{Y} \in \mathcal{Y}^*, \ell(\bar{Y}) < \tau$, $\mathrm{P}_{\bar{\mu}\delta}\left(\mathbf{Y}_\tau = Y\right) = \mathrm{P}_{\mu\delta}\left(\mathbf{Y}_\tau = Y\right)$.

References

[1] R. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[2] S. Sorin and S. Zamir, ""Big Match" with lack of information on one side (III)," in Raghavan *et al.* [20], pp. 101–112.

[3] C. Melolidakis, "Stochastic games with lack of information on one side and positive stop probabilities," in Raghavan *et al.* [20], pp. 113–126.

[4] J. P. Hespanha, M. Prandini, and S. Sastry, "Probabilistic pursuit-evasion games: A one-step Nash approach," in *Proc. of the 39th Conf. on Decision and Contr.*, Dec. 2000.

[5] J. P. Hespanha, Y. S. Ateşkan, and H. H. Kızılocak, "Deception in non-cooperative games with partial information," in *Proc. of the 2nd DARPA-JFACC Symp. on Advances in Enterprise Control*, July 2000.

[6] G. Kimeldorf, "Duels: An overview," in *Mathematics of Conflict* (M. Shubik, ed.), pp. 55–72, Amsterdam: North-Holland, 1983.

[7] P. Bernhard, A.-L. Colomb, and G. P. Papavassilopoulos, "Rabbit and hunter game: Two discrete stochastic formulations," *Comput. Math. Applic.*, vol. 13, no. 1–3, pp. 205–225, 1987.

[8] G. J. Olsder and G. P. Papavassilopoulos, "About when to use a searchlight," *J. Mathematical Analysis and Applications*, vol. 136, pp. 466–478, 1988.

[9] G. J. Olsder and G. P. Papavassilopoulos, "A Markov chain game with dynamic information," *J. Optimization Theory and Applications*, vol. 59, pp. 467–486, Dec. 1988.

[10] I. B. Rhodes and D. G. Luenberger, "Differential games with imperfect state information," *IEEE Trans. Automat. Contr.*, vol. AC-14, pp. 29–38, Feb. 1969.

[11] P. R. Kumar and J. H. V. Schuppen, "On Nash equilibrium solutions in stochastic dynamic games," *IEEE Trans. Automat. Contr.*, vol. AC-25, pp. 1146–1149, Dec. 1980.

[12] P. R. Kumar, "A differential game with jump process observations," *J. Optimization Theory and Applications*, vol. 31, June 1980.

[13] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York: Spinger-Verlag, 1997.

[14] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. No. 23 in Classics in Applied Mathematics, Philadelphia: SIAM, 2nd ed., 1999.

[15] J. Nash, "Non-cooperative games," *Annals of Mathematics*, vol. 54, pp. 286–295, 1951.

[16] H. V. Stackelberg, *Marktform und Gleichgewicht*. Vienna: Springer-Verlag, 1934.

[17] H. V. Stackelberg, *The Theory of the Market Economy*. Oxford: Oxford University Press, 1952. Translation of [16].

[18] J. P. Hespanha and M. Prandini, "Nash equilibria in partial-information games on Markov chains," tech. rep., University of Southern California, Los Angeles, CA, Mar. 2001.

[19] Y. S. Chow and H. Teicher, *Probability Theory: independence, interchangeability, martingales*. Springer Texts in Statistics, New York: Springer-Verlag, 3rd ed., 1997.

[20] T. E. S. Raghavan, T. S. Ferguson, and T. Parthasarathy, eds., *Stochastic Games and Related Topics: In Honor of Professor L. S. Shapley*, vol. 7 of *Theory and Decision Library, Series C, Game Theory, Mathematical Programming and Operations Research*. Dordrecht: Kluwer Academic Publishers, 1991.