Problems of Information Transmission, Vol. 39, No. 3, 2003, pp. 266–293. Translated from Problemy Peredachi Informatsii, No. 3, 2003, pp. 40–71.
Original Russian Text Copyright © 2003 by Golubtsov, Lyubetsky.

Stochastic Dynamic Games with Various Types of Information

P. V. Golubtsov and V. A. Lyubetsky*****

**M.V. Lomonosov Moscow State University*
p_v_g@mail.ru
***Institute for Information Transmission Problems, RAS, Moscow*
Lyubetsk@iitp.ru

Received April 29, 2002

Abstract—Dynamic discrete-time games are generalized to a stochastic environment, in order to examine the influence of various types of information structures on the course of a game. It is shown that the information structure of a game, i.e., type and amount of information available to players and, in particular, asymmetry of information, may lead to unexpected and sometimes counter-intuitive effects on the game result, i.e., the players' payoffs. The paper also develops algorithms for obtaining the Nash equilibrium strategies in such games. These involve reducing optimal reaction policies to the corresponding dynamic programming algorithms and generalizing the classical optimal control technique. Results of computer simulations for a variant of fishery harvesting game are presented.

1. INTRODUCTION

The paper studies a class of stochastic dynamic discrete-time games with explicit account of information available to the players.

We explore a wide range of information structures in our games in order to study the *role of information available to players in constructing optimal strategies* (in other words, we want to study effects of *contextual use of information in the behavior* of players). Thus, we are dealing with the semantic aspect of information (i.e., *semantic information*).

Two independent players perform control of a common discrete-time dynamic system, which is also affected by random disturbing effects. At every time step, the players make decisions based on some information about these effects. Specifically, each player selects his strategy in order to maximize his total discounted payoff for a long enough time period provided that the other player follows some fixed strategy. This means that the course of the game is described by the dynamic Nash equilibrium.

Random effects mentioned above are represented by a random Markov parameter (which may be multidimensional).

We shall explore a wide range of information structures [1] in our games. In particular, players may possess various levels of knowledge about realizations of a random parameter, e.g., current or delayed information [2, 3], or even information obtained from imperfect observation of a random parameter. Moreover, information structure may be asymmetric; e.g., one player may possess full current information while the other has only delayed or imperfect information. We shall also consider cooperative games with different types of information structures, etc.

For each variant of the game, we propose an algorithm that reduces the original problem to a multiple construction of the Nash equilibrium in a certain finite-dimensional space. Moreover, in the most complicated case considered in Section 5.3, we construct an algorithm that reduces the

0032-9460/03/3903-0266 $25.00 © 2003 MAIK "Nauka/Interperiodica"

problem to a multistage construction of the Nash equilibrium points in a function space, which in turn leads to an iterative stepwise optimization procedure (at every stage).

In Sections 4.1, 5.3, and 8, we provide a set of simple propositions, whose major goal is to clarify the main concepts and obtained algorithms. Instead of a theoretical analysis of our algorithms, we performed extensive computations, which demonstrated fast convergence of these algorithms to natural results for a substantial amount of initial data.

Then we present results of computer simulation for games in which information available to players influences their optimal strategies (behavior) in a nontrivial way. The authors believe that examination of effects that demonstrate the influence of available information on the course of the game (i.e., the result of information utilization in the context of "organism activity") clarifies the very concept of "semantic informativeness."

For a simple example of dynamic games and for demonstrative interpretation of results, following the long tradition [4, 5], the specific context is taken to be marine fishery.

The authors consider the following observations (obtained in simulations) of information influence on the players' behavior to be interesting and (to a certain extent) unexpected.

In the game with symmetric current information ("complete information"), reduction of the cost of the harvest may lead to reduction of the average payoff (Fig. 2, the area around $c = 0.3$, solid curve). In the game with minimum information, a similar situation takes place as well. Moreover, there exists an area where minimum information leads to higher payoffs than complete information. However, for the cooperative behavior of players (and symmetric access to information), we obtain natural and expected results.

For all noncooperative games and all types of information structures, an increase in information accuracy (starting from a certain level) leads to a decrease in average payoffs. However, cooperative behavior of players leads to natural results again (see Figs. 3 and 4).

Figure 5 shows that, at certain levels of the environmental variability, information structure may have unexpected influence on a game result. For example, sometimes minimal information may be more advantageous for players than complete information.

Figure 6 demonstrates that additional information is beneficial for players only if their degree of cooperation is high enough.

All the peculiarities of the contextual influence of information mentioned above do not seem to be evident (to the authors) or to follow from some general considerations. Besides, they seem to be of interest for a possible formalization of a (non-Shannon and non-Kolmogorov) concept of information, i.e., *semantic information.*

Let us say a few words about the interpretation (of the problems examined) that the authors kept in mind. Intuitively, we have some apprehension of a certain confrontation and cooperation of two forces, e.g., organism (or cell, population) and environment, individual and society, man and destiny, etc. A cell (say, of a bacteria) confronts the environment and cooperates with it on the basis of the genome, which is a certain list of a great number of instructions on biochemical production (and transportation from the environment) of specific chemical agents (as well as special instructions that are contained in regulatory parts of the genome and control the instructions mentioned above, etc.). Sophisticated complexes of instructions turn on or off simultaneously, depending on the current situation (in the cell or around it). The goal of the cell is maintaining its certain parameters, while the goal of the environment is also to maintain some other parameters for the convenience of population (this may be restrictive for the cell and partially antagonistic to its interests). It is important that both sides at every moment of their action (decision making) possess incomplete, partly common and partly separate (sometimes, contradictory) information. Thus, the question on the character of their cooperation and, in particular, on the (optimal) strategy of information exchange between them is of great evolutional importance.

For example, the environment may evolve a signal about the number of other cells (of the same or other types) in a neighborhood of a given cell (feromon). As a result, the cell may stop splitting, change the virulence, produce an antibiotic (to confront competitors in the environment), collaborate in creating a biofilm, etc.

Another example of complex "game-like" interactions is the sporulation mechanism, when cells of a population form a complex configuration (a head and a stalk). Cells that form the head transform to spores and fly away, while the other cells die. Also mention symbiotic bacteria that live in the luminous organ of a cuttlefish and interact with it; or seaweeds which secret analogs of bacterial feromons in order to destroy a population of bacteria colonizing them.

In these examples, various interaction strategies are possible, depending on the level of interaction: a separate cell and its environment, interaction within a clone (descendants of one cell), within a community of clones, etc. In addition, results of the interaction may be quite unexpected: for example, one successful cell within a population-clone may deceive all the other cells, spawn, and destroy the community; as a result, the whole clone will die out. Besides, interaction between clones may be symbiotic or antagonistic. Priority may be given to a particular person or to the common weal.

It may seem that standard mathematical approaches cannot adequately describe this kind of systems (about which we presently have extensive and concrete scientific knowledge). One can consider a space of states "cell (population)–environment" with transfers from a current state R to the next one R_+, where, for a given state R, its estimation v is computed. An area of this space is attainable depending on the pair $S = \langle S^\alpha, S^\beta \rangle$, where S^α is the reaction of the cell on R^α, while S^β is the result of the environment activity. The authors plan to discuss interpretations of this kind for a cell–environment system in further publications.

The structure of the paper is as follows: Sections 2 and 3 contain the description of a stochastic discrete-time two-player game and a brief problem statement. Sections 4 and 5 contain specification of problem statements and descriptions of algorithms that construct optimal strategies, which give the Nash equilibrium. These algorithms reduce the original problem (of finding optimal control policies) to the corresponding dynamic programming algorithms, generalizing the classical optimal control technique (see, e.g., [6]). Section 6 considers cases of partial or complete cooperation of players. Section 7 presents computer simulation results. Finally, Section 8 contains some general facts about games, which are used in the main body of the paper.

2. DESCRIPTION OF A DYNAMIC GAME

In this section, we give a rather general description of a discrete-time nonantagonistic two-person dynamic game

$$R \xrightarrow{\rho(R,\nu,p^\alpha,p^\beta)} R_+.$$

Here, R and R_+ represent the current and the next-step states of the system respectively, ν is a random disturbance depending on the step (time), and p^α and p^β are decisions (actions) of players. We assume that the next-step system state R_+ and current payoffs v^α and v^β of the players depend on the current state R, decisions p^α and p^β, and realization of the random disturbance ν for the current step, i.e.,

$$R_+ = \rho(R,\nu,p^\alpha,p^\beta),$$
$$v^\alpha = v^\alpha(R,\nu,p^\alpha,p^\beta), \qquad v^\beta = v^\beta(R,\nu,p^\alpha,p^\beta).$$

The players' actions are determined by the current state R of the system and, possibly, by some additional information ξ^α and ξ^β about disturbance ν:

$$p^\alpha = P^\alpha(R, \xi^\alpha), \qquad p^\beta = P^\beta(R, \xi^\beta),$$

where P^α and P^β are the players' control strategies (policies), i.e., functions that depend on the information that players α and β possess, respectively.

In what follows, we assume that random elements ν_t (where t is time) are independent and identically distributed or, in a more general case, constitute a Markov chain with a known transition distribution. In particular, if ν_t assumes a finite number n of values, then its transition distribution is determined by an $n \times n$ stochastic matrix.

3. THE INFINITE HORIZON GAME

Let us briefly describe a formulation of the optimal decision-making (control) problem for a nonantagonistic dynamic game. More precise specifications and technical details of reducing such a dynamic game to the corresponding dynamic programming formulation can be found in Sections 4 and 6.

For the sake of simplicity, in this section we consider a game with an infinite horizon. Typically, in infinite horizon games it is assumed that the situation is stationary, that is, the functions ρ, p^α, and p^β and the annual payoff functions v^α and v^β do not explicitly depend on t. We also assume this in the current section.

Each player's objective in the game is to choose an optimal stationary policy to maximize the expected discounted sum of his annual payoffs given the policy of his competitor. Thus, player α will choose P^α conditional on P^β to maximize

$$U^\alpha(R_0, P^\alpha, P^\beta) = \mathsf{E}\sum_{t=0}^{\infty} \gamma_\alpha^t v^\alpha(R_t^\alpha, p_t^\alpha, p_t^\beta, \nu_t) = \mathsf{E}\sum_{t=0}^{\infty} \gamma_\alpha^t v^\alpha(R_t^\alpha, P^\alpha(R_t, \xi_t^\alpha), P^\beta(R_t, \xi_t^\beta), \nu_t),$$

while simultaneously β attempts to conditionally maximize $U^\beta(R_0, P^\alpha, P^\beta)$. Here, the expectation is taken over all the random variables ν_t and (in the case of imperfect observation) over all the measurements ξ_t^α and ξ_t^β.

Thus, we have a two-player game, which consists in finding a Nash equilibrium solution

$$\begin{cases} \max\limits_{P^\alpha} U^\alpha(R, P^\alpha, P^\beta), \\ \max\limits_{P^\beta} U^\beta(R, P^\alpha, P^\beta) \end{cases}$$

for all possible R. If a pair $\widehat{P}^\alpha, \widehat{P}^\beta$ constitutes the Nash equilibrium then, by the definition, for all possible values of R and all policy functions P^α and P^β, we have

$$\begin{cases} U^\alpha(R, \widehat{P}^\alpha, \widehat{P}^\beta) \geq U^\alpha(R, P^\alpha, \widehat{P}^\beta), \\ U^\beta(R, \widehat{P}^\alpha, \widehat{P}^\beta) \geq U^\beta(R, \widehat{P}^\alpha, P^\beta). \end{cases}$$

The optimal policies $\widehat{P}^\alpha$ and $\widehat{P}^\beta$ (and the corresponding optimal average discounted payoffs $\widehat{U}^\alpha$ and $\widehat{U}^\beta$) can be constructed by taking the limit as $T \to \infty$ for the corresponding game with finite horizon T. Algorithms for the corresponding finite horizon games (and also for the nonstationary case) are considered in Sections 4 and 5.

4. FINITE HORIZON GAME: PLAYERS HAVE RESULTS OF PERFECT OBSERVATIONS

Let us consider a finite horizon game with final season T. Since, in a finite horizon game, different moments are not equivalent, we will not restrict ourselves to stationarity. Thus, the functions v_τ^α and v_τ^β may be different at different moments τ. Denote a decision policy for player α at time τ by P_τ^α and a sequence of decision functions P_τ^α from the moment t to T by

$$\boldsymbol{P}_t^\alpha = \langle P_t^\alpha, P_{t+1}^\alpha, \ldots, P_T^\alpha \rangle = \langle P_t^\alpha, \boldsymbol{P}_{t+1}^\alpha \rangle.$$

Similarly for player β.

In this section, we assume that random elements ν_t form a Markov chain (with a finite or infinite number of states) with a given transition probability.

At time t, each player knows the current state of the system, R_t, and has some information about the stochastic parameter ν_t (or a preceding one $\nu_{t-1} \ldots$). Information on ν_t includes, at the minimum, the probability transition matrix for this Markovian random sequence.

In this section, we consider two simplest cases: at time t, a player knows the realization of the stochastic element ν_{t-1} and all the preceding values $\nu_{t-2}, \ldots$, or also knows the current value of ν, i.e., ν_t.

In the next section, we will study the case where players get imperfect information about ν_t, which is obtained from measurements. The knowledge structure may be symmetric, with both players having the same information, or asymmetric, e.g., when one player has current knowledge of ν_t while the other has only delayed knowledge ν_{t-1}.

4.1. Players Have Current Information

Assume that information held by both players at time t includes the current value of the random element ν_t.

Then, at moment t, the expected discounted payoff for player α is

$$V_t^\alpha(R_t, \nu_t, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta) = \mathsf{E}_{(\nu_{t+1}, \nu_{t+2}, \ldots, \nu_T \,|\, \nu_t)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v_\tau^\alpha(R_\tau, \nu_\tau, p_\tau^\alpha, p_\tau^\beta) \tag{1}$$

(a similar expression for player β), where

$$p_\tau^\alpha = P_\tau^\alpha(R_\tau, \nu_\tau), \qquad p_\tau^\beta = P_\tau^\beta(R_\tau, \nu_\tau).$$

A pair $\langle \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta \rangle$ provides the Nash equilibrium for a pair $\langle V_t^\alpha, V_t^\beta \rangle$ if, for all possible values of R_t and ν_t, we have

$$\begin{cases} V_t^\alpha(R_t, \nu_t, \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta) = \max\limits_{\boldsymbol{P}_t^\alpha} V_t^\alpha(R_t, \nu_t, \boldsymbol{P}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta), \\ V_t^\beta(R_t, \nu_t, \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta) = \max\limits_{\boldsymbol{P}_t^\beta} V_t^\beta(R_t, \nu_t, \widehat{\boldsymbol{P}}_t^\alpha, \boldsymbol{P}_t^\beta). \end{cases} \tag{2}$$

In what follows, we will denote the corresponding Nash equilibrium discounted payoffs as

$$\begin{cases} \widehat{V}_t^\alpha(R_t, \nu_t) = V_t^\alpha(R_t, \nu_t, \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta), \\ \widehat{V}_t^\beta(R_t, \nu_t) = V_t^\beta(R_t, \nu_t, \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta). \end{cases}$$

Since the random parameter ν is a Markov process, which is completely determined by its current value and single-stage transition distribution (i.e., the distribution of ν_{t+1} for any given ν_t), the mathematical expectation $\mathsf{E}_{(\nu_{t+1},\nu_{t+2},\ldots,\nu_T\,|\,\nu_t)}$ can be presented as a sequence of conditional expectations:

$$\mathsf{E}_{(\nu_{t+1},\nu_{t+2},\ldots,\nu_T\,|\,\nu_t)} = \mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)}\mathsf{E}_{(\nu_{t+2}\,|\,\nu_{t+1})}\ldots\mathsf{E}_{(\nu_T\,|\,\nu_{T-1})}.$$

It follows that V_t^α can be expressed through the immediate payoff v_t^α and V_{t+1}^α:

$$\begin{aligned}V_t^\alpha(R_t,\nu_t,\boldsymbol{P}_t^\alpha,\boldsymbol{P}_t^\beta) &= v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta)\\ &\quad+\gamma_\alpha\mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)}V_{t+1}^\alpha(\rho(R_t,\nu_t,p_t^\alpha,p_t^\beta),\nu_{t+1},\boldsymbol{P}_{t+1}^\alpha,\boldsymbol{P}_{t+1}^\beta).\end{aligned}\tag{3}$$

Note that we can also use expression (3) as an alternative (recursive) definition of the discounted payoff.

Expression (3) allows one to reduce problem (2) to a series of considerably simpler problems through a dynamic programming procedure. Assume that $\langle\widehat{\boldsymbol{P}}_{t+1}^\alpha,\widehat{\boldsymbol{P}}_{t+1}^\beta\rangle$ are the Nash equilibrium policies starting from the moment $t+1$ and $\langle\widehat{V}_{t+1}^\alpha,\widehat{V}_{t+1}^\beta\rangle$ are the corresponding optimal discounted payoffs.

By the analogy with classical dynamic programming, define the function

$$\widetilde{V}_t^\alpha(R_t,\nu_t,P_t^\alpha,P_t^\beta) = V_t^\alpha\Big(R_t,\nu_t,\langle P_t^\alpha,\widehat{\boldsymbol{P}}_{t+1}^\alpha\rangle,\langle P_t^\beta,\widehat{\boldsymbol{P}}_{t+1}^\beta\rangle\Big),$$

i.e., the discounted payoff for player α corresponding to arbitrary policies P_t^α and P_t^β at time t and optimal "tails" $\widehat{\boldsymbol{P}}_t^\alpha$ and $\widehat{\boldsymbol{P}}_t^\beta$, and a similar function for player β.

Then the optimal policies $\langle\widehat{P}_t^\alpha,\widehat{P}_t^\beta\rangle$ for time t can be obtained by solving, for all possible values of R_t and ν_t, the Nash equilibrium problem for the functions

$$\begin{cases}\widetilde{V}_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta) = v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta)+\gamma_\alpha\mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)}\widehat{V}_{t+1}^\alpha(\rho(R_t,p_t^\alpha,p_t^\beta,\nu_t),\nu_{t+1}),\\ \widetilde{V}_t^\beta(R_t,\nu_t,p_t^\alpha,p_t^\beta) = v_t^\beta(R_t,\nu_t,p_t^\alpha,p_t^\beta)+\gamma_\beta\mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)}\widehat{V}_{t+1}^\beta(\rho(R_t,p_t^\alpha,p_t^\beta,\nu_t),\nu_{t+1})\end{cases}\tag{4}$$

with respect to $\langle p_t^\alpha,p_t^\beta\rangle$. Namely, $\widehat{P}_t^\alpha(R_t,\nu_t)=\widehat{p}_t^\alpha$ and $\widehat{P}_t^\beta(R_t,\nu_t)=\widehat{p}_t^\beta$, where the pair $\langle\widehat{p}_t^\alpha,\widehat{p}_t^\beta\rangle$ attains the Nash equilibrium for these functions with given R_t and ν_t. Thus, the Nash equilibrium policies can be obtained recursively as

$$\widehat{\boldsymbol{P}}_t^\alpha = \langle\widehat{P}_t^\alpha,\widehat{\boldsymbol{P}}_{t+1}^\alpha\rangle,\qquad \widehat{\boldsymbol{P}}_t^\beta = \langle\widehat{P}_t^\beta,\widehat{\boldsymbol{P}}_{t+1}^\beta\rangle.\tag{5}$$

In order to express this more precisely, assume that $R\in\mathcal{R}$, the space of the system states; $\nu\in\mathcal{N}$, the space of random parameter values; and $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ are the spaces of decisions p^α and p^β of players α and β respectively. We will assume that $\mathcal{R}$, $\mathcal{N}$, $\mathcal{D}^\alpha$, and $\mathcal{D}^\beta$ are measurable spaces and $v_t^\alpha,v_t^\beta\colon\mathcal{R}\times\mathcal{N}\times\mathcal{D}^\alpha\times\mathcal{D}^\beta\to\mathbb{R}$ and $\rho\colon\mathcal{R}\times\mathcal{N}\times\mathcal{D}^\alpha\times\mathcal{D}^\beta\to\mathcal{R}$ are measurable mappings.

At each stage t, the strategy P_t^α of player α is a measurable mapping from $\mathcal{R}\times\mathcal{N}$ to $\mathcal{D}^\alpha$. Denote by $\widetilde{\mathcal{D}}^\alpha$ the space of all measurable mappings $\mathcal{R}\times\mathcal{N}\to\mathcal{D}^\alpha$. Similarly, let $\widetilde{\mathcal{D}}^\beta$ be the space of all measurable mappings $\mathcal{R}\times\mathcal{N}\to\mathcal{D}^\beta$.

Now the game with finite horizon T can be considered as the two-player game

$$G_{t,R,\nu} = \langle\widetilde{\mathcal{D}}_t^\alpha,\widetilde{\mathcal{D}}_t^\beta,V_t^\alpha,V_t^\beta\rangle_{R,\nu},\quad R\in\mathcal{R},\quad \nu\in\mathcal{N},\quad t=1,\ldots,T,$$

with the states of strategies

$$\widetilde{\mathcal{D}}_t^\alpha = \prod_{\tau=t}^T\widetilde{\mathcal{D}}^\alpha,\qquad \widetilde{\mathcal{D}}_t^\beta = \prod_{\tau=t}^T\widetilde{\mathcal{D}}^\beta$$

and with payoff functions V_t^α and V_t^β defined according to (1).

More precisely, at any fixed moment t, we have a family of games $\{G_{t,R,\nu} \mid R \in \mathcal{R},\ \nu \in \mathcal{N}\}$, which is parametrized by R and ν. A pair of strategies $\langle \widehat{\boldsymbol{P}}_t^{\alpha}, \widehat{\boldsymbol{P}}_t^{\beta} \rangle$ is optimal for this family of games (is a Nash equilibrium point) if, for any strategies $\boldsymbol{P}_t^{\alpha}$ and $\boldsymbol{P}_t^{\beta}$ and any $R \in \mathcal{R}$ and $\nu \in \mathcal{N}$, we have the following inequalities:

$$\begin{aligned}V_t^{\alpha}(R,\nu,\widehat{\boldsymbol{P}}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta}) \geq V_t^{\alpha}(R,\nu,\boldsymbol{P}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta}),\\ V_t^{\beta}(R,\nu,\widehat{\boldsymbol{P}}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta}) \geq V_t^{\beta}(R,\nu,\widehat{\boldsymbol{P}}_t^{\alpha},\boldsymbol{P}_t^{\beta}).\end{aligned}$$

Now let us consider the family of games determined by the payoff functions $\widetilde{V}_t^{\alpha}$ and $\widetilde{V}_t^{\beta}$,

$$\widetilde{G}_{t,R,\nu} = \langle \mathcal{D}^{\alpha}, \mathcal{D}^{\beta}, \widetilde{V}_t^{\alpha}, \widetilde{V}_t^{\beta} \rangle_{R,\nu}, \quad R \in \mathcal{R}, \quad \nu \in \mathcal{N}, \quad t = 1,\ldots,T.$$

The following proposition reduces the problem of constructing optimal strategies for the game $G_{t,R,\nu}$ to a series of constructions of optimal strategies for the simpler games $\widetilde{G}_{t,R,\nu}$.

Proposition 4.1. *Assume that, for all t starting from $t = T$ down to $t = 1$ and for arbitrary $R \in \mathcal{R}$ and $\nu \in \mathcal{N}$, the game $\widetilde{G}_{t,R,\nu}$ has a Nash equilibrium point $\langle \widehat{p}^{\alpha}_{R,\nu}, \widehat{p}^{\beta}_{R,\nu} \rangle$, and that the mappings $\widehat{P}_t^{\alpha}$ and $\widehat{P}_t^{\beta}$ defined as*

$$\widehat{P}_t^{\alpha}(R,\nu) = \widehat{p}^{\alpha}_{R,\nu}, \qquad \widehat{P}_t^{\beta}(R,\nu) = \widehat{p}^{\beta}_{R,\nu}, \qquad \forall R \in \mathcal{R}, \quad \forall \nu \in \mathcal{N}, \tag{6}$$

are measurable with respect to R and ν. Then, for each t, there exists a pair of strategies $\langle \widehat{\boldsymbol{P}}_t^{\alpha}, \widehat{\boldsymbol{P}}_t^{\beta} \rangle$ which provides a Nash equilibrium point for the family of games $\{G_{t,R,\nu} \mid R \in \mathcal{R},\ \nu \in \mathcal{N}\}$.

The optimal strategies $\widehat{\boldsymbol{P}}_t^{\alpha}$ and $\widehat{\boldsymbol{P}}_t^{\beta}$ are determined recursively by expressions (5) *and* (6) *through the optimal strategies $\widehat{\boldsymbol{P}}_{t+1}^{\alpha}$ and $\widehat{\boldsymbol{P}}_{t+1}^{\beta}$ at the next moment and through the equilibrium points $\langle \widehat{p}^{\alpha}_{R,\nu}, \widehat{p}^{\beta}_{R,\nu} \rangle$ for the games $\widetilde{G}_{t,R,\nu}$.*

The payoff functions $\widetilde{V}_t^{\alpha}$ and $\widetilde{V}_t^{\beta}$ for the game $\widetilde{G}_{t,R,\nu}$ are expressed through the payoff functions at the next moment (i.e., for the game $\widetilde{G}_{t+1,R,\nu}$) by formulas (4) *and the equalities*

$$\begin{aligned}\widehat{V}_t^{\alpha}(R,\nu) = \widetilde{V}_t^{\alpha}(R,\nu,\widehat{p}^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}),\\ \widehat{V}_t^{\beta}(R,\nu) = \widetilde{V}_t^{\beta}(R,\nu,\widehat{p}^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}).\end{aligned}$$

Proof. To prove the proposition, it suffices to verify the following "inductive" assertion for any $t = 1,\ldots,T$: If there exists a Nash equilibrium point $\langle \widehat{\boldsymbol{P}}_{t+1}^{\alpha}, \widehat{\boldsymbol{P}}_{t+1}^{\beta} \rangle$ for the family of games $\{G_{t+1,R,\nu} \mid R \in \mathcal{R},\ \nu \in \mathcal{N}\}$ and, for any $R \in \mathcal{R}$ and $\nu \in \mathcal{N}$, there exists a Nash equilibrium point $\langle \widehat{p}^{\alpha}_{R,\nu}, \widehat{p}^{\beta}_{R,\nu} \rangle$ for the game $\widetilde{G}_{t,R,\nu}$, then $\langle \widehat{\boldsymbol{P}}_t^{\alpha}, \widehat{\boldsymbol{P}}_t^{\beta} \rangle$ is a Nash equilibrium point for the family of games $\{G_{t,R,\nu} \mid R \in \mathcal{R},\ \nu \in \mathcal{N}\}$, where $\widehat{\boldsymbol{P}}_t^{\alpha}$ and $\widehat{\boldsymbol{P}}_t^{\beta}$ are determined by formulas (5) and (6).

Now, assume that, for some moment t, the pair $\langle \widehat{\boldsymbol{P}}_{t+1}^{\alpha}, \widehat{\boldsymbol{P}}_{t+1}^{\beta} \rangle$ determines optimal strategies for the family of games $\{G_{t+1,R,\nu} \mid R \in \mathcal{R},\ \nu \in \mathcal{N}\}$. Let $\boldsymbol{P}_t^{\alpha}$ be an arbitrary strategy for player α. Consider the payoff function $V_t^{\alpha}(R_t,\nu_t,\boldsymbol{P}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta})$. Using equation (3), we get

$$\begin{aligned}V_t^{\alpha}(R,\nu,\boldsymbol{P}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta}) &= v_t^{\alpha}(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}) + \gamma_{\alpha}\mathsf{E}_{(\nu_+\,|\,\nu)}V_{t+1}^{\alpha}(\rho(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}),\nu_+,\boldsymbol{P}_{t+1}^{\alpha},\widehat{\boldsymbol{P}}_{t+1}^{\beta})\\ &\leq v_t^{\alpha}(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}) + \gamma_{\alpha}\mathsf{E}_{(\nu_+\,|\,\nu)}V_{t+1}^{\alpha}(\rho(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}),\nu_+,\widehat{\boldsymbol{P}}_{t+1}^{\alpha},\widehat{\boldsymbol{P}}_{t+1}^{\beta})\\ &= v_t^{\alpha}(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}) + \gamma_{\alpha}\mathsf{E}_{(\nu_+\,|\,\nu)}\widehat{V}_{t+1}^{\alpha}(\rho(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}),\nu_+)\\ &= \widetilde{V}_t^{\alpha}(R,\nu,p^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}) \leq \widetilde{V}_t^{\alpha}(R,\nu,\widehat{p}^{\alpha}_{R,\nu},\widehat{p}^{\beta}_{R,\nu}) = V_t^{\alpha}(R,\nu,\widehat{\boldsymbol{P}}_t^{\alpha},\widehat{\boldsymbol{P}}_t^{\beta}).\end{aligned}$$

Here, $p^\alpha_{R,\nu} = P^\alpha_t(R,\nu)$.

Similarly, $V^\beta_t(R,\nu,\widehat{\boldsymbol{P}}^\alpha_t,\boldsymbol{P}^\beta_t) \le V^\beta_t(R,\nu,\widehat{\boldsymbol{P}}^\alpha_t,\widehat{\boldsymbol{P}}^\beta_t)$ for any strategy $\boldsymbol{P}^\beta_t$ of player β.

This yields that the pair $\langle \widehat{\boldsymbol{P}}^\alpha_t,\widehat{\boldsymbol{P}}^\beta_t\rangle$ is a Nash equilibrium point in the game $G_{t,R,\nu}$ for all $R \in \mathcal{R}$ and $\nu \in \mathcal{N}$. △

Note that the Nash equilibrium points $\langle \widehat{p}^\alpha_{R,\nu},\widehat{p}^\beta_{R,\nu}\rangle$ for the games $\widetilde{G}_{t,R,\nu}$ can be constructed by the use of the iteration procedure described in Proposition 8.1.

4.2. Players Obtain Delayed Information

This case looks very similar to the case of current knowledge, except for the fact that the policies and discounted values at time t depend on not ν_t but ν_{t-1}, i.e.,

$$p^\alpha_t = P^\alpha_t(R_t,\nu_{t-1}), \qquad p^\beta_t = P^\beta_t(R_t,\nu_{t-1}),$$

and

$$\begin{aligned} V^\alpha_t(R_t,\nu_{t-1},\boldsymbol{P}^\alpha_t,\boldsymbol{P}^\beta_t) &= \mathsf{E}_{(\nu_t,\nu_{t+1},\ldots,\nu_T\,|\,\nu_{t-1})}\sum_{\tau=t}^{T}\gamma_\alpha^{\tau-t}v^\alpha_\tau(R_\tau,\nu_\tau,p^\alpha_\tau,p^\beta_\tau) \\ &= \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\mathsf{E}_{(\nu_{t+1},\ldots,\nu_T\,|\,\nu_t)}\sum_{\tau=t}^{T}\gamma_\alpha^{\tau-t}v^\alpha_\tau(R_\tau,\nu_\tau,p^\alpha_\tau,p^\beta_\tau) \\ &= \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\dot{V}^\alpha_t(R_t,\nu_t,\boldsymbol{P}^\alpha_t,\boldsymbol{P}^\beta_t), \end{aligned}$$

where $\dot{V}^\alpha_t$ denotes the discounted value function for the "current information" case.

Thus, the case where both players have delayed information leads to the dynamic programming procedure with the following Nash equilibrium problem at each step:

$$\begin{cases} \widetilde{V}^\alpha_t(R_t,\nu_{t-1},p^\alpha_t,p^\beta_t) = \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\Big[v^\alpha_t(R_t,\nu_t,p^\alpha_t,p^\beta_t) + \gamma_\alpha\widehat{V}^\alpha_{t+1}(\rho(R_t,p^\alpha_t,p^\beta_t,\nu_t),\nu_t)\Big], \\ \widetilde{V}^\beta_t(R_t,\nu_{t-1},p^\alpha_t,p^\beta_t) = \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\Big[v^\beta_t(R_t,\nu_t,p^\alpha_t,p^\beta_t) + \gamma_\beta\widehat{V}^\beta_{t+1}(\rho(R_t,p^\alpha_t,p^\beta_t,\nu_t),\nu_t)\Big], \end{cases}$$

where $\widehat{V}^\alpha_t(R_t,\nu_{t-1})$ and $\widehat{V}^\beta_t(R_t,\nu_{t-1})$ are the Nash equilibrium values for these functions. Optimal decision policies $\widehat{P}^\alpha_t$ and $\widehat{P}^\beta_t$ are defined as $\widehat{P}^\alpha_t(R_t,\nu_{t-1}) = \widehat{p}^\alpha_t$ and $\widehat{P}^\beta_t(R_t,\nu_{t-1}) = \widehat{p}^\beta_t$, where the pair $\langle\widehat{p}^\alpha_t,\widehat{p}^\beta_t\rangle$ attains the Nash equilibrium for these functions with given R_t and ν_{t-1}.

4.3. Players Get Asymmetric Information: Current Versus Delayed

Now assume that the first player has current knowledge of ν and the second one has delayed information. Thus, the first player's policy depends on ν_t and ν_{t-1}, while the second player's policy depends on ν_{t-1} only, i.e.,

$$p^\alpha_t = P^\alpha_t(R_t,\nu_{t-1},\nu_t), \qquad p^\beta_t = P^\beta_t(R_t,\nu_{t-1}).$$

In this case, at each dynamic programming step, we have the following (asymmetric) Nash equilibrium problem:

$$\begin{cases} \widetilde{V}^\alpha_t(R_t,\nu_{t-1},\nu_t,p^\alpha_t,p^\beta_t) = v^\alpha_t(R_t,\nu_t,p^\alpha_t,p^\beta_t) + \gamma_\alpha\mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)}\widehat{V}^\alpha_{t+1}(\rho(R_t,p^\alpha_t,p^\beta_t,\nu_t),\nu_{t+1}), \\ \widetilde{V}^\beta_t(R_t,\nu_{t-1},p^\alpha_t,p^\beta_t) \qquad = \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\Big[v^\beta_t(R_t,\nu_t,p^\alpha_t,p^\beta_t) + \gamma_\beta\widehat{V}^\beta_{t+1}(\rho(R_t,p^\alpha_t,p^\beta_t,\nu_t),\nu_t)\Big]. \end{cases}$$

Here, the first player utilizes the knowledge of ν_{t-1} to compute the second player's policy at time t. However, while the second player can calculate the first player's policy, he cannot know his opponent's actual response (since he does not know ν_t). Instead, he can only assign to it a (conditional) probability distribution of ν_t.

Strictly speaking, a solution to this asymmetric Nash equilibrium problem is a pair of functions $\widehat{P}_t^\alpha(R_t,\nu_{t-1},\nu_t)$ and $\widehat{P}_t^\beta(R_t,\nu_{t-1})$ which, among all functions $P_t^\alpha(R_t,\nu_{t-1},\nu_t)$ and $P_t^\beta(R_t,\nu_{t-1})$, for all values of R_t, ν_{t-1}, and ν_t, attain the Nash equilibrium for the following pair of functions:

$$\begin{cases} \widetilde{V}_t^\alpha(R_t,\nu_{t-1},\nu_t,P_t^\alpha,P_t^\beta) = v_t^\alpha(R_t,\nu_t,P_t^\alpha(R_t,\nu_{t-1},\nu_t),P_t^\beta(R_t,\nu_{t-1})) \\ \qquad\qquad + \gamma_\alpha \mathsf{E}_{(\nu_{t+1}\,|\,\nu_t)} \widehat{V}_{t+1}^\alpha(\rho(R_t,\nu_t,P_t^\alpha(R_t,\nu_{t-1},\nu_t),P_t^\beta(R_t,\nu_{t-1})),\nu_{t+1}), \\ \widetilde{V}_t^\beta(R_t,\nu_{t-1},P_t^\alpha,P_t^\beta) \quad = \mathsf{E}_{(\nu_t\,|\,\nu_{t-1})}\Big[v_t^\beta(R_t,\nu_t,P_t^\alpha(R_t,\nu_{t-1},\nu_t),P_t^\beta(R_t,\nu_{t-1})) \\ \qquad\qquad + \gamma_\beta \widehat{V}_{t+1}^\beta(\rho(R_t,\nu_t,P_t^\alpha(R_t,\nu_{t-1},\nu_t),P_t^\beta(R_t,\nu_{t-1})),\nu_t)\Big]. \end{cases}$$

Note that each player can adjust his policy "pointwise," i.e., for all possible values of his policy arguments, provided that the other player's policy is fixed. This, in fact, can be used for computing optimal policies iteratively. Namely, for some α-policy, we can find an optimum response β-policy. Then we fix this β-policy and find the corresponding optimum α-policy, and so on.

5. FINITE HORIZON GAME: PLAYERS OBTAIN RESULTS OF IMPERFECT OBSERVATIONS

Our assumption in Section 4 that the players possess precise knowledge of the realization of a stochastic element ν_t is clearly idealization. Typically, its value cannot be determined precisely but only with a certain error. In this section, we will introduce the notion of a measurement error in observation of stochastic parameters. In particular, this allows us to formalize various levels of players' possession of information.

An imperfect observation of ν_t can be characterized through the transition probability from the space of states of the parameter ν_t to the space of states of the observation ξ_t. In the case where these spaces are finite, say, the numbers of states for ν_t and ξ_t are n and m, respectively, the measurement is completely determined by an $m \times n$ stochastic matrix. The ith column of this matrix represents the conditional distribution of the observation ξ_t when ν_t is in the ith state.

Thus, assume that information that a player possesses at time t consists of the current system state R_t and the result of the imperfect measurement ξ_t of the current parameter ν_t. Different players may have results of different measurements ξ^α and ξ^β, or these may be results of the same measurement. Now, policies of players α and β at time t depend on R_t and on ξ_t^α or ξ_t^β respectively; that is,

$$p_t^\alpha = P_t^\alpha(R_t,\xi_t^\alpha), \qquad p_t^\beta = P_t^\beta(R_t,\xi_t^\beta).$$

If players have the same measurement information, i.e., $\xi_t^\alpha = \xi_t^\beta$, we will denote it by ξ_t. Besides, we will assume that the random elements ν are independent and identically distributed.

5.1. Players Have Symmetric Information: Optimization of Average Payoff

Assume that the players' policies depend on the common measurement ξ_t:

$$p_t^\alpha = P_t^\alpha(R_t,\xi_t), \qquad p_t^\beta = P_t^\beta(R_t,\xi_t),$$

and each of them maximizes his average discounted payoff U^α and U^β, where

$$U_t^\alpha(R_t,\boldsymbol{P}_t^\alpha,\boldsymbol{P}_t^\beta) = \mathsf{E}_{(\nu_t,\xi_t,\nu_{t+1},\xi_{t+1},\ldots,\nu_T,\xi_T)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v_\tau^\alpha(R_\tau,\nu_\tau,P_\tau^\alpha(R_\tau,\xi_\tau),P_\tau^\beta(R_\tau,\xi_\tau)). \qquad (7)$$

As in the previous cases, we can express U_t^α through the immediate payoff v_t^α and the next-step average discounted payoff U_{t+1}^α, thus obtaining the following recursive expression:

$$\begin{aligned}
U_t^\alpha(R_t,\boldsymbol{P}_t^\alpha,\boldsymbol{P}_t^\beta) &= \mathsf{E}_{\nu_t}\mathsf{E}_{(\xi_t\,|\,\nu_t)}\mathsf{E}_{\nu_{t+1}}\mathsf{E}_{(\xi_{t+1}\,|\,\nu_{t+1})}\ldots\mathsf{E}_{\nu_T}\mathsf{E}_{(\xi_T\,|\,\nu_T)}\sum_{\tau=t}^{T}\gamma_\alpha^{\tau-t}v_\tau^\alpha(R_\tau,\nu_\tau,p_\tau^\alpha,p_\tau^\beta)\\
&= \mathsf{E}_{\nu_t}\mathsf{E}_{(\xi_t\,|\,\nu_t)}\Big[v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta)\\
&\quad + \gamma_\alpha\mathsf{E}_{(\xi_{t+1},\nu_{t+1})}\ldots\mathsf{E}_{(\xi_T,\nu_T)}\sum_{\tau=t+1}^{T}\gamma_\alpha^{\tau-(t+1)}v_\tau^\alpha(R_\tau,\nu_\tau,p_\tau^\alpha,p_\tau^\beta)\Big]\\
&= \mathsf{E}_{\nu_t}\mathsf{E}_{(\xi_t\,|\,\nu_t)}\Big[v_t^\alpha(R_t,\nu_t,P_t^\alpha(R_t,\xi_t),P_t^\beta(R_t,\xi_t))\\
&\quad + \gamma_\alpha U_{t+1}^\alpha(\rho_t(R_t,\nu_t,P_t^\alpha(R_t,\xi_t),P_t^\beta(R_t,\xi_t)),\boldsymbol{P}_{t+1}^\alpha,\boldsymbol{P}_{t+1}^\beta)\Big].
\end{aligned}$$

Denote, as usual, by $\widehat{\boldsymbol{P}}_t^\alpha$ and $\widehat{\boldsymbol{P}}_t^\beta$ the optimal policy sequences from the moment t and by $\widehat{U}_t^\alpha(R_t)$ and $\widehat{U}_t^\beta(R_t)$ the corresponding average discounted payoffs, i.e.,

$$\widehat{U}_t^\alpha(R_t) = U_t^\alpha(R_t,\widehat{\boldsymbol{P}}_t^\alpha,\widehat{\boldsymbol{P}}_t^\beta),\qquad \widehat{U}_t^\beta(R_t) = U_t^\beta(R_t,\widehat{\boldsymbol{P}}_t^\alpha,\widehat{\boldsymbol{P}}_t^\beta).$$

Let $\widetilde{U}_t^\alpha(R_t,P_t^\alpha,P_t^\beta)$ and $\widetilde{U}_t^\beta(R_t,P_t^\alpha,P_t^\beta)$ be the average discounted payoffs for optimal "tail" policies, i.e.,

$$\begin{aligned}
\widetilde{U}_t^\alpha(R_t,P_t^\alpha,P_t^\beta) &= U_t^\alpha(R_t,\langle P_t^\alpha,\widehat{\boldsymbol{P}}_{t+1}^\alpha\rangle,\langle P_t^\alpha,\widehat{\boldsymbol{P}}_{t+1}^\alpha\rangle)\\
&= \mathsf{E}_{\nu_t}\mathsf{E}_{(\xi_t\,|\,\nu_t)}\Big[v_t^\alpha(R_t,\nu_t,P_t^\alpha(R_t,\xi_t),P_t^\beta(R_t,\xi_t))\\
&\quad + \gamma_\alpha\widehat{U}_{t+1}^\alpha(\rho_t(R_t,\nu_t,P_t^\alpha(R_t,\xi_t),P_t^\beta(R_t,\xi_t)))\Big].
\end{aligned}$$

Thus, the optimum policies $\widehat{P}_t^\alpha$ and $\widehat{P}_t^\beta$ at moment t attain the Nash equilibrium for the pair of functions

$$\begin{cases}\widetilde{U}_t^\alpha(R_t,P_t^\alpha,P_t^\beta),\\ \widetilde{U}_t^\beta(R_t,P_t^\alpha,P_t^\beta).\end{cases}\tag{8}$$

These optimal policies can be found iteratively, pointwise in R (but not in ξ), by using the following recursive procedure: For a given state R and some fixed policy $P_{(1)}^\alpha$ for player α, find the optimal response $P_{(1)}^\beta$ for player β, then find the optimal response $P_{(2)}^\alpha$ for α with respect to $P_{(1)}^\beta$, etc., i.e.,

$$\begin{aligned}
P_{(i)}^\beta &= \arg\max_{P^\beta}\widetilde{U}_t^\beta(R,P_{(i)}^\alpha,P^\beta),\\
P_{(i+1)}^\alpha &= \arg\max_{P^\alpha}\widetilde{U}_t^\alpha(R,P^\alpha,P_{(i)}^\beta).
\end{aligned}$$

Note that, at each iteration step, we have to find the whole function, e.g., $P_{(i)}^\alpha(R,\xi)$, which, in fact, can be considered as a function of one variable ξ (since we can fix R but have to take the average over all values of ξ).

However, it is still possible to specify a pointwise procedure for obtaining strategies that provide the Nash equilibrium in function spaces. In other words, we propose a way of reducing the construction of the Nash equilibrium in function spaces to a similar one in finite-dimensional arithmetic spaces. To do this, define

$$W_t^\alpha(R,\nu,\xi,P^\alpha,P^\beta) = v_t^\alpha(R,\nu,P^\alpha(R,\xi),P^\beta(R,\xi)) + \gamma_\alpha\widehat{U}_{t+1}^\alpha(\rho_t(R,\nu,P^\alpha(R,\xi),P^\beta(R,\xi)))$$

and a similar function $W_t^\beta(R,\nu,\xi,P^\alpha,P^\beta)$ for player β. Then the function $\widetilde{U}_t^\alpha$ can be written in the following form:

$$\begin{aligned}\widetilde{U}_t^\alpha(R,P^\alpha,P^\beta) &= \mathsf{E}_\nu \mathsf{E}_{(\xi\,|\,\nu)} W_t^\alpha(R,\nu,\xi,P^\alpha,P^\beta)\\ &= \mathsf{E}_\xi \mathsf{E}_{(\nu\,|\,\xi)} W_t^\alpha(R,\nu,\xi,P^\alpha,P^\beta),\end{aligned}$$

where $\mathsf{E}_{(\nu\,|\,\xi)}$ is the conditional distribution of ν for a given ξ. Here we use the fact that the mathematical expectation over the joint distribution of ν and ξ can be presented in the forms

$$\mathsf{E}_{(\nu,\xi)} = \mathsf{E}_\nu \mathsf{E}_{(\xi\,|\,\nu)} = \mathsf{E}_\xi \mathsf{E}_{(\nu\,|\,\xi)}.$$

Now assume that, for a fixed ξ, the policies $\widehat{P}^\alpha$ and $\widehat{P}^\beta$ attain the Nash equilibrium for the functions

$$\begin{cases}\widetilde{V}^\alpha(R,\xi,P^\alpha,P^\beta) = \mathsf{E}_{(\nu\,|\,\xi)} W_t^\alpha(R,\nu,\xi,P^\alpha,P^\beta),\\ \widetilde{V}^\beta(R,\xi,P^\alpha,P^\beta) = \mathsf{E}_{(\nu\,|\,\xi)} W_t^\beta(R,\nu,\xi,P^\alpha,P^\beta).\end{cases} \tag{9}$$

This means that, for all R and ξ and for arbitrary policies P^α and P^β, we have

$$\begin{cases}\widetilde{V}^\alpha(R,\xi,\widehat{P}^\alpha,\widehat{P}^\beta) \ge \widetilde{V}^\alpha(R,\xi,P^\alpha,\widehat{P}^\beta),\\ \widetilde{V}^\beta(R,\xi,\widehat{P}^\alpha,\widehat{P}^\beta) \ge \widetilde{V}^\beta(R,\xi,\widehat{P}^\alpha,P^\beta).\end{cases}$$

Then

$$\begin{cases}\widetilde{U}^\alpha(R,\widehat{P}^\alpha,\widehat{P}^\beta) = \mathsf{E}_\xi \widetilde{V}^\alpha(R,\xi,\widehat{P}^\alpha,\widehat{P}^\beta) \ge \mathsf{E}_\xi \widetilde{V}^\alpha(R,\xi,P^\alpha,\widehat{P}^\beta) = \widetilde{U}^\alpha(R,P^\alpha,\widehat{P}^\beta),\\ \widetilde{U}^\beta(R,\widehat{P}^\alpha,\widehat{P}^\beta) = \mathsf{E}_\xi \widetilde{V}^\beta(R,\xi,\widehat{P}^\alpha,\widehat{P}^\beta) \ge \mathsf{E}_\xi \widetilde{V}^\beta(R,\xi,\widehat{P}^\alpha,P^\beta) = \widetilde{U}^\beta(R,\widehat{P}^\alpha,P^\beta).\end{cases}$$

Thus, the original Nash equilibrium problem for (8) can be reduced to a similar problem for functions (9), for which the Nash equilibrium policies can be computed pointwise.

On the other hand, we arrive at the same Nash equilibrium problem if we consider the game for *conditional* discounted payoffs

$$V_t^\alpha(R_t,\xi_t,\boldsymbol{P}_t^\alpha,\boldsymbol{P}_t^\beta) = \mathsf{E}_{(\nu_t,\nu_{t+1},\xi_{t+1},\ldots\,|\,\xi_t)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v_\tau^\alpha(R_\tau,\nu_\tau,p_\tau^\alpha,p_\tau^\beta), \tag{10}$$

and a similar expression for β. This will be done below.

5.2. Players Have Symmetric Information: Optimization of Conditional Payoff

In the previous subsection, we stated the optimum-decision problem as a problem of optimizing *average* discounted payoffs, where the averaging is taken over all random parameters ν_t and over all measurement results ξ_t.

However, it may seem to be more natural to optimize not *average* but *conditional* (with respect to measurement results) discounted payoffs according to expression (10).

By rewriting V_t^α in a recursive form, we get

$$\begin{aligned}V_t^\alpha(R_t,\xi_t,\boldsymbol{P}_t^\alpha,\boldsymbol{P}_t^\beta) &= \mathsf{E}_{(\nu_t\,|\,\xi_t)} \mathsf{E}_{(\nu_{t+1},\xi_{t+1},\ldots)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v_\tau^\alpha(R_\tau,\nu_\tau,p_\tau^\alpha,p_\tau^\beta)\\ &= \mathsf{E}_{(\nu_t\,|\,\xi_t)} \left[v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta) + \gamma_\alpha \mathsf{E}_{(\nu_{t+1},\xi_{t+1})} \sum_{\tau=t+1}^{T} \gamma_\alpha^{\tau-(t+1)} v_\tau^\alpha(R_\tau,\nu_\tau,p_\tau^\alpha,p_\tau^\beta) \right]\end{aligned}$$

$$= \mathsf{E}_{(\nu_t\,|\,\xi_t)}\Big[v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta) \\ + \gamma_\alpha \mathsf{E}_{(\xi_{t+1})} V_{t+1}^\alpha\big(\rho_t(R_t,\nu_t,p_t^\alpha,p_t^\beta),\xi_{t+1},\boldsymbol{P}_{t+1}^\alpha,\boldsymbol{P}_{t+1}^\beta\big)\Big].$$

Here, the mathematical expectation $\mathsf{E}_{(\nu_{t+1},\xi_{t+1})}$ over the joint distribution of the pair (ν_{t+1},ξ_{t+1}) can be represented as a sequence of expectation operations, i.e.,

$$\mathsf{E}_{(\nu_{t+1},\xi_{t+1})} = \mathsf{E}_{(\xi_{t+1})}\mathsf{E}_{(\nu_{t+1}\,|\,\xi_{t+1})}.$$

This leads to a dynamic programming algorithm, which involves computation of a Nash equilibrium pair $\langle p_t^\alpha, p_t^\beta\rangle$ for the following functions:

$$\begin{cases} \widetilde V_t^\alpha(R_t,\xi_t,p_t^\alpha,p_t^\beta) = \mathsf{E}_{(\nu_t\,|\,\xi_t)}\Big[v_t^\alpha(R_t,\nu_t,p_t^\alpha,p_t^\beta) + \gamma_\alpha \mathsf{E}_{(\xi_{t+1})}\widehat V_{t+1}^\alpha\big(\rho_t(R_t,\nu_t,p_t^\alpha,p_t^\beta),\xi_{t+1}\big)\Big], \\ \widetilde V_t^\beta(R_t,\xi_t,p_t^\alpha,p_t^\beta) = \mathsf{E}_{(\nu_t\,|\,\xi_t)}\Big[v_t^\beta(R_t,\nu_t,p_t^\alpha,p_t^\beta) + \gamma_\beta \mathsf{E}_{(\xi_{t+1})}\widehat V_{t+1}^\beta\big(\rho_t(R_t,\nu_t,p_t^\alpha,p_t^\beta),\xi_{t+1}\big)\Big]. \end{cases}$$

Both optimal strategies $\langle P_t^\alpha, P_t^\beta\rangle$ can now be found pointwise by setting $\widehat P_t^\alpha(R_t,\xi_t) = \widehat p_t^\alpha$ and $\widehat P_t^\beta(R_t,\xi_t) = \widehat p_t^\beta$, where $\langle \widehat p_t^\alpha, \widehat p_t^\beta\rangle$ attains the Nash equilibrium for the above functions.

More precisely, computation of the optimal policies $\langle \widehat P_t^\alpha, \widehat P_t^\beta\rangle$ at each step can be performed in two different ways:

(a) Fix R and ξ, find the Nash equilibrium point $\langle \widehat p^\alpha, \widehat p^\beta\rangle$ for the functions $\widetilde V_t^\alpha(R,\xi,p^\alpha,p^\beta)$ and $\widetilde V_t^\beta(R,\xi,p^\alpha,p^\beta)$, and set $\widehat P_t^\alpha(R,\xi) = \widehat p^\alpha$ and $\widehat P_t^\beta(R,\xi) = \widehat p^\beta$. Thus, the problem reduces to a pointwise computation of the Nash equilibrium for all possible values of R and ξ.

Note that the equilibrium pair $\langle \widehat p^\alpha, \widehat p^\beta\rangle$ can be found iteratively: for a fixed initial iteration $p_{(1)}^\alpha$, find the optimal response $p_{(1)}^\beta$, i.e.,

$$p_{(1)}^\beta = \arg\max_{p^\beta} \widetilde V_t^\beta(R,\xi,p_{(1)}^\alpha,p^\beta),$$

then find the optimal response $p_{(2)}^\alpha$ for $p_{(1)}^\beta$, etc. According to Proposition 8.1, the sequence $\langle p_{(i)}^\alpha, p_{(i)}^\beta\rangle$ (if converges) converges to the Nash equilibrium point $\langle \widehat p^\alpha, \widehat p^\beta\rangle$.

(b) Fix some policy $P_{(1)}^\alpha$ and find the optimal response $P_{(1)}^\beta$, i.e., a policy such that, for all R and ξ, the function $P_{(1)}^\beta$ maximizes the functional $\widetilde V_t^\beta(R,\xi,P_{(1)}^\alpha(R,\xi),P^\beta(R,\xi))$ with respect to P^β. Then similarly find $P_{(2)}^\alpha$, etc. It is obvious that, at each step, the optimal response can be found pointwise (separately for each combination R,ξ) for any (!) policy used by the other player.

5.3. Players Have Different Information

In a more general situation, players may obtain information based on *different* measurements $\xi_t^\alpha \in \mathcal{X}^\alpha$ and $\xi_t^\beta \in \mathcal{X}^\beta$ of the random parameter $\nu_t \in \mathcal{N}$. Then each player's policy depends on the respective information available to him, i.e., $p_t^\alpha = P_t^\alpha(R,\xi^\alpha)$ and $p_t^\beta = P_t^\beta(R,\xi^\beta)$, where P_t^α and P_t^β are measurable mappings:

$$P_t^\alpha\colon \mathcal{R}\times\mathcal{X}^\alpha \to \mathcal{D}^\alpha, \qquad P_t^\beta\colon \mathcal{R}\times\mathcal{X}^\beta \to \mathcal{D}^\beta.$$

Denote the spaces of measurable mappings from $\mathcal{R}\times\mathcal{X}^\alpha$ to $\mathcal{D}^\alpha$ and from $\mathcal{R}\times\mathcal{X}^\beta$ to $\mathcal{D}^\beta$ by $\widetilde{\mathcal{D}}^\alpha$ and $\widetilde{\mathcal{D}}^\beta$, respectively. Then the complete strategies $\boldsymbol{P}_t^\alpha$ and $\boldsymbol{P}_t^\beta$ are elements of the corresponding spaces

$$\widetilde{\mathcal{D}}_t^\alpha = \prod_{\tau=t}^T \widetilde{\mathcal{D}}^\alpha, \qquad \widetilde{\mathcal{D}}_t^\beta = \prod_{\tau=t}^T \widetilde{\mathcal{D}}^\beta.$$

Consider the game

$$G_{t,R} = \langle \widetilde{\mathcal{D}}^\alpha_t, \widetilde{\mathcal{D}}^\beta_t, U^\alpha_t, U^\beta_t \rangle_R, \quad R \in \mathcal{R}, \quad t = 1, \ldots, T,$$

in which each player maximizes his average discounted payoff:

$$U^\alpha_t(R_t, \boldsymbol{P}^\alpha_t, \boldsymbol{P}^\beta_t) = \mathsf{E}_{(\nu_t, \xi^\alpha_t, \xi^\beta_t, \ldots)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v^\alpha_\tau\Big(R_\tau, \nu_\tau, P^\alpha_\tau(R_\tau, \xi^\alpha_\tau), P^\beta_\tau(R_\tau, \xi^\beta_\tau)\Big),$$

and a similar expression for $U^\beta_t(R_t, \boldsymbol{P}^\alpha_t, \boldsymbol{P}^\beta_t)$.

Since U^α_t can be expressed through U^α_{t+1},

$$\begin{aligned} U^\alpha_t(R_t, \boldsymbol{P}^\alpha_t, \boldsymbol{P}^\beta_t) &= \mathsf{E}_{(\nu_t, \xi^\alpha_t, \xi^\beta_t, \ldots)} \Big[v^\alpha_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)) \\ &\qquad + \gamma_\alpha \sum_{\tau=t+1}^{T} \gamma_\alpha^{\tau-t} v^\alpha_\tau(R_\tau, \nu_\tau, P^\alpha_\tau(R_\tau, \xi^\alpha_\tau), P^\beta_\tau(R_\tau, \xi^\beta_\tau)) \Big] \\ &= \mathsf{E}_{(\nu_t, \xi^\alpha_t, \xi^\beta_t)} \Big[v^\alpha_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)) \\ &\qquad + \gamma_\alpha U^\alpha_{t+1}(\rho_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)), \xi^\alpha_{t+1}, \boldsymbol{P}^\alpha_{t+1}, \boldsymbol{P}^\beta_{t+1}) \Big], \end{aligned}$$

the problem of constructing optimal strategies $\langle \widehat{\boldsymbol{P}}^\alpha_t, \widehat{\boldsymbol{P}}^\beta_t \rangle$ can also be solved by a dynamic programming procedure. In this procedure, for every moment t, one constructs the optimal strategy pair $\langle \widehat{P}^\alpha_t, \widehat{P}^\beta_t \rangle$ for a certain (more simple) game provided that the solution $\langle \widehat{\boldsymbol{P}}^\alpha_{t+1}, \widehat{\boldsymbol{P}}^\beta_{t+1} \rangle$ for the game $G_{t+1,R}$ is already found.

Specifically, assume that

$$\widehat{U}^\alpha_t(R_t) = U^\alpha_t(R_t, \widehat{\boldsymbol{P}}^\alpha_t, \widehat{\boldsymbol{P}}^\beta_t), \qquad \widehat{U}^\beta_t(R_t) = U^\beta_t(R_t, \widehat{\boldsymbol{P}}^\alpha_t, \widehat{\boldsymbol{P}}^\beta_t), \qquad R \in \mathcal{R}, \quad t = 1, \ldots, T,$$

are the optimal payoffs. The functions

$$\begin{aligned} \widetilde{U}^\alpha_t(R_t, P^\alpha_t, P^\beta_t) &= \mathsf{E}_{(\nu_t, \xi^\alpha_t, \xi^\beta_t)} \Big[v^\alpha_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)) \\ &\qquad + \gamma_\alpha \widehat{U}^\alpha_{t+1}(\rho_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)), \xi^\alpha_{t+1}) \Big], \\ \widetilde{U}^\beta_t(R_t, P^\alpha_t, P^\beta_t) &= \mathsf{E}_{(\nu_t, \xi^\alpha_t, \xi^\beta_t)} \Big[v^\beta_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)) \\ &\qquad + \gamma_\beta \widehat{U}^\beta_{t+1}(\rho_t(R_t, \nu_t, P^\alpha_t(R_t, \xi^\alpha_t), P^\beta_t(R_t, \xi^\beta_t)), \xi^\beta_{t+1}) \Big] \end{aligned} \tag{11}$$

represent the players' payoffs provided that the players use arbitrary strategies P^α_t and P^β_t at time t and optimal strategies $\widehat{\boldsymbol{P}}^\alpha_{t+1}$ and $\widehat{\boldsymbol{P}}^\beta_{t+1}$ at all the following stages.

Let us now consider the game that appears at step t (we assume that the steps from $\tau = T$ down to $\tau = t+1$ are already performed) for any fixed $R \in \mathcal{R}$:

$$\widehat{G}_{t,R} = \langle \widehat{\mathcal{D}}^\alpha, \widehat{\mathcal{D}}^\beta, \widetilde{U}^\alpha_t, \widetilde{U}^\beta_t \rangle_R, \qquad R \in \mathcal{R}, \quad t = 1, \ldots, T, \tag{12}$$

where $\widehat{\mathcal{D}}^\alpha$ and $\widehat{\mathcal{D}}^\beta$ are the spaces of measurable mappings $\mathcal{X}^\alpha \to \mathcal{D}^\alpha$ and $\mathcal{X}^\beta \to \mathcal{D}^\beta$ respectively.

Proposition 5.1. *Assume that, for all t from $t = T$ down to $t = 1$ and for all $R \in \mathcal{R}$ and $\nu \in \mathcal{N}$, the game $\widehat{G}_{t,R}$ has a Nash equilibrium point $\langle \widehat{p}^\alpha_R, \widehat{p}^\beta_R \rangle$ (where $\widehat{p}^\alpha_R \in \widehat{\mathcal{D}}^\alpha$ and $\widehat{p}^\beta_R \in \widehat{\mathcal{D}}^\beta$) and the mappings*

$$\widehat{P}^\alpha_t(R, \xi^\alpha) = \widehat{p}^\alpha_R(\xi^\alpha), \qquad \widehat{P}^\beta_t(R, \xi^\beta) = \widehat{p}^\beta_R(\xi^\beta), \qquad \forall R \in \mathcal{R}, \quad \forall \xi^\alpha \in \mathcal{X}^\alpha, \quad \forall \xi^\beta \in \mathcal{X}^\beta, \tag{13}$$

are measurable with respect to R, ξ^α, and ξ^β. Then, for every t, there exists a pair of strategies $\langle \widehat{\boldsymbol{P}}_t^\alpha, \widehat{\boldsymbol{P}}_t^\beta \rangle$ that provides the Nash equilibrium for the family of games $\{G_{t,R,\nu} \mid R \in \mathcal{R}\}$.

The optimal strategies $\widehat{\boldsymbol{P}}_t^\alpha$ and $\widehat{\boldsymbol{P}}_t^\beta$ are determined recursively through the next-stage optimal strategies $\widehat{\boldsymbol{P}}_{t+1}^\alpha$ and $\widehat{\boldsymbol{P}}_{t+1}^\beta$ and through the equilibrium points $\langle \widehat{p}_R^\alpha, \widehat{p}_R^\beta \rangle$ for the games $\widehat{G}_{t,R}$ by relations (5) *and* (13).

The payoff functions $\widetilde{U}_t^\alpha$ and $\widetilde{U}_t^\beta$ for the game $\widehat{G}_{t,R}$ can be expressed through the payoff functions for the game $\widehat{G}_{t+1,R}$ by equation (11) *and the following expressions:*

$$\widehat{U}_t^\alpha(R) = \widetilde{U}_t^\alpha(R, \widehat{p}_R^\alpha, \widehat{p}_R^\beta), \qquad \widehat{U}_t^\beta(R) = \widetilde{U}_t^\beta(R, \widehat{p}_R^\alpha, \widehat{p}_R^\beta), \qquad R \in \mathcal{R}.$$

Proof. The scheme of the proof is similar to that of Proposition 4.1. △

Unlike the procedure in Proposition 4.1, where constructing optimal strategies reduces to finding equilibrium points for games of the simplest kind (where the spaces of strategies coincide with the spaces of elementary decisions $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$) and optimal strategies $\widehat{P}_t^\alpha$ and $\widehat{P}_t^\beta$ could be constructed pointwise (over R and ν), Proposition 5.1 leads to games $\widehat{G}_{t,R}$ (where the spaces of strategies $\widehat{\mathcal{D}}^\alpha$ and $\widehat{\mathcal{D}}^\beta$ are actually function spaces) and involves the construction of mappings $\widehat{p}_R^\alpha(\xi^\alpha)$ and $\widehat{p}_R^\beta(\xi^\beta)$ (pointwise over R only). However, construction of these mappings itself can be performed pointwise (over ξ^α and ξ^β) if we apply the iteration procedure of Proposition 8.3 (see the Appendix). Indeed, for a fixed R and t, game (12) is a game of the type (14).

Combination of Propositions 5.1 and 4.1 leads to the following procedure of constructing optimal strategies.

Consider the conditional (with respect to ξ_t^α) discounted payoff for player α:

$$V_t^\alpha(R_t, \xi_t^\alpha, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta) = \mathsf{E}_{(\nu_t, \xi_t^\beta, \ldots \,|\, \xi_t^\alpha)} \sum_{\tau=t}^{T} \gamma_\alpha^{\tau-t} v_\tau^\alpha(R_\tau, \nu_\tau, P_\tau^\alpha(R_\tau, \xi_\tau^\alpha), P_\tau^\beta(R_\tau, \xi_\tau^\beta))$$

and a similar expression for $V_t^\beta(R_t, \xi_t^\beta, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta)$.

Obviously, U_t^α is expressed through V_t^α in the following way:

$$U_t^\alpha(R_t, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta) = \mathsf{E}_{\xi_t^\alpha} V_t^\alpha(R_t, \xi_t^\alpha, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta),$$

where $\mathsf{E}_{\xi_t^\alpha}$ is the expectation over the distribution of the random element ξ_t^α.

Further, V_t^α is expressed through V_{t+1}^α:

$$\begin{aligned} V_t^\alpha(R_t, \xi_t^\alpha, \boldsymbol{P}_t^\alpha, \boldsymbol{P}_t^\beta) = \mathsf{E}_{(\nu_t, \xi_t^\beta \,|\, \xi_t^\alpha)} \Big[& v_t^\alpha(R_t, \nu_t, P_\tau^\alpha(R_\tau, \xi_\tau^\alpha), P_\tau^\beta(R_\tau, \xi_\tau^\beta)) \\ & + \gamma_\alpha \mathsf{E}_{\xi_{t+1}^\alpha} V_{t+1}^\alpha(\rho_t(R_t, \nu_t, P_\tau^\alpha(R_\tau, \xi_\tau^\alpha), P_\tau^\beta(R_\tau, \xi_\tau^\beta)), \xi_{t+1}^\alpha, \boldsymbol{P}_{t+1}^\alpha, \boldsymbol{P}_{t+1}^\beta) \Big]. \end{aligned}$$

Together with Propositions 5.1 and 4.1, this leads to the dynamic programming procedure in which, at every step, one constructs a pair of strategies $\langle \widehat{P}_t^\alpha, \widehat{P}_t^\beta \rangle$ that provide optimal responses for each other with respect to the functions

$$\begin{cases} \widetilde{V}_t^\alpha(R_t, \xi_t^\alpha, p_t^\alpha, P_t^\beta) = \mathsf{E}_{(\nu_t, \xi_t^\beta \,|\, \xi_t^\alpha)} \Big[v_t^\alpha(R_t, \nu_t, p_t^\alpha, P_t^\beta(R_t, \xi_t^\beta)) \\ \qquad\qquad + \gamma_\alpha \mathsf{E}_{\xi_{t+1}^\alpha} \widehat{V}_{t+1}^\alpha(\rho_t(R_t, \nu_t, p_t^\alpha, P_t^\beta(R_t, \xi_t^\beta)), \xi_{t+1}^\alpha) \Big], \\ \widetilde{V}_t^\beta(R_t, \xi_t^\beta, P_t^\alpha, p_t^\beta) = \mathsf{E}_{(\nu_t, \xi_t^\alpha \,|\, \xi_t^\beta)} \Big[v_t^\beta(R_t, \nu_t, P_t^\alpha(R_t, \xi_t^\alpha), p_t^\beta) \\ \qquad\qquad + \gamma_\beta \mathsf{E}_{\xi_{t+1}^\beta} \widehat{V}_{t+1}^\beta(\rho_t(R_t, \nu_t, P_t^\alpha(R_t, \xi_t^\alpha), p_t^\beta), \xi_{t+1}^\beta) \Big]. \end{cases}$$

In particular, the optimal pair $\langle \widehat{P}_t^\alpha, \widehat{P}_t^\beta \rangle$ satisfies the following conditions:

$$\begin{cases} \widetilde{V}_t^\alpha(R_t, \xi_t^\alpha, p_t^\alpha, \widehat{P}_t^\beta) \le \widetilde{V}_t^\alpha(R_t, \xi_t^\alpha, \widehat{P}_t^\alpha(R_t, \xi_t^\alpha), \widehat{P}_t^\beta), \\ \widetilde{V}_t^\beta(R_t, \xi_t^\beta, \widehat{P}_t^\alpha, p_t^\beta) \le \widetilde{V}_t^\beta(R_t, \xi_t^\beta, \widehat{P}_t^\beta, \widehat{P}_t^\alpha(R_t, \xi_t^\beta)). \end{cases}$$

Note that, in this case, one cannot use algorithm (a) from the previous subsection since now players (even if they know the strategies P_t^α and P_t^β for each other) cannot predict the real decisions p_t^α and p_t^β for each other (since these decisions are determined by the results of different measurements ξ_t^α and ξ_t^β available to the players). However, iteration algorithm (b) from the previous subsection can be used.

6. COOPERATIVE BEHAVIOR OF PLAYERS

It is easy to introduce some sort of cooperation (or contradiction) in our model by modifying the discounted payoffs V^α and V^β in a simple way, which reflects the "care" of one player about the other.

Specifically, player α "takes care" of the interests of β by optimizing the linear combination of payoffs

$$\boldsymbol{V}^\alpha = c_{\alpha\alpha} V^\alpha + c_{\alpha\beta} V^\beta$$

instead of his original payoff V^α. Similarly, β can optimize the linear combination

$$\boldsymbol{V}^\beta = c_{\beta\alpha} V^\alpha + c_{\beta\beta} V^\beta.$$

In fact, this describes (in the case $c_{\alpha\alpha} + c_{\beta\alpha} = 1$ and $c_{\alpha\beta} + c_{\beta\beta} = 1$) a game with side payments, where one player knows that he will get a certain fraction of another player's payoff.

If $c_{\alpha\beta} > 0$, player α tries to increase the income of β and, if $c_{\alpha\beta} > c_{\alpha\beta}$, then α cares about β more than about himself. Conversely, $c_{\alpha\beta} < 0$ means that α tries to "disserve" to β, possibly in order to exclude him from the business.

An interesting particular case is $c_{\alpha\alpha} = c_{\alpha\beta} = c_{\beta\alpha} = c_{\beta\beta} = \frac{1}{2}$. In fact, it represents a sole operator (monopolist) case.

Note that, if the discount factors are equal, i.e., $\gamma_\alpha = \gamma_\beta = \gamma$, then introducing cooperation coefficients in the problem statement scarcely influences the dynamic programming solution algorithm. Indeed, expressions like $v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta)$ are simply replaced by $c_{\alpha\alpha} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) + c_{\alpha\beta} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta)$. For example, in the "current information" case, we will have the following Nash equilibrium problem at each step:

$$\begin{cases} \widetilde{V}_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) = c_{\alpha\alpha} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) + c_{\alpha\beta} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) \\ \qquad\qquad + \gamma \mathsf{E}_{(\nu_{t+1} \mid \nu_t)} \widehat{V}_{t+1}^\alpha(\rho(R_t, p_t^\alpha, p_t^\beta, \nu_t), \nu_{t+1}), \\ \widetilde{V}_t^\beta(R_t, \nu_t, p_t^\alpha, p_t^\beta) = c_{\beta\alpha} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) + c_{\beta\beta} v_t^\alpha(R_t, \nu_t, p_t^\alpha, p_t^\beta) \\ \qquad\qquad + \gamma \mathsf{E}_{(\nu_{t+1} \mid \nu_t)} \widehat{V}_{t+1}^\beta(\rho(R_t, p_t^\alpha, p_t^\beta, \nu_t), \nu_{t+1}). \end{cases}$$

7. COMPUTER SIMULATION

7.1. Split-Stream Harvesting

In the split-stream harvesting model, we assume that each player (fishery fleet) harvests in his own stream and that the random split factor θ between streams may be unknown or imperfectly known to the players. The split-stream harvesting game is illustrated by the following diagram:

$$\begin{tikzcd}
& R^\alpha \arrow[r] & S^\alpha \arrow[dr] & & \\
R \arrow[ur] \arrow[dr] & & & S \arrow[r] & R_+ \\
& R^\beta \arrow[r] & S^\beta \arrow[ur] & &
\end{tikzcd}$$

Here, R is the current year's harvestable stock level, or "recruitment," and R^α and R^β are partial recruitments, in the separate streams, accessible to players α and β respectively. Thus,

$$R^\alpha = \theta^\alpha R, \qquad R^\beta = \theta^\beta R,$$

where

$$\theta^\alpha = \theta, \qquad\qquad \theta^\beta = 1 - \theta,$$

and θ is a random split factor. The residual substream stock, or "escapement," following the harvest is denoted by S^α or S^β respectively. These are determined by

$$S^\alpha = \sigma^\alpha(R^\alpha, p^\alpha), \qquad S^\beta = \sigma^\beta(R^\beta, p^\beta).$$

Here, p^α and p^β are the players' harvesting policies for this year (season), which determine what fraction of the available stock is harvested. We shall define policies as escapement fractions, so that

$$S^\alpha = p^\alpha R^\alpha, \qquad S^\beta = p^\beta R^\beta.$$

Finally, the substream escapements combine to form the current year's total escapement

$$S = S^\alpha + S^\beta,$$

which is the brood stock, for determining the following year's recruitment R_+ through the so-called "stock-recruitment relation":

$$R_+ = F(S).$$

Each player's strategy (harvesting policy) depends on the mutually known information structure of the game and on specific information that a player has when he makes his annual harvest decisions. We will always assume that both players know the current total recruitment R and also that each one has some information ξ^α and ξ^β about current and past random disturbances θ. Thus,

$$p^\alpha = P^\alpha(R, \xi^\alpha), \qquad p^\beta = P^\beta(R, \xi^\beta),$$

where P^α and P^β are the players' decision strategies.

The degree of players' knowledge about the random split factor θ may vary. In all the cases, we assume that both players know at least the stochastic properties of the random parameter θ. In addition, a player may have additional information: e.g., full current knowledge (this season's value), or only delayed knowledge (the previous season's value), or the result of imperfect observation of a current parameter value. Alternatively, he may possess no additional knowledge at all. Furthermore, the structure of the knowledge may be asymmetric; that is, the players may have different levels of knowledge.

In each season, a player gets a net return (annual payoff) v^α_{spl} or v^β_{spl}, which is a given function of his stream's recruitment and his own policy, i.e.,

$$v^\alpha_{\mathrm{spl}} = v^\alpha_{\mathrm{spl}}(R^\alpha, p^\alpha), \qquad v^\beta_{\mathrm{spl}} = v^\beta_{\mathrm{spl}}(R^\beta, p^\beta).$$

The player's payoff in the dynamic game is taken to be a discounted sum of his seasonal returns over the time span of the game.

It is easily seen that split-stream harvesting can be considered as a particular case of the dynamic game studied above. Indeed, recalling the notation of Section 2, we put $\nu = \theta$, the function ρ has the form

$$\rho(R, \theta, p^\alpha, p^\beta) = F\left((\theta p^\alpha + (1-\theta)p^\beta)R\right),$$

and the functions v^α and v^β are expressed through the corresponding functions $v^\alpha_{\rm spl}$ and $v^\beta_{\rm spl}$ as follows:

$$v^\alpha(R, \theta, p^\alpha, p^\beta) = v^\alpha_{\rm spl}(\theta R, p^\alpha), \qquad v^\beta(R, \theta, p^\alpha, p^\beta) = v^\beta_{\rm spl}((1-\theta)R, p^\beta).$$

In all simulation examples considered below, it is assumed that the random variables θ_t are independent and identically distributed.

7.2. Details of the Computer Model

In simulations, we use algorithms described above in Sections 4 and 5. Here, we specify a number of numerical and functional parameters, in particular, the immediate payoff function, described in Section 7.2.1; growth function, described in Section 7.2.2; and other numerical model parameters described in Section 7.2.3. Special attention in our simulations is paid to different variants of the game information structure, which is our primary goal (these variants are described in Section 7.2.4).

All the parameters were varied in simulations, with certain "steps" and in rather wide ranges. Thus, about a thousand various combinations were studied. Note that the effects described in Section 7.3 were observed more or less evidently in all of these combinations. Below, we demonstrate the results for one combination of parameters only, which is typical with respect to the effects observed. Besides, this combination is natural for the model described in Section 7.1 (i.e., the numerical values specified below are natural from the bioeconomic point of view).

As usual, in the situation of computer experiment, we cannot claim that in all the computations the algorithm converged to the sought-for equilibrium point. This is connected with the absence of a simple and convenient criterion for the algorithm convergence. We have processed about 30 thousands variants of games and, in 96% of the cases, algorithms converged in the following "practical" sense.

Let δ_t be an ordinary residual for an iteration procedure. If, for some iteration, this residual becomes lower than a fixed threshold δ_0 and remains to be so during d_0 following iterations, then it is assumed that the algorithm converges and computation is stopped. If the above condition is not fulfilled during D_0 iterations, then the algorithm is considered to be divergent.

Note that, even for one combination, every iteration involves computation of an optimal strategy as a function of two variables: a continuous one, R (we used a grid with 21 or more nodes, with linear interpolation between them), and a discrete one, θ or ξ (taking several—usually two—values). This required a considerable amount of computations.

A theoretical analysis of uniqueness of a pair of Nash equilibrium strategies and of the algorithm stability with respect to the initial pair of strategies requires a separate deep study.

In the simulations presented in this paper, the initial iteration for a pair of strategies (corresponding to the final moment T) was taken to reflect the absence of the harvest. For all the other moments t, a pair of optimal strategies obtained at the previous step (i.e., for the subsequent moment $t+1$) was used as an initial approximation.

At all steps of the algorithm, each player selected his strategy as an optimal reaction with respect to the strategy of the other player. Specifically, this reaction was constructed by global optimization on a certain grid.

In order to check the stability of the algorithm with respect to the initial strategy (and the uniqueness of the Nash equilibrium point), we used the following conventional approach. Initial pairs of strategies where selected at random in the space of all such pairs. In all the studied cases, the algorithm converged to one and the same Nash equilibrium point; i.e., the result did not depend on the pair of initial strategies.

7.2.1. Immediate payoff function. We define the players' decisions p^α and p^β as the corresponding escapement fractions (for the current season), i.e.,

$$S^\alpha = p^\alpha R^\alpha, \qquad S^\beta = p^\beta R^\beta.$$

Then the players' seasonal *harvests* are

$$H^\alpha = (1 - p^\alpha) R^\alpha, \qquad H^\beta = (1 - p^\beta) R^\beta.$$

In our simulations, the cost of harvesting is taken into account. We assume that the unit cost of harvesting for player α is inversely proportional to the current fish stock x in his stream and equals $\frac{c}{x}$, where c is a fixed coefficient. Thus, the seasonal price of harvesting for player α can be obtained as an integral from the player's escapement S^α up to his current recruitment R^α:

$$\text{cost} = \int\limits_S^R \frac{c}{x}\,dx = -c\log(p).$$

Then his total immediate payoff (for the current season) is

$$v(R,p) = H - \text{cost} = H^\alpha + c\log(p).$$

7.2.2. Growth function. In our studies, we used the growth function $F(S)$ that reflects the possibility of complete stock extinction (even without harvesting) if its amount falls below a certain critical level. This means that $F(S) < S$ for small enough S.

A wide enough class of growth functions $F(S)$ can be represented by 3rd-order polynomials that pass through the point $(0,0)$:

$$F(S) = a_1 S + a_2 S^2 + a_3 S^3.$$

Since a_1 is equal to the derivative of this function at zero, for $0 < a_1 < 1$ we have the possibility of stock extinction mentioned above. Below, we present simulation results for the natural case $a_1 = 0.6$. Besides, our function $F(S)$ attains its maximum at the point $(1,1)$. Thus, our growth function has the following form:

$$F(S) = 0.6S + 1.8S^2 - 1.6S^3.$$

Its graph is shown in Fig. 1.

7.2.3. Simulation parameters. All simulations presented in this paper were performed for the split-stream harvesting game from Section 7.1. The default parameter values were the following:

- Payoff function (Section 7.2.1) has $c_\alpha = c_\beta = c = 0.2$;
- Growth function (Section 7.2.2) has the form shown in Fig. 1;
- States of the random factor θ are $\theta_1 = 0.1$ and $\theta_2 = 0.9$. States at different moments are independent and identically distributed with equal probabilities $P(\theta = \theta_1) = P(\theta = \theta_2) = 0.5$;
- Discount factors are $\gamma_\alpha = \gamma_\beta = 0.9$ (for both players α and β);
- Cooperation weights from Section 6 are $c_{\alpha\alpha} = 1$, $c_{\alpha\beta} = 0$, $c_{\beta\alpha} = 0$, and $c_{\beta\beta} = 1$ (for pure competition) or $c_{\alpha\alpha} = 0.5$, $c_{\alpha\beta} = 0.5$, $c_{\beta\alpha} = 0.5$, and $c_{\beta\beta} = 0.5$ (for complete cooperation).

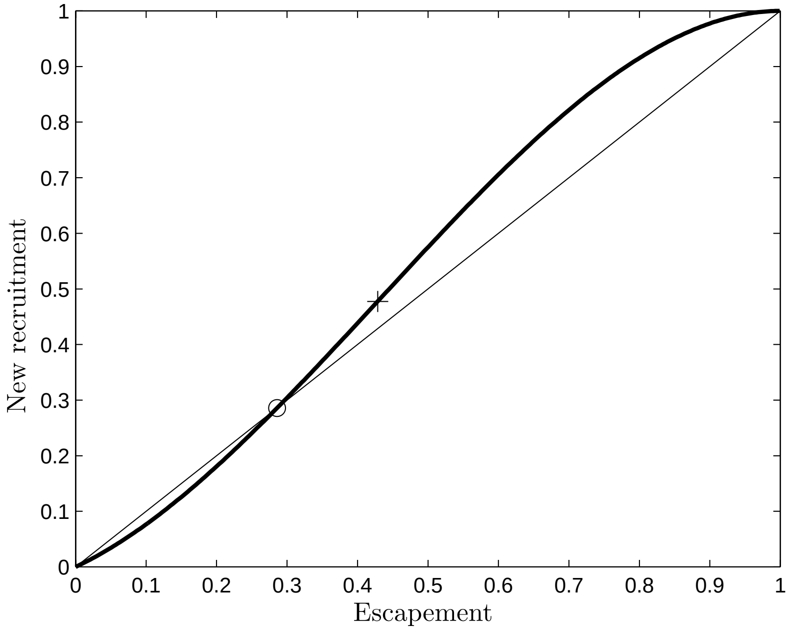

Fig. 1. Growth function. The circle shows the critical escapement.

In order to vary the "completeness" information about θ, we use imperfect observations (see Section 5). We change the "measurement precision" parameter π, which determines the measurement matrix

$$M = \begin{pmatrix} P(\xi = \xi_1 \,|\, \theta = \theta_1) & P(\xi = \xi_1 \,|\, \theta = \theta_2) \\ P(\xi = \xi_2 \,|\, \theta = \theta_1) & P(\xi = \xi_2 \,|\, \theta = \theta_2) \end{pmatrix},$$

(i.e., the matrix of conditional probabilities of the observable ξ for various values of θ) in the following way:

$$M = \begin{pmatrix} (1+\pi)/2 & (1-\pi)/2 \\ (1-\pi)/2 & (1+\pi)/2 \end{pmatrix}.$$

Thus, if $\pi = 1$ (maximum precision), then $M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which corresponds to "identical" measurement. If $\pi = 0$ (minimum precision), then $M = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$, which results in observations "independent" of the states of θ.

All simulations ran through a long enough averaging time period of 2000 time steps.

7.2.4. Major types of the game information structure. In our computer simulations, we studied the following types of the information structure of the game (in parentheses, we give the name and short notation for each of the cases).

1. At every moment t, players have complete symmetric current information about the random parameter θ ("current information," "Cur").
2. Players know the probability distribution of θ only ("minimal information," "Min").
3. Players get asymmetric (current vs. minimal) information: the first player gets current information, as in type 1, while the second one gets minimal information only, as in type 2 ("asymmetric information," "Cur–Min").

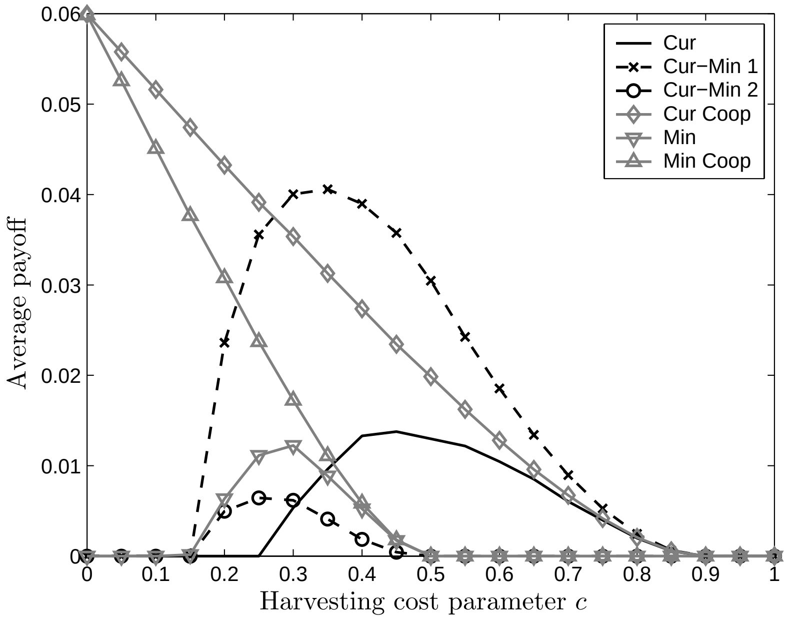

Fig. 2. Influence of the harvesting cost c for different types of information structures: "Cur"—players have complete current information (including all the previous moments), only one curve is shown since both payoffs are the same; "Cur–Min"—players have asymmetric information: the first player has current information, while the second one has minimal knowledge only; "Cur Coop"—cooperative harvesting with current information; "Min"—competitive harvesting with minimal symmetric information; "Min Coop"—cooperative harvesting with minimum information.

4. Players have incomplete information, namely, they observe a realization of the random variable ξ, which is a measurement of the random parameter θ ("incomplete information obtained from measurements," "Meas").
5. Players get incomplete asymmetric information: the first player according to type 4, while the second one—only minimal—according to type 2 ("incomplete asymmetric information," "Meas–Min").

We also provide simulation results of the cooperative behavior for symmetric information structures 1, 2, and 4 (since, under cooperative management, it is natural to assume that players completely interchange all available information).

7.3. Results of Computer Simulations

7.3.1. Influence of the harvesting cost. Here we examine the influence of the harvesting cost parameter c on the outcome of the split-stream harvesting game (see Section 7.2.1).

The graphs displayed in Fig. 2 show *average* (over all seasons) *payoffs* as functions of the harvesting cost parameter c for the following five types of the game: cooperative harvesting with current information, cooperative harvesting with minimum information, competitive harvesting with current information, competitive harvesting with minimal symmetric information, and finally competitive harvesting with asymmetric (current vs. minimal) information.

The most interesting in the cases where the players compete is the reduction of the mean payoff when the harvesting cost c is low. It is explained by the fact that, at low cost of the harvest, competition becomes more aggressive, and this leads to a severe reduction of the stock. This effect

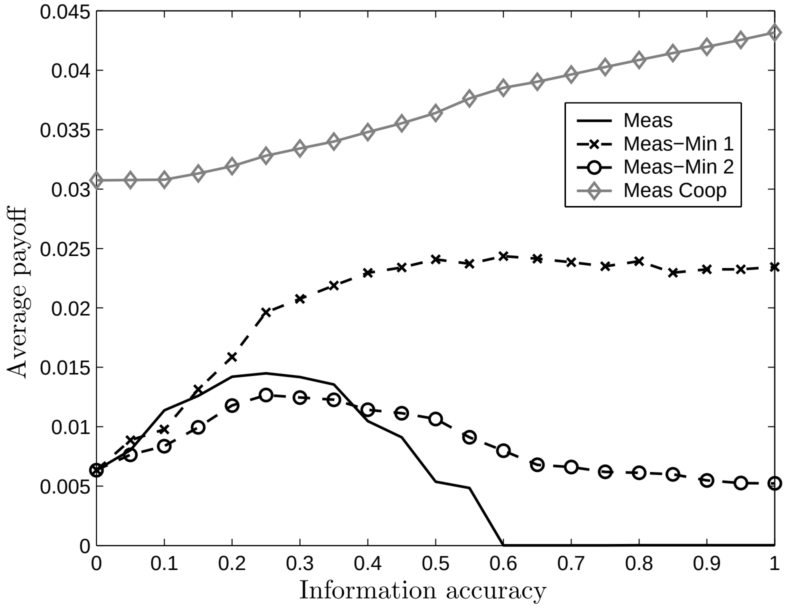

Fig. 3. Influence of the information accuracy. "Meas"—both players have equal incomplete information; "Meas–Min"—the first player ("Meas–Min 1") obtains measurement information, while the second one ("Meas–Min 2") has only minimal knowledge; "Meas Coop"—cooperative harvesting with equal incomplete measurement information.

is not seen when the players cooperate and, hence, are able to keep the stock at a high enough level.

Besides, it is seen in Fig. 2 that an informational advantage in the asymmetric knowledge case is highly beneficial for the first player ("Cur–Min 1"). Moreover, he would not wish to share his additional knowledge with his competitor and thereby switch to the symmetric complete knowledge case ("Cur"). Typically, the second player ("Cur–Min 2") in the asymmetric case would prefer to get the current knowledge, but not for the low cost (< 0.3), where the asymmetric lack of knowledge is more beneficial (even for him) than the symmetric complete current knowledge ("Cur"). Furthermore, at low enough costs (< 0.35), minimal information ("Min") is more beneficial than the complete current knowledge ("Cur"). Apparently, this is due to the fact that, in the absence of a precise knowledge, the risk of an accidental stock destruction becomes higher and, as a consequence, the players behave more "carefully."

As is noted above, competition becomes especially destructive for the low cost of harvest c, when overharvest can completely destroy the stock. However, if the players cooperate, their return is significantly higher, especially when the cost of harvest is low. Cooperatively, in contrast to the competitive case, they are able to hold expected escapements at relatively high levels.

7.3.2. Information from imperfect observations. In this set of simulations (Fig. 3), information about the current θ is obtained from imperfect measurements of the random parameter θ. The measurement accuracy is a variable parameter, increasing from 0 (no information) to 1 (complete information).

Note that, when the players cooperate ("Meas Coop"), the situation is quite natural: the greater their (shared) information, the greater their payoff. The sum of the payoffs here will necessarily be superior to those in competitive situations, with or without informational asymmetry.

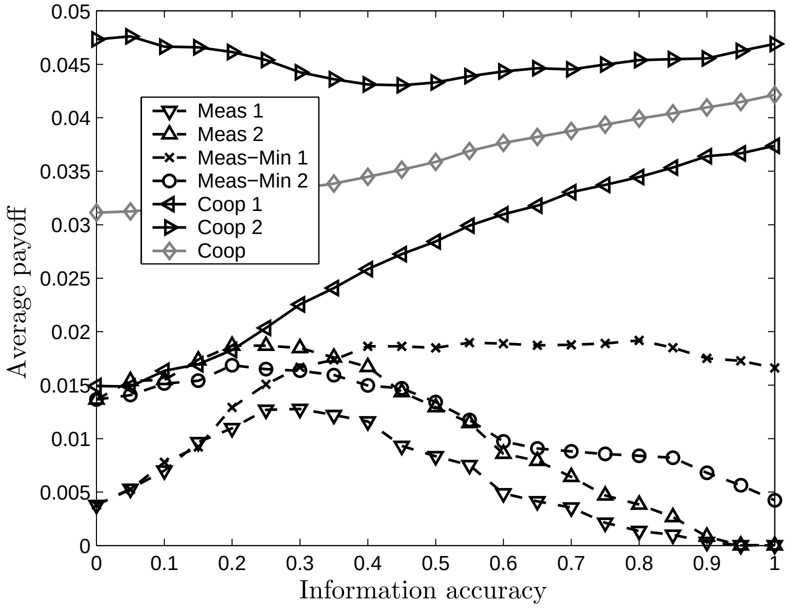

Fig. 4. Influence of the information accuracy under asymmetric environmental conditions. "Meas"—both players have equal incomplete information; "Meas–Min"—the first player obtains measurement information, while the second one has only minimal knowledge; "Coop"—cooperative harvesting with equal incomplete measurement information (in addition to the players' actual payoffs "Coop 1" and "Coop 2," we also show their equal sharing payoff "Coop").

It is seen from Fig. 3 that, in competitive games with symmetric ("Meas") and asymmetric ("Meas–Min") information structures, additional information below a certain level is beneficial to both players, even for the player who does not possess additional information ("Meas–Min 2"). However, further increasing the knowledge level degrades the situation dramatically, presumably by making harvesting policies more aggressive.

In addition, for a low measurement accuracy, the situation where the second player also accesses the measurement information is better for him than the situation where he does not ("Meas" versus "Meas–Min 2"). However, for high enough accuracy levels (> 0.38), he does not benefit from his additional knowledge.

At the same time, cooperative management ("Meas Coop") provides a much higher return, which is constantly growing with the increase of the measurement accuracy.

7.3.3. Balancing asymmetric information against asymmetric environmental conditions. Here, the nature slightly favors the second player (Fig. 4). Specifically, θ takes the values 0.1 and 0.8 with equal probabilities (so, the first player's fraction of the total recruitment is either 0.1 or 0.8 of the whole recruitment, while the second player receives the fraction either 0.9 or 0.2). Thus, the mean recruitment for the first player is lower than for the second one.

As one would expect, when the players possess identical information, player 2 ("Meas 2") will always do better than player 1 ("Meas 1"). But when player 1 ("Meas–Min 1") has a strong informational advantage (measurement accuracy > 0.3), this can overbalance the second player's ("Meas–Min 2") environmental advantages.

On the other hand, the sum of the players' payoffs will be the greatest with cooperation: the players share information, with the common objective of maximizing the sum of their returns. In this case, because of the environmental asymmetry, the direct harvest returns in the two sub-

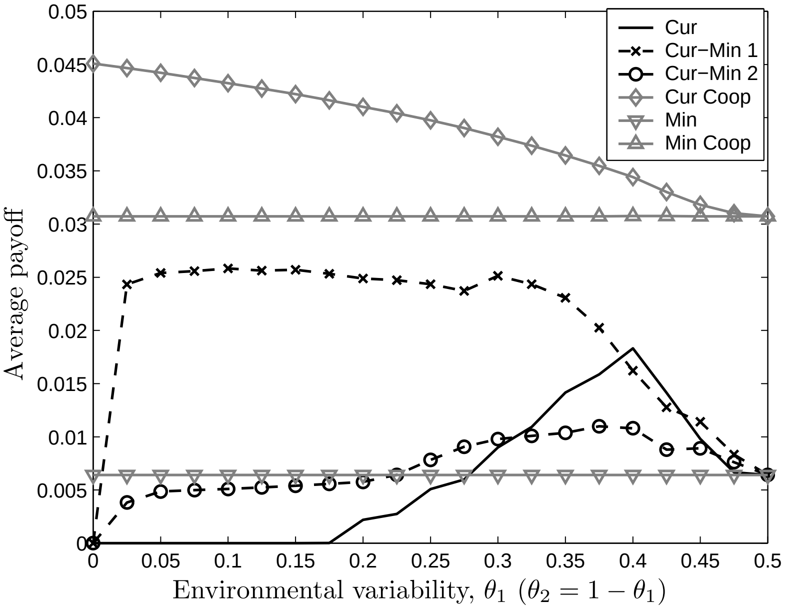

Fig. 5. Influence of the environmental variability (characterized by θ_1). "Cur"—players have complete current information; "Cur–Min"—players have asymmetric information: the first player has current information, while the second one has minimal knowledge only; "Cur Coop"—cooperative harvesting with current information; "Min"—minimal information; "Min Coop"—cooperative harvesting with minimum information.

streams ("Coop 1" and "Coop 1") are not equal. This can be considered as a pure "good-will" cooperation. Alternatively, the two players agree upon a redistribution of this total return, which leads to a compensating "side payment" from one player to the other. The case of equal sharing ("Coop") is shown in the figure as well.

7.3.4. Influence of the environmental variability. In this set of simulations (Fig. 5), we change the variance of the stock-split factor θ. Specifically, θ randomly takes two values: θ_1 and $\theta_2 = 1 - \theta_1$, where θ_1 may be any fraction between 0 and 0.5. For $\theta_1 = 0.5$, there is no variability ($\theta = 0.5$ always), while for $\theta_1 = 0$ the variability is the highest (θ jumps randomly between 0 and 1). For cooperative harvesting with complete knowledge ("Cur Coop"), the increase of the payoff, as well as an increase of the variability of θ (decrease of θ_1) is quite natural. Indeed, with high variability of θ, almost the entire fish stock goes into one of two streams, and this leads to a reduction of harvesting cost per unit (cf. Section 7.2.1). Since this cooperative game is fully symmetric, the average annual payoffs to the two players will be identical.

It seems that the increase of the payoff in competitive games with a decrease of θ_1 from 0.5 to 0.4 may have the same explanation. However, at a higher variability (low θ_1), the effect of competition (especially for complete knowledge, "Cur") becomes dominant. Indeed, if all the stock is in one stream, the corresponding fleet can harvest almost all the stock at a relatively low cost.

7.3.5. Transition from competition to cooperation. In Section 6, we described a simple way to introduce "cooperation" into the competitive game. Figure 6 shows the dependence of average payoffs on the "degree of cooperation" for the cases of complete and minimal information.

Here, the zero degree of cooperation d means the purely competitive behavior, with each player maximizing his own discounted payoff, while degree-1 cooperation means that both players maximize the total discounted payoff. For intermediate values of cooperation, each player maximizes a convex linear combination of his own and his competitor's discounted payoffs.

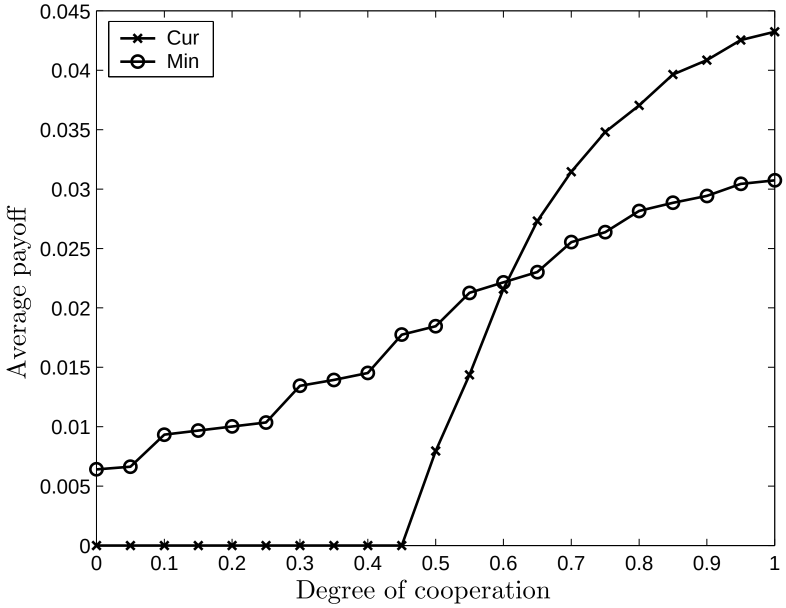

Fig. 6. Increasing degree of cooperation for current and minimum knowledge. “Cur”—current information, “Min”—minimal information.

To be precise, cooperation coefficients are expressed through the cooperation degree, denoted by d, in the following way: $c_{\alpha\alpha} = 1 - d/2$, $c_{\alpha\beta} = d/2$, $c_{\beta\alpha} = d/2$, and $c_{\beta\beta} = 1 - d/2$. Thus, when d goes from 0 to 1, the situation changes from “no cooperation” to “complete cooperation.”

It is clearly seen from Fig. 6 that additional information (“Cur”) is beneficial only if there is a high degree of cooperation. At low degrees of cooperation and especially in the case of pure competition, additional knowledge leads to a critical drop of escapement and to zero average payoff.

8. APPENDIX: GENERAL INFORMATION ABOUT GAMES

Recall that a nonantagonistic two-player game is determined by a tuple $g = \langle \mathcal{D}^\alpha, \mathcal{D}^\beta, v^\alpha, v^\beta \rangle$, where $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ are the spaces of actions (decisions) of players α and β, while v^α and v^β are their payoff functions respectively. More precisely, $v^\alpha, v^\beta \colon \mathcal{D}^\alpha \times \mathcal{D}^\beta \to \mathbb{R}$, $v^\alpha(p^\alpha, p^\beta)$ and $v^\beta(p^\alpha, p^\beta)$ are payoffs for players α and β if their decisions are p^α and p^β respectively.

The classical problem of constructing optimal decisions in such a game consists in finding a Nash equilibrium point, i.e., a pair of (optimal) decisions $\langle \widehat{p}^\alpha, \widehat{p}^\beta \rangle$ such that any one-sided deviation of a player from his optimal decision leads to a reduction of his payoff, i.e.,

$$\begin{aligned} \forall p^\alpha \in \mathcal{D}^\alpha, &\quad v^\alpha(p^\alpha, \widehat{p}^\beta) \le v^\alpha(\widehat{p}^\alpha, \widehat{p}^\beta), \\ \forall p^\beta \in \mathcal{D}^\beta, &\quad v^\alpha(\widehat{p}^\alpha, p^\beta) \le v^\beta(\widehat{p}^\alpha, \widehat{p}^\beta). \end{aligned}$$

By $\pi^\alpha \colon \mathcal{D}^\beta \to \mathcal{P}(\mathcal{D}^\alpha)$, denote the multivalued mapping[1] which determines the set of optimal responses of player α with respect to the decision p^β of player β:

$$\pi^\alpha(p^\beta) = \arg\max_{p^\alpha}(p^\alpha, p^\beta).$$

Similarly define $\pi^\beta \colon \mathcal{D}^\alpha \to \mathcal{P}(\mathcal{D}^\beta)$:

$$\pi^\beta(p^\alpha) = \arg\max_{p^\beta}(p^\alpha, p^\beta).$$

[1] Here and in what follows, $\mathcal{P}(\mathcal{X})$ denotes the set of all subsets of a set $\mathcal{X}$.

Obviously, a pair $\langle \widehat{p}^\alpha, \widehat{p}^\beta \rangle$ is a Nash equilibrium point if

$$\widehat{p}^\alpha \in \pi^\alpha(\widehat{p}^\beta), \qquad \widehat{p}^\beta \in \pi^\beta(\widehat{p}^\alpha).$$

In practice, a Nash equilibrium point $\langle \widehat{p}^\alpha, \widehat{p}^\beta \rangle$ can be constructed by the following iteration procedure.

Let p_0^α be an arbitrary decision of player α (initial approximation). Assume that p_0^β is an optimal response of player β with respect to p_0^α ($p_0^\beta \in \pi^\beta(p_0^\alpha)$). Similarly, let p_1^α be an optimal response of α with respect to p_0^β ($p_1^\alpha \in \pi^\alpha(p_0^\beta)$), etc.:

$$p_n^\alpha \in \pi^\alpha(p_{n-1}^\beta), \qquad p_n^\beta \in \pi^\beta(p_n^\alpha).$$

Proposition 8.1. *Assume that $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ are metric spaces and π^α and π^β are everywhere defined single-valued continuous mappings. If there exists any of the limits*

$$\lim_{n\to\infty} p_n^\alpha = \lim_{n\to\infty} (\pi^\alpha \circ \pi^\beta)^n (p_0^\alpha) = \widehat{p}^\alpha$$

or

$$\lim_{n\to\infty} p_n^\beta = \lim_{n\to\infty} \left((\pi^\beta \circ \pi^\alpha)^n \circ \pi^\beta \right) (p_0^\alpha) = \widehat{p}^\beta,$$

then the other also exists, and the pair $\langle \widehat{p}^\alpha, \widehat{p}^\beta \rangle$ is a Nash equilibrium point for the game $g = \langle \mathcal{D}^\alpha, \mathcal{D}^\beta, v^\alpha, v^\beta \rangle$.

Proof. Assume that there exists $\lim\limits_{n\to\infty} p_n^\alpha = \widehat{p}^\alpha$. Let us prove that the second limit $\lim\limits_{n\to\infty} p_n^\beta$ also exists and is equal to $\widehat{p}^\beta$. Indeed,

$$\widehat{p}^\beta = \lim_{n\to\infty} p_n^\beta = \lim_{n\to\infty} \pi^\beta(p_n^\alpha) = \pi^\beta\left(\lim_{n\to\infty} p_n^\alpha\right) = \pi^\beta(\widehat{p}^\alpha).$$

Similarly, if there exists $\lim\limits_{n\to\infty} p_n^\beta = \widehat{p}^\beta$, then $\lim\limits_{n\to\infty} p_n^\alpha$ also exists and

$$\widehat{p}^\alpha = \lim_{n\to\infty} p_n^\alpha = \pi^\alpha(\widehat{p}^\beta).$$

Thus, the pair $\langle \widehat{p}^\alpha, \widehat{p}^\beta \rangle$ satisfies the conditions $\widehat{p}^\alpha \in \pi^\alpha(\widehat{p}^\beta)$ and $\widehat{p}^\beta \in \pi^\beta(\widehat{p}^\alpha)$, i.e., is a Nash equilibrium point. $\triangle$

Now assume that the result of a game depends not only on players' decisions but also on a random parameter ν (which has a known probability distribution) from a space $\mathcal{N}$. Specifically, assume that $v^\alpha, v^\beta \colon \mathcal{D}^\alpha \times \mathcal{D}^\beta \times \mathcal{N} \to \mathbb{R}$. Since the players do not know the parameter ν, it is natural to formulate the optimal decision problem as the problem of maximizing the *average* payoffs

$$V^\alpha(p^\alpha, p^\beta) = \mathsf{E}_\nu v^\alpha(p^\alpha, p^\beta, \nu)$$

and

$$V^\beta(p^\alpha, p^\beta) = \mathsf{E}_\nu v^\beta(p^\alpha, p^\beta, \nu).$$

Thus, we obtain a new game $\langle \mathcal{D}^\alpha, \mathcal{D}^\beta, V^\alpha, V^\beta \rangle$, which differs from the initial one $\langle \mathcal{D}^\alpha, \mathcal{D}^\beta, v^\alpha, v^\beta \rangle$ by payoff functions only (V^α and V^β instead of v^α and v^β).

The situation becomes more complicated if the players obtain some information about the random parameter ν. Specifically, assume that players α and β obtain results of imperfect observations of ν, i.e., realizations of certain random elements $\xi^\alpha \in \mathcal{X}^\alpha$ and $\xi^\beta \in \mathcal{X}^\beta$. Interrelation between ν, ξ^α, and ξ^β is described by a given joint distribution on the space $\mathcal{N} \times \mathcal{X}^\alpha \times \mathcal{X}^\beta$.

Assume that player α gets an observation of the random element ξ^α. Then his decision p^α may depend on ξ^α, i.e., is defined as $p^\alpha = P^\alpha(\xi^\alpha)$, where P^α is some decision strategy of player α—a measurable mapping from $\mathcal{X}^\alpha$ to $\mathcal{D}^\alpha$. Similarly, player β may make his decisions based on the observation ξ^β, i.e., $p^\beta = P^\beta(\xi^\beta)$.

If payers use strategies P^α and P^β, their average payoffs are defined as

$$V^\alpha(P^\alpha, P^\beta) = \mathsf{E}_{(\nu,\xi^\alpha,\xi^\beta)} v^\alpha\left(P^\alpha(\xi^\alpha), P^\beta(\xi^\beta), \nu\right),$$
$$V^\beta(P^\alpha, P^\beta) = \mathsf{E}_{(\nu,\xi^\alpha,\xi^\beta)} v^\beta\left(P^\alpha(\xi^\alpha), P^\beta(\xi^\beta), \nu\right).$$

Thus, in the situation with imperfect observations of the random parameter, we arrive at a new game

$$G = \langle \widetilde{\mathcal{D}}^\alpha, \widetilde{\mathcal{D}}^\beta, V^\alpha, V^\beta \rangle.$$

Here, $\widetilde{\mathcal{D}}^\alpha$ and $\widetilde{\mathcal{D}}^\beta$ are function spaces, specifically, spaces of measurable mappings of the form $\mathcal{X}^\alpha \to \mathcal{D}^\alpha$ and $\mathcal{X}^\beta \to \mathcal{D}^\beta$, respectively, and V^α and V^β are functionals defined on the space $\widetilde{\mathcal{D}}^\alpha \times \widetilde{\mathcal{D}}^\beta$.

Let us note that the problem of constructing a Nash equilibrium point for this game is considerably more complicated than that for the initial game g. Indeed, in the initial game, strategies are typically elements of finite-dimensional linear spaces (or even finite sets) while, in the corresponding game with imperfect observations, strategies are elements of function spaces.

However, in some cases (for example, when both players obtain equal information), it is possible to reduce constructing optimal strategies for the game $G = \langle \widetilde{\mathcal{D}}^\alpha, \widetilde{\mathcal{D}}^\beta, V^\alpha, V^\beta \rangle$ to those for games of the type $\langle \mathcal{D}^\alpha, \mathcal{D}^\beta, v^\alpha, v^\beta \rangle$.

So, assume that both players obtain results of one and the same observation, i.e., $\xi^\alpha = \xi^\beta = \xi \in \mathcal{X} = \mathcal{X}^\alpha = \mathcal{X}^\beta$.

For a fixed $\xi \in \mathcal{X}$, consider the conditional payoff functions

$$V^\alpha_\xi(p^\alpha, p^\beta) = \mathsf{E}_{(\nu\,|\,\xi)} v^\alpha(p^\alpha, p^\beta, \nu),$$
$$V^\beta_\xi(p^\alpha, p^\beta) = \mathsf{E}_{(\nu\,|\,\xi)} v^\beta(p^\alpha, p^\beta, \nu),$$

where $\mathsf{E}_{(\nu\,|\,\xi)}$ is the operation of conditional (with respect to ξ) mathematical expectation over ν.

Now, for a fixed $\xi \in \mathcal{X}$, consider the "conditional" game

$$G_\xi = \langle \mathcal{D}^\alpha, \mathcal{D}^\beta, V^\alpha_\xi, V^\beta_\xi \rangle,$$

which differs from the game G by replacing the distribution of the random element ν with the conditional distribution of ν with respect to a fixed ξ.

Proposition 8.2 (Bayes principle). *Assume that, for any $\xi \in \mathcal{X}$, there exists a Nash equilibrium point $\langle \widehat{p}^\alpha_\xi, \widehat{p}^\beta_\xi \rangle$ for the game $G_\xi = \langle \mathcal{D}^\alpha, \mathcal{D}^\beta, V^\alpha_\xi, V^\beta_\xi \rangle$, and the mappings $\widehat{P}^\alpha$ and $\widehat{P}^\beta$ defined by the equations*

$$\widehat{P}^\alpha(\xi) = \widehat{p}^\alpha_\xi, \qquad \widehat{P}^\beta(\xi) = \widehat{p}^\beta_\xi$$

are measurable. Then the pair $\langle \widehat{P}^\alpha_\xi, \widehat{P}^\beta_\xi \rangle$ is a Nash equilibrium point for the game

$$G = \langle \widetilde{\mathcal{D}}^\alpha, \widetilde{\mathcal{D}}^\beta, V^\alpha, V^\beta \rangle.$$

Proof. Let $P^\alpha \colon \mathcal{X} \to \mathcal{D}^\alpha$ and $P^\beta \colon \mathcal{X} \to \mathcal{D}^\beta$ be arbitrary measurable mappings. Since the mathematical expectation operation $\mathsf{E}_{(\nu,\xi)}$ can be written as a composition $\mathsf{E}_{(\nu,\xi)} = \mathsf{E}_\xi \mathsf{E}_{(\nu\,|\,\xi)}$ of expectations over the marginal distribution of ξ and the conditional distribution of ν with respect to ξ, we have

$$V^\alpha(P^\alpha, P^\beta) = \mathsf{E}_\xi \mathsf{E}_{(\nu\,|\,\xi)} v^\alpha(P^\alpha(\xi), P^\beta(\xi), \nu),$$
$$V^\beta(P^\alpha, P^\beta) = \mathsf{E}_\xi \mathsf{E}_{(\nu\,|\,\xi)} v^\beta(P^\alpha(\xi), P^\beta(\xi), \nu).$$

Thus,

$$V^\alpha(P^\alpha, P^\beta) = \mathsf{E}_\xi V^\alpha_\xi(P^\alpha(\xi), P^\beta(\xi)),$$
$$V^\beta(P^\alpha, P^\beta) = \mathsf{E}_\xi V^\beta_\xi(P^\alpha(\xi), P^\beta(\xi)).$$

Now, let us consider $V^\alpha(P^\alpha, \widehat{P}^\beta)$. According to the definition, $V^\alpha_\xi(P^\alpha(\xi), \widehat{p}^\beta_\xi) \le V^\alpha_\xi(\widehat{p}^\alpha_\xi, \widehat{p}^\beta_\xi)$ for all $\xi \in \mathcal{X}$. This yields

$$V^\alpha(P^\alpha, \widehat{P}^\beta) = \mathsf{E}_\xi V^\alpha_\xi(P^\alpha(\xi), \widehat{P}^\beta(\xi)) \le \mathsf{E}_\xi V^\alpha_\xi(\widehat{P}^\alpha(\xi), \widehat{P}^\beta(\xi)) = V^\alpha(\widehat{P}^\alpha, \widehat{P}^\beta).$$

This implies $V^\alpha(P^\alpha, \widehat{P}^\beta) \le V^\alpha(\widehat{P}^\alpha, \widehat{P}^\beta)$. By the same argument, we prove that $V^\beta(\widehat{P}^\alpha, P^\beta) \le V^\beta(\widehat{P}^\alpha, \widehat{P}^\beta)$, i.e., $\langle \widehat{P}^\alpha, \widehat{P}^\beta \rangle$ is a Nash equilibrium point for the game G. △

Thus, in the game G with equal information, a pair of optimal strategies $\langle \widehat{P}^\alpha, \widehat{P}^\beta \rangle$ can be constructed "pointwise" by considering games G_ξ for all the possible values $\xi \in \mathcal{X}$. Clearly, optimal decisions for a conditional game can be constructed according to the procedure for the initial game g in Proposition 8.1.

Finally, for a game G in the general case, optimal strategies can also be constructed "pointwise," but with the use of a more complex iteration procedure.

So, consider the general game

$$G = \langle \widetilde{\mathcal{D}}^\alpha, \widetilde{\mathcal{D}}^\beta, V^\alpha, V^\beta \rangle \tag{14}$$

for the case of different observations $\xi^\alpha \in \mathcal{X}^\alpha$ and $\xi^\beta \in \mathcal{X}^\beta$ for players α and β respectively.

For an arbitrary fixed value $\xi^\alpha \in \mathcal{X}^\alpha$, consider the conditional payoff for player α:

$$V^\alpha_{\xi^\alpha}(p^\alpha, P^\beta) = \mathsf{E}_{(\nu, \xi^\beta\,|\,\xi^\alpha)} v^\alpha(p^\alpha, P^\beta(\xi^\beta), \nu).$$

Similarly define $V^\beta_{\xi^\beta}$:

$$V^\beta_{\xi^\beta}(P^\alpha, p^\beta) = \mathsf{E}_{(\nu, \xi^\alpha\,|\,\xi^\beta)} v^\beta(P^\alpha(\xi^\alpha), p^\beta, \nu).$$

Assume that $\pi^\alpha_{\xi^\alpha} \colon \widetilde{\mathcal{D}}^\beta \to \mathcal{P}(\mathcal{D}^\alpha)$ is a multivalued mapping that represents the set of all optimal reactions p^α of player α with respect to a given strategy P^β of player β provided that player α has observation ξ^α. Specifically, put

$$\pi^\alpha_{\xi^\alpha}(P^\beta) = \arg\max_{p^\alpha} V^\alpha_{\xi^\alpha}(p^\alpha, P^\beta).$$

In the same way, define $\pi^\beta_{\xi^\beta} \colon \widetilde{\mathcal{D}}^\alpha \to \mathcal{P}(\mathcal{D}^\beta)$ for all $\xi^\beta \in \mathcal{X}^\beta$.

Now let us consider the following iteration procedure. Take an arbitrary initial strategy P^α_0 (i.e., a measurable mapping $P^\alpha_0 \colon \mathcal{X}^\alpha \to \mathcal{D}^\alpha$) for player α and construct an optimal reaction for player β. To do this, for every $\xi^\beta \in \mathcal{X}^\beta$ put $P^\beta_0(\xi^\beta) \in \pi^\beta_{\xi^\beta}(P^\alpha_0)$. In the same way, define $P^\alpha_1 \colon \mathcal{X}^\alpha \to \mathcal{D}^\alpha$ such that $P^\alpha_1(\xi^\alpha) \in \pi^\alpha_{\xi^\alpha}(P^\beta_0)$ for all $\xi^\alpha \in \mathcal{X}^\alpha$, etc.:

$$P^\alpha_n(\xi^\alpha) \in \pi^\alpha_{\xi^\alpha}(P^\beta_{n-1}), \qquad \forall \xi^\alpha \in \mathcal{X}^\alpha,$$
$$P^\beta_n(\xi^\beta) \in \pi^\beta_{\xi^\beta}(P^\alpha_n), \qquad \forall \xi^\beta \in \mathcal{X}^\beta.$$

Proposition 8.3. *Let $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ be metric spaces. Assume that mappings $\pi^\alpha_{\xi^\alpha}$ and $\pi^\beta_{\xi^\beta}$,*

$$\pi^\alpha_{\xi^\alpha}\colon \widetilde{\mathcal{D}}^\beta \to \mathcal{P}(\mathcal{D}^\alpha), \quad \pi^\alpha_{\xi^\alpha}(P^\beta) = \arg\max_{p^\alpha} V^\alpha_{\xi^\alpha}(p^\alpha, P^\beta), \quad \forall \xi^\alpha \in \mathcal{X}^\alpha,$$
$$\pi^\alpha_{\xi^\beta}\colon \widetilde{\mathcal{D}}^\alpha \to \mathcal{P}(\mathcal{D}^\beta), \quad \pi^\alpha_{\xi^\beta}(P^\alpha) = \arg\max_{p^\beta} V^\beta_{\xi^\beta}(P^\alpha, p^\beta), \quad \forall \xi^\beta \in \mathcal{X}^\beta,$$

are everywhere defined, single-valued, equicontinuous with respect to $\xi^\alpha \in \mathcal{X}^\alpha$ and $\xi^\beta \in \mathcal{X}^\beta$, respectively, and are such that, for all fixed measurable mappings $P^\alpha \in \widetilde{\mathcal{D}}^\alpha$ and $P^\beta \in \widetilde{\mathcal{D}}^\beta$, the mappings $\pi^\alpha_{\xi^\alpha}(P^\beta)$ and $\pi^\beta_{\xi^\beta}(P^\alpha)$ are measurable with respect to ξ^α and ξ^β respectively. If there exists one of the limits

$$\lim_{n\to\infty} P^\alpha_n = \widehat{P}^\alpha, \qquad \lim_{n\to\infty} P^\beta_n = \widehat{P}^\beta,$$

then the other also exists and $\langle \widehat{P}^\alpha, \widehat{P}^\beta \rangle$ is a Nash equilibrium point for the game G.

Proof. Consider a mapping $\widehat{\pi}^\alpha$ defined by $\widehat{\pi}^\alpha(P^\beta)(\xi^\alpha) = \widehat{\pi}^\alpha_{\xi^\alpha}(P^\beta)$. By the condition, for any measurable $P^\beta \in \widetilde{\mathcal{D}}^\beta$, the mapping $\widehat{\pi}^\alpha(P^\beta)\colon \mathcal{X}^\alpha \to \mathcal{D}^\alpha$ is also measurable, i.e., $\widehat{\pi}^\alpha\colon \widetilde{\mathcal{D}}^\beta \to \widetilde{\mathcal{D}}^\alpha$. Similarly define $\widehat{\pi}^\beta\colon \widetilde{\mathcal{D}}^\alpha \to \widetilde{\mathcal{D}}^\beta$, i.e., $\widehat{\pi}^\beta(P^\alpha)(\xi^\beta) = \widehat{\pi}^\beta_{\xi^\beta}(P^\alpha)$.

The spaces $\widetilde{\mathcal{D}}^\alpha$ and $\widetilde{\mathcal{D}}^\beta$ are metric spaces with the uniform metric defined by the metric of the spaces $\mathcal{D}^\alpha$ and $\mathcal{D}^\beta$ respectively. Thus, the equicontinuity condition for the mappings $\pi^\alpha_{\xi^\alpha}$ and $\pi^\beta_{\xi^\beta}$ immediately implies that the mappings $\widehat{\pi}^\alpha$ and $\widehat{\pi}^\beta$ are continuous.

Indeed, the equicontinuity of the mappings $\pi^\alpha_{\xi^\alpha}$, $\xi^\alpha \in \mathcal{X}^\alpha$, at the point $P^\beta \in \widetilde{\mathcal{D}}^\beta$ means that

$$\forall \varepsilon > 0\ \exists \delta\ \forall P \in \widetilde{\mathcal{D}}^\beta\ \left[d_{\widetilde{\mathcal{D}}^\beta}(P, P^\beta) < \delta \Rightarrow d_{\mathcal{D}^\alpha}(\pi^\alpha(P), \pi^\alpha(P^\beta)) < \varepsilon \right].$$

However, since

$$d_{\widetilde{\mathcal{D}}^\alpha}(P_1, P_2) = \sup_{\xi^\alpha} d_{\mathcal{D}^\alpha}(P_1, P_2),$$

this is equivalent to the continuity of the mapping $\widehat{\pi}^\alpha_{\xi^\alpha}$ at the point $P^\beta \in \widetilde{\mathcal{D}}^\beta$:

$$\forall \varepsilon > 0\ \exists \delta\ \forall P \in \widetilde{\mathcal{D}}^\beta\ \left[d_{\widetilde{\mathcal{D}}^\beta}(P, P^\beta) < \delta \Rightarrow d_{\widetilde{\mathcal{D}}^\alpha}(\widehat{\pi}^\alpha(P), \widehat{\pi}^\alpha(P^\beta)) < \varepsilon \right].$$

This implies that the conditions of Proposition 8.1 are satisfied (with the obvious substitutions $\mathcal{D}^\alpha \to \widetilde{\mathcal{D}}^\alpha$, $\mathcal{D}^\beta \to \widetilde{\mathcal{D}}^\beta$, $\pi^\alpha \to \widehat{\pi}^\alpha$, $\pi^\beta \to \widehat{\pi}^\beta$, $v^\alpha \to V^\alpha$, and $v^\beta \to V^\beta$). △

REFERENCES

1. Geanakopolos, J., Common Knowledge, *Handbook of Game Theory with Economic Applications*, Aumann, R.J. and Hart, S., Eds., Amsterdam: Elsevier, 1984, vol. 2, ch. 40.

2. McKelvey, R., The Split-Stream Harvesting Game. Part I: Mathematical Analysis, *NCAR Technical Note, National Center for Atmospheric Research*, Boulder, Colorado, 2001, pp. 1–42.

3. McKelvey, R. and Cripe, G., The Split-Stream Harvesting Game. Part II: Numerical and Simulation Studies, *NCAR Technical Note, National Center for Atmospheric Research*, Boulder, Colorado, 2001, pp. 43–57.

4. Clark, C.W., Restricted Access to Common-Property Fishery Resources: A Game-Theoretic Analysis, *Dynamic Optimization and Mathematical Economics*, Liu, P.-T., Ed., New York: Plenum, 1980, pp. 117–132.

5. Levhari, D. and Mirman, L.J., The Great Fish War: An Example Using a Dynamic Cournot–Nash Solution, *Bell J. Econ.*, 1980, vol. 11, pp. 322–344.

6. Hernández-Lerma, O. and Lasserre, J.B., *Discrete-Time Markov Control Processes. Basic Optimality Criteria*, New York: Springer, 1996.