Games and Economic Behavior 78 (2013) 31–43

Contents lists available at SciVerse ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

A dynamic epistemic characterization of backward induction without counterfactuals ☆

Giacomo Bonanno

Department of Economics, University of California, Davis, CA 95616-8578, USA

ARTICLE INFO

Article history:
Received 21 March 2012
Available online 13 December 2012

JEL classification:
C7

Keywords:
Perfect-information game
Backward induction
Dynamic interactive beliefs
Rationality
Kripke frame

ABSTRACT

We propose a dynamic framework where the rationality of a player's choice is judged on the basis of the actual beliefs that he has at the time he makes that choice. The set of "possible worlds" is given by state-instant pairs (ω, t), where each state specifies the entire play of the game. At every (ω, t) the beliefs of the active player provide an answer to the question "what will happen if I take action a?", for every available action a. A player is rational at (ω, t) if either he is not active or the action he takes is optimal given his beliefs. We characterize backward induction in terms of the following event: the first mover (i) is rational and has correct beliefs, (ii) believes that the active player at date 1 is rational and has correct beliefs, (iii) believes that the active player at date 1 believes that the active player at date 2 is rational and has correct beliefs, etc.

1. Introduction

The analysis of rational play in dynamic games is usually done within a *static* framework that specifies, for every player, his initial beliefs as well as his disposition to revise those beliefs conditional on hypothetical states of information that the player might find himself in. This is done by means of interactive structures which model a rather complex web of beliefs: for example, Player 2 might initially believe that Player 1 will end the game right away and yet have very detailed beliefs about what Player 1 would believe about Player 2's revised beliefs if Player 1 were instead to give the move to Player 2. In these models each player is assumed to have not only a disposition to revise his own beliefs, should he be faced with unexpected information, but also to have (conditional) beliefs about the disposition of the other players to revise their beliefs. This seems to constitute a rather "heavy" approach to modeling the players' states of mind in a dynamic game. It is shown in this literature (Battigalli et al., 2012; Battigalli and Siniscalchi, 2002; Ben-Porath, 1997; Samet, 1996; Stalnaker, 1998) that common *initial* belief of rationality does not imply a backward-induction outcome in perfect-information games.

In this paper we suggest an alternative and "lighter" approach, where the rationality of a player's choice is judged on the basis of the *actual beliefs* that the player has *at the time he makes that choice*. We propose a dynamic analysis of perfect-information games where the set of "possible worlds" is given by state-instant pairs (ω, t). Each state ω specifies the entire play of the game and, for every instant t, (ω, t) specifies the history that is reached at that instant (in state ω). A player is said to be *active* at (ω, t) if the history reached in state ω at time t is a decision history of his. At every state-instant

☆ I am grateful to two anonymous referees and the Advisory Editor for helpful comments and suggestions. A first draft of this paper was presented at the *First CSLI Workshop on Logic, Rationality and Interaction*, Stanford University, June 2012, and at the *Tenth Conference on Logic and the Foundations of Game and Decision Theory* (LOFT10), University of Sevilla, June 2012.

E-mail address: gfbonanno@ucdavis.edu.
URL: http://www.econ.ucdavis.edu/faculty/bonanno/.

0899-8256/$ – see front matter
http://dx.doi.org/10.1016/j.geb.2012.12.004

pair (ω, t) the beliefs of the active player provide an answer to the question "what will/might happen if I take action a?", for every available action a. A player is said to be rational at (ω, t) if either he is not active there or the action he ends up taking at state ω is optimal given his beliefs at (ω, t). We provide a characterization of backward induction in terms of the following event: the first mover (i) is rational and has correct beliefs, (ii) believes that the active player at date 1 is rational and has correct beliefs, (iii) believes that the active player at date 1 believes that the active player at date 2 is rational and has correct beliefs, etc.

This can be stated more precisely as follows. First we define a time-t belief operator B_t which captures the beliefs of the active player and enables us to express a player's belief that the next player will respond rationally to his choice. Let $\mathbf{T}_t$ be the set of states where the active player at date t (if there is any) has correct beliefs and let $\mathbf{R}_t$ be the set of states where the choice of the active player at date t is rational. In keeping with the literature, we focus on perfect-information games with no relevant ties where there is a unique backward-induction solution. We prove the following characterization. For every m greater than or equal to the depth of the game, if $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap B_0 B_1(\mathbf{T}_2 \cap \mathbf{R}_2) \cap \cdots \cap B_0 B_1 \cdots B_{m-2}(\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ then the play associated with ω is the backward-induction play. Conversely, if z is the backward-induction play then there is a model of the game and a state ω such that $\omega \in (\mathbf{T}_0 \cap \mathbf{R}_0) \cap B_0(\mathbf{T}_1 \cap \mathbf{R}_1) \cap \cdots \cap B_0 B_1 \cdots B_{m-2}(\mathbf{T}_{m-1} \cap \mathbf{R}_{m-1})$ and the play associated with ω is z.

Thus we provide an epistemic characterization of backward induction that does not rely on (objective or subjective) counterfactuals or on dispositional belief revision. Furthermore, strategies do not play any role in our framework.

The analysis is developed in Sections 2 and 3, while Section 4 is devoted to a discussion of conceptual aspects of the proposed approach and of related literature. The proofs are given in Appendix A.

2. Perfect-information games and models

We use the history-based definition of extensive-form game. If A is a set, we denote by A^* the set of finite sequences in A. If $h = \langle a_1, \ldots, a_k \rangle \in A^*$ and $1 \leqslant j \leqslant k$, the sequence $\langle a_1, \ldots, a_j \rangle$ is called a *prefix* of h. If $h = \langle a_1, \ldots, a_k \rangle \in A^*$ and $a \in A$, we denote the sequence $\langle a_1, \ldots, a_k, a \rangle \in A^*$ by ha.

A *finite extensive form with perfect information* (without chance moves) is a tuple $\langle A, H, N, \iota \rangle$ whose elements are:

- A finite set of *actions* A.
- A finite set of *histories* $H \subseteq A^*$ which is closed under prefixes (that is, if $h \in H$ and $h' \in A^*$ is a prefix of h, then $h' \in H$). The null history $\langle\rangle$, denoted by $\emptyset$, is an element of H and is a prefix of every history. A history $h \in H$ such that, for every $a \in A$, $ha \notin H$, is called a *terminal history*. The set of terminal histories is denoted by Z. $D = H \setminus Z$ denotes the set of non-terminal or *decision* histories. For every history $h \in D$, we denote by $A(h)$ the set of actions available at h, that is, $A(h) = \{a \in A\colon\ ha \in H\}$.
- A finite set N of *players*.
- A function $\iota : D \to N$ that assigns a player to each decision history. Thus $\iota(h)$ is the player who moves at history h. For every $i \in N$, let $D_i = \iota^{-1}(i)$ be the set of histories assigned to player i.

Given an extensive form, one obtains an *extensive game* by adding, for every player $i \in N$, a *utility* (or *payoff*) *function* $U_i : Z \to \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers; recall that Z is the set of terminal histories).

Given a history $h \in H$, we denoted by $\ell(h)$ the length of h, which is defined recursively as follows: $\ell(\emptyset) = 0$ and if $h \in D$ and $a \in A(h)$ then $\ell(ha) = \ell(h) + 1$. Thus $\ell(h)$ is equal to the number of actions that appear in h; for example, if $h = \langle \emptyset, a_1, a_2, a_3 \rangle$ then $\ell(h) = 3$. We denote by $\ell^{\max}$ the length of the maximal histories in H: $\ell^{\max} = \max_{h \in H}\{\ell(h)\}$. Clearly, if $\ell(h) = \ell^{\max}$ then $h \in Z$. Given a history $h \in H$ and an integer t with $0 \leqslant t \leqslant \ell^{\max}$, we denote by h_t the prefix of h of length t. For example, if $h = \langle \emptyset, a, b, c, d \rangle$, then $h_0 = \emptyset$, $h_2 = \langle \emptyset, a, b \rangle$, etc.

From now on histories will be denoted more succinctly by listing the corresponding actions, without angled brackets and without commas: thus instead of writing $\langle \emptyset, a_1, a_2, a_3, a_4 \rangle$ we will simply write $a_1 a_2 a_3 a_4$.

Let Ω be a set of states and $T = \{0, 1, \ldots, m\}$ a set of instants or dates. We call $\Omega \times T$ the set of *state-instant pairs*. If $E \subseteq \Omega \times T$ and $t \in T$, we denote by E_t the set of states $\{\omega \in \Omega\colon\ (\omega, t) \in E\}$.

Definition 1. Given an extensive form with perfect-information $G = \langle A, H, N, \iota \rangle$, a *state-time representation of* G is a triple $\langle \Omega, T, \zeta \rangle$ where Ω is a set of states, $T = \{0, 1, \ldots, m\}$ with $m \geqslant \ell^{\max}$ (recall that $\ell^{\max}$ is the depth of the game) and $\zeta : \Omega \to Z$ is a function that assigns to every state a terminal history. Given a state-instant pair $(\omega, t) \in \Omega \times T$, let

$$\zeta_t(\omega) = \begin{cases} \text{the prefix of } \zeta(\omega) \text{ of length } t & \text{if } t < \ell(\zeta(\omega)), \\ \zeta(\omega) & \text{if } t \geqslant \ell(\zeta(\omega)). \end{cases}$$

Interpretation: the play of the game unfolds over time; the first move is made at date 0, the second move at date 1, etc. A state $\omega \in \Omega$ specifies a particular play of the game (that is, a complete sequence of moves leading to terminal history $\zeta(\omega)$); $\zeta_t(\omega)$ denotes the "state of play at time t" in state ω, that is, the partial history of the play up to date t [if t is less than the length of $\zeta(\omega)$, otherwise – once the play is completed – the state of the system remains at $\zeta(\omega)$].