Substantive Rationality and Backward Induction

Joseph Halpern
Cornell University

December 9, 2011

Rationality

Making sense out of rationality is one of the major preoccupations of game theory.

- What is the "right" way to play a game if you are rational?
- What if you know your opponent is rational too?
- What if he knows you know and you know he knows?
- ...

At the 1998 TARK Conference (Theoretical Aspects of Rationality and Knowledge), there was a 2.5 hour round table discussion on "Common knowledge of rationality and the backward induction solution for games of perfect information"

Robert Aumann and Robert Stalnaker stated the following theorems:

Aumann's Theorem Common knowledge of substantive rationality implies the backwards induction solution in games of perfect information.

Stalnaker's Theorem: Common knowledge of substantive rationality does not imply the backwards induction solution in games of perfect information.

Both Aumann and Stalnaker seem to be using the words the same way.

The goal of this talk:

- Explain all the relevant notions
- Define a formal model in which both theorems are correct
- Show that Aumann and Stalnaker interpret "substantive rationality" differently
- Bring out the key role of counterfactuals

Games

A game is best thought of in terms of a *game tree*.

- Each vertex is associated with an agent (the player who moves)
- The edges represent possible moves
- The numbers at the leaves represent the utilities (rewards/payoffs) to the agent if that leaf is reached

Games of Perfect Information

A game of *perfect information* is one where all the players know what vertex in the tree they are at at all times during the play.

- Chess is a game of perfect information (if we assume perfect recall)
- Poker is not

Strategies

A *strategy* for player i describes what player i does at his move.

- Formally, a strategy for player i in game Γ is a function from player i's vertices in Γ to actions
- a strategy is a *universal plan*
- it contains *counterfactual* information
  - Even if player i goes left at the root, a strategy says what player i would do at vertices in the right subtree

In an n-player game, a *strategy profile* is a tuple $s = (s_1, \ldots, s_n)$, with a strategy for each player.

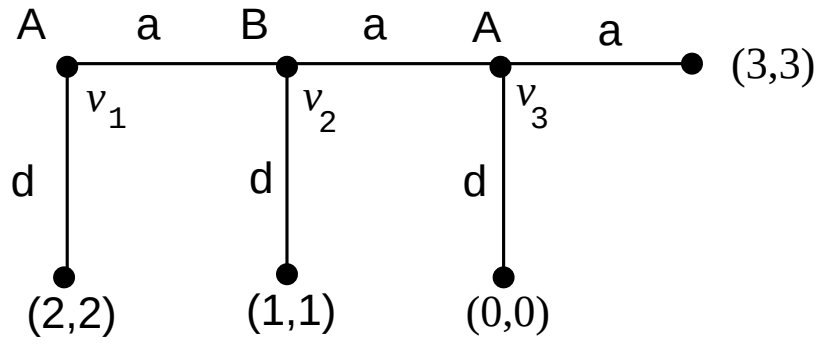
- if the strategies are deterministic, the profile determines a unique outcome

The Backward Induction Solution

This goes back to Zermelo:

- Working backwards from the leaves of the tree, make the optimal move under the assumption that all the other players will do the same

Example:



The Centipede Game

Backward induction seems reasonable, but consider the *centipede game* [Rosenthal]:

What assumptions force the backward induction strategy?

Rationality

Intuitively, a strategy for player i is rational at vertex v if player i does not make a "silly" move at vertex v, relative to his beliefs.

(Probabilistic) rationality: Player i's strategy is rational if it maximizes i's expected utility, given i's beliefs.

- if i has a probability distribution over the strategies being used by other agents, this his strategy is rational if no other strategy gives a higher expected utility, subject to i's beliefs.

Non-probabilistic version: Player i's strategy is rational if there is no other strategy that he knows will give him a better payoff.

Both Aumann and Stalnaker use the nonprobabilistic definition of rationality.

Knowledge and Common Knowledge

Intuitively, a player knows a fact p if it is true at all the worlds he considers possible.

- Our "possible worlds" will correspond to different strategy profiles

p is common knowledge if everyone knows p, everyone knows that everyone knows p, ...

A Formal Model

Fix a game Γ. A *model of* Γ is a tuple:

$$(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$$

- Ω is a set of states of the world
- $\mathcal{K}_1, \ldots, \mathcal{K}_n$ are partitions of Ω (equivalence relations) each player $i = 1, \ldots, n$
- $\mathbf{s}(\omega)$ is a strategy profile

Agent i knows p at state ω if p is true at all the states in $\mathcal{K}_i(\omega)$

We assume that players know their strategies:

- if $\omega' \in \mathcal{K}_i(\omega)$, then $\mathbf{s}_i(\omega) = \mathbf{s}_i(\omega')$.

p is common knowledge if everyone knows p, everyone knows that everyone knows, ...

- easy to show: p is common knowledge if p is true at all states reachable by any suquence of $\mathcal{K}_i$'s

Once we define rationality, we will be able to say

- at state ω, everyone knows that player 1 is rational, rationality is common knowledge, etc.

Rationality at a Vertex

Note that a strategy profile s and vertex v uniquely determine an outcome.

- $h_i^v(s)$ is i's payoff if s is used starting at v

Player i is rational at v if there is no strategy that i could have used that i knows would net him a higher payoff than the strategy he actually uses.

- Formal version: Player i *is rational at vertex* v *in* ω if, for all strategies $s^i \neq \mathbf{s}_i(\omega)$, $h_i^v(\mathbf{s}(\omega')) \geq h_i^v((\mathbf{s}_{-i}(\omega'), s^i)$ for some $\omega' \in \mathcal{K}_i(\omega)$.

Note that whether i is rational at vertex v in ω depends on i's beliefs at ω.

Both Aumann and Stalnaker use this definition of rationality.

Substantive Rationality: Aumann's Version

Aumann and Stalnaker both give the same informal definition for *substantive rationality*

- player i is *substantively rational* if i rational at all vertices in the tree

They formalize the notion in different ways though.

- This accounts for the different theorems

Aumann's definition: player i is *A-rational* at ω if i rational at vertex v in ω for all v.

Can now formally state and prove Aumann's theorem.

Aumann's Theorem

Aumann's Theorem: If Γ is a nondegenerate game of perfect information, and ω is a state in a model of Γ where it is common knowledge that all players are A-rational, then the backward induction strategy is used at ω.

Proof: Suppose $s = \mathbf{s}(\omega)$. We show by backward induction on the height of vertex v that the action taken according to s at v is the one dictated by the backward induction strategy.

This is immediate from rationality if $h(v) = 1$.

Suppose $h(v) = k > 1$ and player i moves at vertex v. The induction hypothesis says that the backward induction action is taken at all vertices below v. Since i is rational, it's true for v too.

Substantive Rationality: Stalnaker's Version

Stalnaker's intuition for substantive rationality:

For each vertex v, *if* i were actually to reach v, then what he would do in that case would be rational.

- **Problem:** how do we decide what i would do in state ω if he were to reach v if he doesn't reach v in ω?
  - Note: this becomes a counterfactual.
- Stalnaker's solution: we must consider what i would do in the state ω' "closest" to ω where he actually reaches v.
  - In particular, we must consider whether i is rational at vertex v in ω'.
  - This means we must consider i's beliefs at ω', not just i's beliefs at ω.
  - This is the standard approach for dealing with counterfactuals.

Modeling Counterfactuals

To capture Stalnaker's notion of substantive rationality formally, we need to be able to model counterfactual reasoning.

- In particular, we need to be able to describe the world "closest" to ω

An *extended model of* Γ is a tuple $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s}, f)$:

- $(\Omega, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathbf{s})$ is a model of Γ
- $f : \Omega \times G_i \to \Omega$.
  - if $f(\omega, v) = \omega'$, then state ω' is the state closest to ω where vertex v is reached.

Reasonable constraints on f:

F1. v is reached in $f(\omega, v)$

- v is on the path determined by $\mathbf{s}(f(\omega, v))$.

F2. If v is reached in ω, then $f(\omega, v) = \omega$.

F3. $\mathbf{s}(f(\omega, v))$ and $\mathbf{s}(\omega)$ agree on the subtree of Γ below v.

Stalnaker's Theorem

Stalnaker's version of substantive rationality:

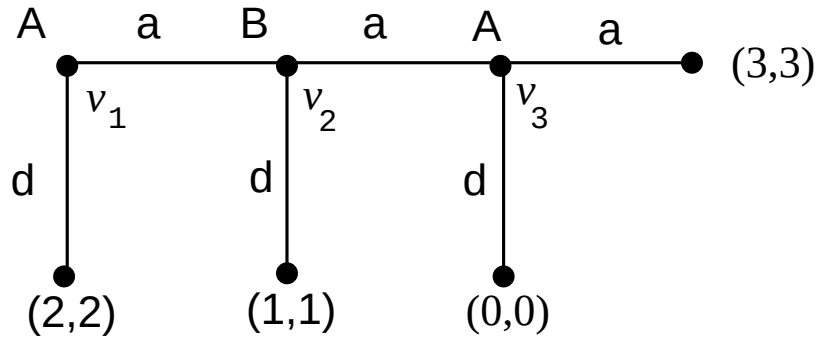
Player i is *S-rational* at vertex v in state ω if i is rational at v in $f(\omega, v)$.

- i's action is the same at v in $f(\omega, v)$ and ω (F3)
- But i's beliefs may be different at ω and $f(\omega, v)$
- We check whether i's action is rational given i's beliefs at $f(\omega, v)$.

S-rationality holds at ω if S-rationality holds at all vertices v in ω.

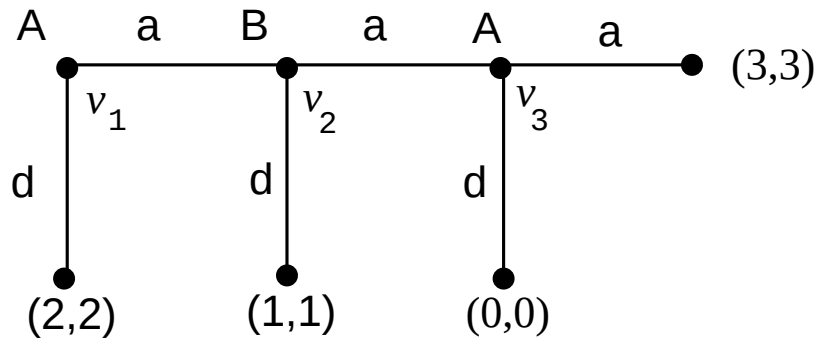
Stalnaker's Theorem: There exists a nondegenerate game Γ of perfect information and an extended model M of Γ in which the selection function satisfies F1–F3 such that S-rationality is common knowledge at some state ω in M but the backward induction solution is not played at ω.

Suppose Ann and Bob play the following game:



Consider the extended model $(\omega, \mathcal{K}_{Ann}, K_{Bob}, \mathbf{s}, f)$:

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$;
- $\mathcal{K}_{Ann}(\omega_i) = \{\omega_i\}$, for $i = 1, \ldots, 5$;
  - Ann knows what strategy is being used
- $\mathcal{K}_{Bob}(\omega_i) = \{\omega_i\}$ for $i = 1, 4, 5$; $\mathcal{K}_{Bob}(\omega_2) = \mathcal{K}_{Bob}(\omega_3) = \{\omega_2, \omega_3\}$;
  - If Bob plays d, he doesn't know whether Ann plays ad or aa
  - If p is true at ω_1, it is common knowledge
- $\mathbf{s}(\omega_1) = (da, d)$;
  $\mathbf{s}(\omega_2) = (aa, d)$;
  $\mathbf{s}(\omega_3) = (ad, d)$;
  $\mathbf{s}(\omega_4) = (aa, a)$: backwards induction solution
  $\mathbf{s}(\omega_5) = (ad, a)$.
- f satisfies F1–F3.



- $\mathbf{s}(\omega_1) = (da, d)$; $\mathbf{s}(\omega_2) = (aa, d)$; $\mathbf{s}(\omega_3) = (ad, d)$;
- $\mathcal{K}_{Bob}(\omega_2) = \{\omega_2, \omega_3\}$;

Claim: At ω_1, Ann and Bob have common knowledge of S-rationality, but do not play the backward induction strategy.

- Bob is rational at v_2 in ω_2
  - Bob considers it possible at ω_2 that Ann may go down at v_3.
- Ann is rational at v_3 in ω_4
- Since $f(\omega_1, v_2) = \omega_2$ and $f(\omega_1, v_3) = \omega_4$, we have S-rationality at ω_1
- So we have c.k. of S-rationality at ω_1.
- But $\mathbf{s}(\omega_1)$ is not the backward induction strategy.

A-rationality is *not* common knowledge at ω_1:

- Bob is not rational at v_2 in ω_1, since he plays down
- Therefore, Bob is not A-rational
- So A-rationality is not common knowledge.

Discussion

In an extended model of the Ann-Bob game, Ann can say

> Although it is common knowledge that I would play across if v_3 were reached, if I were to play across at v_1, Bob would consider it possible that I would play down at v_3.

Can't say this in Aumann's framework, without selection functions.

- Aumann has no way of allowing the agents to revise their beliefs in the context of counterfactual reasoning.

Recovering Aumann's Theorem

With an extra condition on selection functions, we can recover Aumann's Theorem in extended models.

F4. For all players i and vertices v, if $\omega' \in \mathcal{K}_i(f(\omega, v))$ then there is a state $\omega'' \in \mathcal{K}_i(\omega)$ such that $\mathbf{s}(\omega')$ and $\mathbf{s}(\omega'')$ agree on the subtree of Γ below v.

Intuitively, player i considers at least as many strategies possible in ω as at $f(\omega, v)$.

- beliefs don't get revised radically

Theorem If Γ is a nondegenerate game of perfect information, and ω is a state in an extended model of Γ in which the selection function satisfies F1–F4 and it is common knowledge that all players are S-rational, then the backward induction strategy is used at ω.

Rationality and Strategies

What does it mean to say that Ann's strategy at ω is s?

- What does it mean that $s(v_3) = a$ if Ann's beliefs guarantee that v_3 is never reached?
- Does it mean that, given her current beliefs about Bob, Ann would play a if she were at v_3, or given the beliefs she would have if for some strange reason she found herself at v_3, she would play a?

Modeling Counterfactual Reasoning in Games

- Strategies involve counterfactuals.
- Extended models have another source of counterfactuals (selection functions)

Quote from Stalnaker:

> To clarify the causal and epistemic concepts that interact in strategic reasoning, it is useful to break them down into their component parts.

This suggests it is useful to have a model where strategies are not primitive, but are defined in terms of counterfactuals.

- Samet [1996] does this
- He also gives conditions under which Aumann's Theorem holds in his framework.
  - Common knowledge of rationality isn't enough.