Lecture VII: Common Knowledge

Markus M. Möbius

March 4, 2003

Readings for this class: Osborne and Rubinstein, sections 5.1,5.2,5.4

Today we formally introduce the notion of common knowledge and discuss the assumptions underlying players' knowledge in the two solution concepts we discussed so far - IDSDS and Nash equilibrium.

1 A Model of Knowledge

There is a set of states of nature $\Omega = \{\omega_1, \omega_2, .., \omega_n\}$ which represent the uncertainty which an agent faces when making a decision.

Example 1 *Agents* $1, 2$ *have a prior over the states of nature*

$$\begin{aligned} \Omega &= \{\omega_1 = \textit{It will rain today}, \omega_2 = \textit{It will be cloudy today}, \\ \omega_3 &= \textit{It will be sunny today}\} \end{aligned}$$

where each of the three events is equally likely ex ante.

The knowledge of every agent i is represented by an information partition H_i of the set Ω.

Definition 1 *An information partition* H_i *is a collection* $\{h_i(\omega) | \omega \in \Omega\}$ *of disjoint subsets of* Ω *such that*

- *(P1)* $\omega \in h_i(\omega)$,
- *(P2) If* $\omega' \in h_i(\omega)$ *then* $h_i(\omega') = h_i(\omega)$.

Note, that the subsets $h_i(\omega)$ span Ω. We can think of $h_i(\omega)$ as the knowledge of agent i if the state of nature is in fact ω. Property P1 ensures that the true state of nature ω is an element of an agent's information set (or knowledge) - this is called the axiom of knowledge. Property P2 is a consistency criterion. Assume for example, that $\omega' \in h_i(\omega)$ and that there is a state $\omega'' \in h_i(\omega')$ but $\omega'' \notin h_i(\omega)$. Then in the state of nature is ω the decision-maker could argue that because ω'' is inconsistent with his information the true state can not be ω'.

Example 1 *(cont.) Agent* 1 *has the information partition*

$$H_1 = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$$

So the agent has good information if the weather is going to be sunny but cannot distinguish between bad weather.

We next define a knowledge function K.

Definition 2 *For any event E (a subset of Ω) we have*

$$K(E) = \{\omega \in \Omega | h_i(\omega) \subseteq E\}.$$

So the set $K(E)$ is the collection of all states in which the decision maker knows E.

We are now ready to define common knowledge (for simplicity we only consider two players).

Definition 3 *Let K_1 and K_2 be the knowledge functions of both players. An event $E \subseteq \Omega$ is common knowledge between 1 and 2 in the state $\omega \in \Omega$ if ω if a member of every set in the infinite sequence $K_1(E)$, $K_2(E)$, $K_1(K_2(E))$, $K_2(K_1(E))$ and so on.*

This definition implies that player 1 and 2 knows E, they know that the other player knows it, and so on.

There is an equivalent definition of common knowledge which is frequently easier to work with.

Definition 4 *An event $F \subseteq \Omega$ is self-evident between both players if for all $\omega \in F$ we have $h_i(\omega) \subseteq F$ for $i = 1, 2$. An event $E \subseteq \Omega$ is common knowledge between both players in the state $\omega \in \Omega$ if there is a self-evident event F for which $\omega \in F \subseteq E$.*

Example 1 *(cont.) Agent 2 has information function*

$$H_2 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}\}$$

In this case the event $E = \{\omega_1, \omega_2\}$ is common knowledge if the state of nature is ω_1 or ω_2. Both definition can be applied - E survives iterated deletion, but is also self-evident.

We finally show that both definitions of common knowledge are equivalent. We need the next proposition first.

Proposition 1 *The following are equivalent:*

1. *$K_i(E) = E$ for $i = 1, 2$*
2. *E is self evident between 1 and 2.*
3. *E is a union of members of the partition induced by H_i for $i = 1, 2$.*

Proof: Assume a). Then for every $\omega \in E$ we have $h_i(\omega) \subseteq E$ and b) follows. c) follows because immediately. c) implies a).

We can now prove the following theorem.

Theorem 1 *Definitions 3 and 4 are equivalent.*

Proof: Assume that the event E is common knowledge in state ω according to definition 3. First note, that

$$E \supseteq K_i(E) \supseteq K_j(K_i(E)) ...$$

Because Ω is finite and ω is a member of those subsets the infinite regression must eventually produce a set F such that $K_i(F) = F$. Therefore, F is self-evident and we are done.

Next assume that the event E is common knowledge in state ω according to definition 4. Then $F \subseteq E$, $K_i(F) = F \subseteq K_i(E)$ etc, and F is a member of every of the regressive subsets $K_i(K_j(...E...))$ and so is ω. This proves the theorem.

2 Dirty Faces

We have played the game in class. Now we are going to analyze it by using the mathematical language we just developed. Recall, that any agent can only see the faces of all $n-1$ other agents but not her own. Furthermore, at least one face is dirty.

First of all we define the states of nature Ω. If there are n players there are $2^n - 1$ possible states (since all faces clean cannot be a state of nature). It's convenient to denote the states by the n-tuples $\omega = (C, C, D, .., C)$. We also denote the number of dirty faces with $|\omega|$ and note that $|\omega| \geq 1$ by assumption (there is at least one dirty face).

The initial information set of each agent in some state of nature ω has at most two elements. The agent knows the faces of all other agents $\omega(-i)$ but does not know if she has a clean or dirty face, i.e. $h_i(\omega) = \{(C, \omega(-i)), (D, \omega(-i))\}$. Initially, all agents information set has two elements except in the case $|\omega| = 1$ and one agent sees only clean faces - then she know for sure that she has a dirty face because all clean is excluded. You can easily show that the event "There is at least one dirty face" is common knowledge as you would expect.

The game ends when at least player knows the state of the world for sure, i.e. her knowledge partition $h_i(\omega)$ consists of a single element. In the first this will only be the case if the state of world is such that only a single player has a dirt face.

What happens if no agent raises her hand in the first period? All agents update their information partition and exclude all states of nature with just one dirty face. All agents who see just one dirty face now know for sure the state of nature (they have a dirty face!). They raise their hand and the game is over. Otherwise, all agents can exclude states of nature with at most two dirty faces. Agents who see two dirty faces now know the state of nature for sure (they have a dirty face!). etc.

The state of nature with k dirty faces therefore gets revealed at stage k of the game. At that point all guys with dirty faces know the state of nature for sure.

Question: What happens if it is common knowledge that neither all faces are clean, nor all faces are dirty?

Remark 1 *The game crucially depends on the fact that it is common knowledge that at least one agent has a dirty face. Assume no such information would be known - so the state of the world where all faces are clean would*

be a possible outcome. Then no agent in the first round would ever know for sure if she had a dirty face. Hence the information partition would never get refined after any number of rounds.

3 Coordinated Attack

This story shows that 'almost common knowledge' can be very different from common knowledge.

Two divisions of an army are camped on two hilltops overlooking a common valley. In the valley waits the enemy. If both divisions attack simultaneously they will win the battle, whereas if only one division attacks it will be defeated. Neither general will attack unless he is sure that other will attack with him.

Commander A is in peace negotiations with the enemy. The generals agreed that if the negotiations fail, commander A will send a message to commander B with an attack plan. However, there is a small probability ϵ that the messenger gets intercepted and the message does not arrive. The messenger takes one hour normally. How long will it take to coordinate on the attack?

The answer is: never! Once commander B receives the message he has to confirm it - otherwise A is not sure that he received it and will not attack. But B cannot be sure that A receives his confirmation and will not attack until he receives another confirmation from A. etc. The messenger can run back and forth countless times before he is intercepted but the generals can never coordinate with certainty.

Let's define the state of nature to be (n, m) if commander A sent n messages and received $n-1$ confirmation from commander B, and commander B sent m messages and received $m-1$ confirmations. We also introduce the state of nature $(0, 0)$. In that state the peace negotiations have succeeded, and no attack is scheduled.[1]

The information partition of commander A is

$$H_A = \{\{(0,0)\}, \{(1,0),(1,1)\}, \{(2,1),(2,2)\}, ..\}. \tag{1}$$

[1] As was pointed out by an alert student in class this state is necessary to make this exercise interesting. Otherwise, the generals could agree on an attack plan in advance, and no communication would be necessary at all - the attack would be common knowledge already.

The information partition of commander B is

$$H_B = \{\{(0,0),(1,0)\},\{(1,1),(2,1)\},..\}. \tag{2}$$

Both commanders only attack in some state of the world ω if it is common knowledge that commander B has sent a message, i.e. $n \geq 1$ (the negotiations have failed and an attack should occur). However, this event can never be common knowledge for any state of nature (i.e. after any sequence of messages) because there is no self-evident set F contained in the event E. This is easy to verify: take the union of any collection of information sets of commander A (only those can be candidates for a self-evident F). Then ask yourself whether such a set can be also the union of a collection of information sets of commander B. The answer is no - there will always some information set of B which 'stick out' at either 'end' of the candidate set F.

4 Knowledge and Solution Concepts

We finally apply our formal definition of knowledge to redefine IDSDS and Nash equilibrium for 2-player games. We first have to define what we mean with a 'state of the world'. Formally, the set Ω is the environment in which the game is played. In each state we have:

- The action $a_i(\omega)$ taken by agent i in state ω.
- $\mu_i(\omega)$ is agent i's belief about the actions taken by his opponent.
- Some information partition $h_i(\omega)$ which describes player i's knowledge in state ω.

One way to generate the state of the world would be to take each possible action profile $s \in S$ and combine it with each possible belief $\mu \in \Sigma_1 \times \Sigma_2$ and call this combination (s, μ) a 'state'.

4.1 Nash Equilibrium

Nash equilibrium restricts the information partition of players very strongly. Basically, players (i) need to know other players' actions, (ii) have beliefs which are consistent with their knowledge, (iii) be rational.

Proposition 2 *Suppose that in the state $\omega \in \Omega$ each player i*

- *knows the other players' actions:* $h_i(\omega) \subseteq \{\omega' \in \Omega | a_{-i}(\omega') = a_{-i}(\omega)\}$
- *has a belief which is consistent with his knowledge: the support* $\mu_i(\omega)$ *is a subset of* $\{a_{-i}(\omega') \in A_{-i} : \omega' \in h_i(\omega)\}$
- *is rational:* $a_i(\omega)$ *is a best-response of player* i *to* $\mu_i(\omega)$

Then $(a_i(\omega))$ *is a Nash equilibrium of* G*.*

The proof is not difficult. The action of player i is a best-response to his beliefs which in turn put probability 1 on the other players playing strategy a_{-i}.

4.2 IDSDS

For applying iterated deletion of strictly dominated strategies we can use a less restrictive information partition.

Proposition 3 *Suppose that in state* ω *in a two-player game it is common knowledge between players that each player's belief is consistent with his knowledge and that each player is rational. That is, suppose that there is a self-evident event* $\omega \in F$ *such that for every* $\omega' \in F$ *and each player* i *we have:*

- *the support* $\mu_i(\omega')$ *is a subset of* $\{a_{-i}(\omega'') \in A_{-i} : \omega'' \in h_i(\omega')\}$
- *is rational:* $a_i(\omega')$ *is a best-response of player* i *to* $\mu_i(\omega')$

Then the action $a_i(\omega)$ *is an element of* S_i^∞*.*