The centipede game

In the two-period game discussed in section, the final outcome was fairly intuitive: agent 2 acts in her own self-interest and agent 1 acts in his, leading to a well-motivated predicted outcome. The purpose of the following exercise is to show that this method of reasoning can sometimes lead us to counterintuitive results.[1]

We are going to consider a class of games called *Centipede games*.[2] This is a *class of* games since particular versions can vary in different ways, but they should all share some certain qualitative properties.

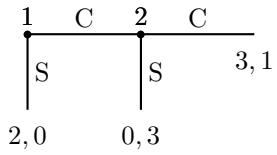

Figure 1: A two-period centipede game

The simplest version is the two-period sequential game shown in Figure 1. Agent 1 acts first, choosing whether to **C**ontinue or **S**top; if he chooses S, the game ends and payoffs are $(1, 0)$. If he chooses C, the game continues and agent 2 chooses C or S.

Using backward induction, we can predict behavior in this game fairly simply: agent 2 should choose S rather than C, as it yields payoff 3 rather than 1. Anticipating that agent 2 will do this, agent 1 should choose S rather than C, as it yields payoff 2 rather than 0. Thus in equilibrium, agent 1 stops the game immediately, and payoffs are $(2, 0)$; strategies may be written as $s_1 = (S), s_2 = (S)$ — recall that agent 2 will choose S if she is given the option.

We can see that this is a breakdown of trust in the face of rational, strategic agents: if they could both agree to choose C, the ultimate payoffs would be $(3, 1)$ and each agent would be better off! Unfortunately, rationality implies that agent 1 cannot trust agent 2 to fulfill her promise to choose C, hence we must end up halting play immediately.

This may not seem too terrible; after all, there is not much being sacrificed. However, consider the centipede game in Figure 2.[3] Moreover, appealing to our vast powers of human cognition, notice that we can make centipede games *arbitrarily long*: i.e., they can last 2 periods, 8 periods, 1000 periods, or more.

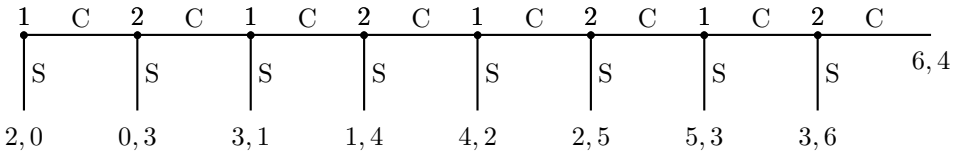

Figure 2: An eight-period centipede game

What does game theory predict here? Agent 1's choice today depends on agent 2's choice tomorrow, which in turn depends on agent 1's choice the day after tomorrow, &c. Although this phrasing very quickly becomes unwieldy, backward induction is still a very tractable method of analyzing this problem: in the last stage,

[1] Jernej referred to this game in passing in lecture, when he mentioned chess grandmasters playing each other.

[2] Economists *love* naming things. Although it is ridiculous on its face, this gives us a useful shorthand for discussing strategic scenarios without having to spell out all of the parameters of a particular model.

[3] Notice now that it is a little clearer why this game is called "centipede".

agent 2 will choose S rather than C since 6 is bigger than 4. In the penultimate stage, then, agent 1 will choose S rather than C since 5 is bigger than 3. When we iterate this logic back to the beginning of the game, we see that agent 1 must choose S immediately! We call this *unraveling*, after the manner in which "good" behavior is undone by the inevitability of "bad" behavior in the future.

This result is often considered counterintuitive because of its generality: if we played a 1000-period centipede game, we could both obtain payoffs of nearly 500 (or over 500) if we cooperated, but by stopping immediately we get payoffs of only $(2, 0)$. As an optimist, this is frustrating. Fortunately, this game also exposes some of the weaknesses of game theory: in reality, people often cooperate for a certain period of time, then stop a little before the final periods of the game.

Lest you feel too positive about that, consider the following real-world scenario: you are working with a classmate on a group project, for which you both will receive the same grade. You find out that your partner is a slacker; it's a situation we all recognize. However, consider this: is slacking in this scenario a character flaw, or is it simply the rational thing to do? There is a reason economics is called "the dismal science".

Missing information

This section is beyond the scope of this Monday's (October 8) lecture, however this discussion will be central later in the quarter.

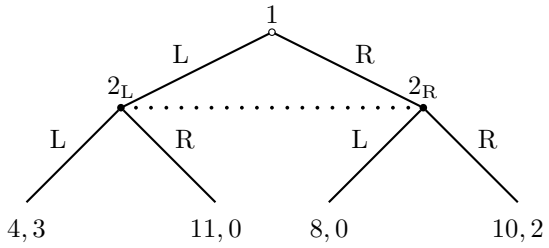

Figure 3: The two-agent sequential game from section

Recall the game discussed in section, shown in Figure 3; here it has been modified slightly, to include a dashed line between the two decision points of agent 2, 2_L and 2_R. We made a sizable assumption in class: agent 2 can distinguish between the two situations in which she may find herself. In many contexts, this is a very reasonable assumption; we can think of agent 1 acting first, agent 2 observing this action, then choosing an action herself.

As is always the case in economics, there are also many settings in which this assumption *is not* valid. For example, agents 1 and 2 may be attending a potluck but don't have each other's contact info; they would like to bring complementary dishes (nobody wants to eat ten kinds of cole slaw for dinner), but they will need to choose what to bring without observing what the other is preparing. A more literal interpretation in terms of the game in Figure 3 is that agent 1 moves first by writing his chosen action on an index card, then agent 2 moves *without witnessing what is written on the card*, writing her own choice on an index card. The outcome of the game is determined by both players simultaneously revealing what they have written.

In a setting such as this, agent 2 cannot distinguish 2_L from 2_R when she is making her choice. Notationally, we indicate this by drawing a dashed line between her decision points. Further, since she can't tell which is which, there is no sense labeling each point individually, so the convention has us labeling only the dashed line (as in Figure 4; this line is referred to as her *information set*.

The question we would like to address remains: what should agent 1 do? What should agent 2 do? This is a useful thought experiment. Consider the following:

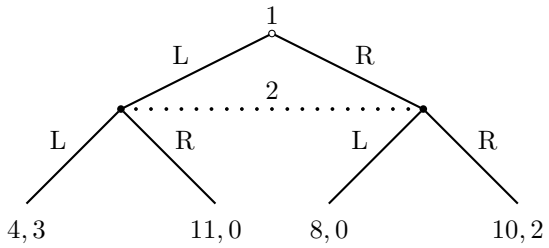

Figure 4: The two-agent sequential game from section, with agent 2 unsure of where she is

- If agent 2 chooses L, she may get 3 or she may get 0.
- If agent 2 chooses R, she may get 1 or she may get 2.

Therefore we can't really say much: 3 is better than 1, but 0 is worse than 2. We will need some more tools. So let's rephrase her choice:

- If agent 2 *thinks that* agent 1 is choosing L, she can get 3 by choosing L or 1 by choosing R.
- If agent 2 *thinks that* agent 1 is choosing R, she can get 0 by choosing L or 2 by choosing R.

It is reasonable then to say that if agent 2 thinks that agent 1 is choosing L, she should choose L, and if she thinks that agent 1 is choosing R, she should choose R.

What about agent 1? Before — when agent 2 knew where she was — agent 1 was able to condition on agent 2's inevitable decision. Now, agent 2 doesn't necessarily have such an inevitable decision. Still, we can consider agent 1's decision in the same way that we considered agent 2's above:

- If agent 1 *thinks that* agent 2 is choosing L, he can get 4 by choosing L or 8 by choosing R.
- If agent 1 *thinks that* agent 2 is choosing R, he can get 11 by choosing L or 10 by choosing R.

It is reasonable then to say that if agent 1 thinks that agent 2 is choosing L, he should choose R, and if he thinks that agent 2 is choosing R, he should choose L.

Let L_1 and R_1 denote a choice of L and R, respectively, by agent 1, and let L_2 and R_2 denote a choice of L and R, respectively, by agent 2. Then the above discussion can be summed up by

$$L_1 \rightsquigarrow L_2 \rightsquigarrow R_1 \rightsquigarrow R_2 \rightsquigarrow L_1 \rightsquigarrow \dots$$

Uh-oh!

It is not clear what game theory should predict happening in this setting. Keep this in the back of your mind; if you like, think about what we might be able to say should happen — there is more than one answer! (although, only one that we will use in Econ 101) Does this problem get any easier if the '8' and '10' are reversed in agent 1's payoffs?

Notice that this complexity arises by changing one assumption very slightly. Game theory offers a very rich set of predictions in many settings, and we will spend the quarter learning tools to apply to a wide variety of real-world (and not-so-real-world) situations.