Refinements of Sequential Equilibrium

Debraj Ray, November 2006

Sometimes sequential equilibria appear to be supported by "implausible" beliefs off the equilibrium path. These notes briefly discuss this problem and some possible fixes.

1. An Example

Consider a sequential equilibrium and a non-equilibrium announcement (such as a nonequilibrium choice of education in the Spence model). What is the other recipient of such a signal (the employer in the example above) to believe when she sees that signal?

Sequential equilibrium imposes little or no restrictions on such beliefs in signalling models. [We have seen, of course, that in other situations — such as those involving moves by Nature — that it does impose several restrictions, but not in the signalling games that we have been studying.] The reason is that sequential equilibrium is unwilling to place *any* strategic value on trembles. That's what they are, just trembles.

There is an entirely different view of trembles as strategic. We have encountered this situation before, in the context of the centipede game. To be sure, we've also appreciated the problematic nature of this strategic viewpoint when there is absolutely no information to be conveyed by the deviation (the centipede game we initially studied was one of complete and perfect information). One then needs to confront higher-order theory systems or belief systems, to make sense of the deviation.

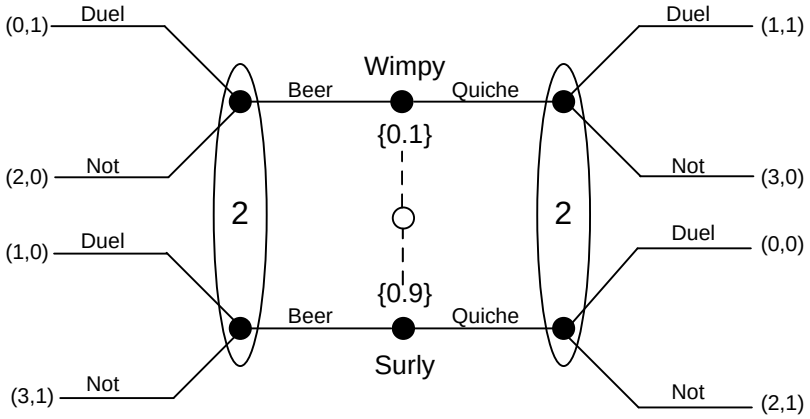
But those insisting on deviations as potentially strategic trembles will find much more of interest in situations in which there *is* information to be conveyed. The perfect example is signaling games.

Consider the following example (suffused with a nice American viewpoint on the world) due to Kreps:

In this example player 1 is endowed by Nature to be wimpy or surly, with probabilities 1/10 and 9/10 respectively. Wimps like quiche. Surlies like beer. (This yields an additive payoff of 1.) Player 1 — whetever his type — doesn't like to fight. As for player 2, he likes to duel the wimpy but not the surly. This should help you read the game tree.

Now, there is one sequential equilibrium in which both types of player 1 choose Beer, and player 2 does not duel if he sees Beer. At the (unreached) information set at which 1 chooses Quiche, player 2 believes that player 1 is wimpy with probability at least 1/2, and fights with probability at least 1/2.

There is another sequential equilibrium in which both types of player 1 choose Quiche, and player 2 does not duel if he sees Quiche. At the (unreached) information set at which 1 chooses Beer, player 2 believes that player 1 is wimpy with probability at least 1/2, and fights with probability at least 1/2.



It is easy to scheck (but please do check) that both these specifications generate bonafide sequential equilibria. But there is something odd about the beliefs supporting the second equilibrium. Notice that the wimpy type is in heaven in the second equilibrium: he gets to eat his quiche and what is more, player 2 does not fight him. In deviating to beer he *cannot* be better off, even if player 2 were not to fight him. On the other hand, the surly type is certainly better off if all beliefs about the guy being wimpy are eliminated by the argument of the previous sentence. (Notice how we must ascribe strategic motivations to deviations in order to make this argument.) Therefore player 2 should not believe that a deviation to Beer would be made by the wimpy, at least if there is room for a profitable deviation (as there is here) by the surly. This sort of reasoning underlines signaling games.

2. Signaling and Belief Restrictions

Formally, consider a two-stage game of the following type. First Nature moves and selects a type from a finite set of types T for player 1 (the "sender"), using some probability distribution μ. The sender knows her type; player 2 (the "receiver") does not. Following the realization of t, the sender chooses an action a from a finite set $A(t)$. The receiver observes this action a, and then chooses his own action from the finite set $B(a)$. Then payoffs are received: player i's payoffs are given by the function $f_i(t, a, b)$.

This is a special case of a game with incomplete information and one can simply apply, without any alternations, the concept of sequential equilibrium that we've defined earlier. In particular, in any equilibrium, player 2 must use Bayes' rule to update her beliefs at any information set (identifiable with an action a) that is reached with positive probability. But at unreached information sets, player 2 is not restricted at all. *Any* belief is compatible with sequential equilibrium. Our purpose is to examine conditions under which such beliefs can be restricted. One such condition is the *intuitive criterion*, advanced by Cho and Kreps.

To define this, fix an equilibrium and let $f^*(t)$ stand for the equilibrium payoff of the sender of type t.

Consider some non-equilibrium action (or signal) a. Consider some type of a player, and suppose even if she were to be treated in the best possible way following a, she *still* would

prefer to stick to her equilibrium action. Then we will say that signal a is *equilibrium-dominated* for the type in question. She would never want to emit that signal, except purely by error. Not strategically.

At each signal a, let $T(a)$ be the set of types that could have taken action a, and let $D(a)$ be the subset of T that is equilibrium-dominated. Define $S(a) \equiv T(a) - D(a)$.

Now say that the equilibrium is *fails the Intuitive Criterion at* a if for some $t \in S(a)$,

$$f^*(t) < f_1(t, a, b)$$

for every conceivable best response b by the receiver under beliefs restricted to $S(a)$.

If this does not happen for any a, then say that the equilibrium *satisfies the Intuitive Criterion* (IC).

Notice that IC places no restrictions on beliefs over the types that are *not* equilibrium dominated, and in addition it also places no restrictions if *every* type is equilibrium-dominated. For then — so goes the logic — the deviation signal must surely be regarded as an error, and once that possibility is admitted, all bets about who is emitting that signal are off.

Let us apply IC to the Spence model.

PROPOSITION **1.** *In the Spence signaling model with two types, a single equilibrium outcome survives the IC, and it is the separating equilibrium in which L plays 0 while H plays* e_1*, where* e_1 *solves*

$$H - \frac{e_1}{L} = L.$$

Proof. First we rule out all equilibria in which types H and L play the same value of e with positive probability. [This deals with all the pooling and all the hybrid equilibria.] At such an e, the payoff to each type θ is

$$\lambda H + (1 - \lambda)L - \frac{e}{\theta},$$

where λ represents the employer's posterior belief after seeing e. Now, there always exists an $e' > e$ such that

$$\lambda H + (1 - \lambda)L - \frac{e}{L} = H - \frac{e'}{L},$$

while at the same time,

$$\lambda H + (1 - \lambda)L - \frac{e}{H} < H - \frac{e'}{H}.$$

It is easy to see that if we choose e'' very close to e' but slightly bigger than it, it will be equilibrium-dominated for the low type —

$$\lambda H + (1 - \lambda)L - \frac{e}{L} > H - \frac{e''}{L},$$

while it is not equilibrium-dominated for the high type:

$$\lambda H + (1 - \lambda)L - \frac{e}{H} < H - \frac{e''}{H}.$$

But now the equilibrium is broken by having the high type deviate to e''. By IC, the employer must believe that the type there is high for sure and so must pay out H. But then the high type benefits from this deviation relative to playing e.

Next, consider all separating equilibria in which L plays 0 while H plays some $e > e_1$. Then a value of e' which is still bigger than e_1 but smaller than e can easily be seen to be equilibrium-dominated for the low type but not for the high type. So such values of e' must be rewarded with a payment of H, by IC. But then the high type will indeed deviate, breaking the equilibrium. This proves that the only equilibrium that can survive the IC (in the Spence model) is the one in which the low type plays 0 and the high type chooses e_1. □

Once we accept that deviations may be strategic, the IC is pretty conservative as restrictions go, and there is a variety of other stronger conditions that have been imposed. For instance, it can be checked that all sorts of fresh equilibria appear in the Spence model when there are three or more types, but stronger versions of the IC serve to rule them out.

3. Forward Induction

Related to restrictions of this sort is the *forward induction* argument. The basic idea is that an off-equilibrium signal can be due to one of two things: an error, or strategic play. If at all strategic play can be suspected, the error theory must play second fiddle: that's what a forward induction argument would have us believe. The IC and related concepts restrict types in this way; a forward induction argument signals an *intention* to play "future" stages of the game in a particular way.

Consider the following example:

A multinational decides whether to come "In" or stay "Out". If "Out", it picks up a sure payoff of 2. If "In", it plays a BOS-style game with a domestic company (the latter chooses columns below):

	L	R
T	$3, 1$	$0, 0$
B	$0, 0$	$1, 3$

One equilibrium involves the multinational staying out, while the domestic plays R is there is entry. But if all deviations are strategic, what might the domestic think if it sees the multinational enter? The multinational gave up a *sure* payoff of 2; if it enters now it must be expecting to play T and pick up 3. If this reasoning is credible, the domestic may be induced to play L. We've inducted forward.

In this example the multinational comes in or out depending on whether he takes the outside option. The following example is a bit more disturbing because it involves the play of a truly irrelevant action before the "actual game" is played. (Of course, this is is misleading: the point is that the whole thing has to be viewed as "the game".) In this situation, player 1 decides whether or not to burn a dollar. Player 2 sees this. Then they play exactly the same game above, in which 1 chooses rows and 2 chooses columns.

Note that by *not* burning the dollar and then playing T with suitable probability p, player 1 can guarantee herself 3/4. To see this, let q be any mixed strategy for player 2; i.e., her probability of playing L. Then player 1's expected return is $p(3q) + (1-p)(1-q)$. Put $p = 3/4$. Now use forward induction. If player 1 burns money, she's lost 1 then and there. The *only* way she can make this up is if the action pair (T, L) is played. So if player 2 were to treat all moves as strategic, we must conclude that if player 1 burns the dollar, player 2 must play L, thereby guaranteeing player 1 a (net) payoff of 2.

This is a bit disturbing, but it gets worse. If we agree that player 1 can guarantee herself 2 by burning the dollar, then by *not* burning that dollar she must be signaling her intention of gettingt a *still* higher payoff! So if our logic is taken to the limit, the unique outcome of this game must be that player 1 does not even burn the dollar, and then they play (T, L). Just the availability of that dollar to burn creates a selection from the subsequent equilibria of the "second-stage" game.

The normal form of this game allows us to look at it from another angle.

	L if burn, L if not	R if burn, L if not	L if burn, R if not	R if burn, R if not
Not, T	$3,1$	$3,1$	0,0	0,0
Not, B	$0,0$	$0,0$	$1,3$	$1,3$
Burn, T	$2,1$	$-1,0$	$2,1$	$-1,0$
Burn, B	$-1,0$	$0,3$	$-1,0$	$0,3$

For player 1, (Burn, B) is *strictly* dominated by a combination of (Not, T) and (Not, B), say with probability 3/4 and 1/4. If you delete this last row, then for player 2, the last column is *weakly* dominated by the third column. (Please note: it cannot be strictly dominated.) If the fourth column is thereby eliminated, this is a crucial step. Now the second row, (Not, B) can be strictly dominated by a suitable mix of the first and third rows. Once this is taken out, the second and third columns for player 2 can be removed by weak domination (another crucial step). This leaves player 1 to freely play her top row, and we "implement" the outcome discussed earlier.

Iterated elimination of weakly dominated strategies and forward induction are closely connected. There is a large literature on this.

Forward induction arguments can generate some interesting interactions — conflicts — with backward induction, and one should apply these solution concepts (if at all) with a great deal of care. Consider the following variation on our first example, due to Kohlberg. Enrich the baseline model a bit (though throw out the multinational interpretation):

Player 1 can play Out(1) and generate payoffs $(2, 0)$, or can play In(1) whereupon the following game is played, as before:

	L	R
T	$3,1$	$0,0$
B	$0,0$	$1,3$

We've seen that the unique forward-induction solution to this game is for player 1 to play In(1), with (T, L) played subsequently.

Now add an earlier node. Player 2 moves first, and can play In(2) or Out(2). If she plays Out(2), payoffs are $(0, 2)$; the game is over. If she plays In(2), then the above game is played.

Like 1, player 2 has a forward-induction signal. By playing In(2), she is "threatening" to play R in the subsequent game. But by *backward* induction, player 1 can do a similar thing when it is *his* turn to move, which is next. So who trumps whom? Use iterated weak-dominance to get a sense of this.