Economics 703
Advanced Microeconomics **Prof. Peter Cramton**

Lecture Note 5: Signaling[1]

Outline

A. Signaling Games
   1. The Intuitive Criterion
   2. Forward Induction Equilibrium
   3. D1, D2, Divinity, Universal Div.

B. Cheap Talk
   1. Strategic Information Transmission
   2. Neologisms
   3. Perfect Sequential Equilibria

1. The Intuitive Criterion (Cho & Kreps, 1987)

This paper analyzes a simple signaling model that has applications throughout information economics. There are two players: a Sender of information (S) and a Receiver of information (R). The timing of the game is: (1) nature draws a type for S, denoted $t \in T$, according to the probability distribution p(t); (2) S privately observes the type t and then sends the message $m \in M$ to R; and (3) R observes m and then takes the action $a \in A$. The three sets T, M, and A are all finite. The payoffs are $U^S(t,m,a)$ and $U^R(t,m,a)$. Everything about the game except nature's choice of t is common knowledge.

The paper proposes a test for identifying unreasonable sequential equilibria of this game. Part of the paper concerns the use of this test as an intuitive implication of the not-so-intuitive equilibrium concept designed by Kohlberg and Mertens [1986]. But the test is of substantial interest in its own right as a way to refine the set of sequential equilibria.

To explain the test Cho & Kreps propose, consider a simple version of the game: T={t,t'} and M={m,m'}. Suppose that in a particular equilibrium both types send the message m with probability one. Then the message m' is off the equilibrium path, so R's beliefs after observing m' cannot be derived from Bayes' rule. Instead, these beliefs need only satisfy Kreps and Wilson's definition of consistency in order to be part of a sequential equilibrium. (It is left as an exercise to show that in a signaling game *any* beliefs off the equilibrium path satisfy consistency.)

Sequential rationality dictates that the action R taken after observing m' must be optimal given R's beliefs. That is,

---

[1]These notes are based without restraint on notes by Robert Gibbons, MIT.

$$a(m') \in \underset{a \in A}{\operatorname{argmax}} \sum_{t \in T} m(t|m') U^R(t, m', a).$$

Suppose that (1) no matter what belief R holds, the resulting action a(m') makes type t worse off than t is in the equilibrium, and that (2) if R infers from m' that S is type t', then R's optimal action will make t' better off than t' is in the equilibrium. Then if S is type t', the following speech should be believed by R:

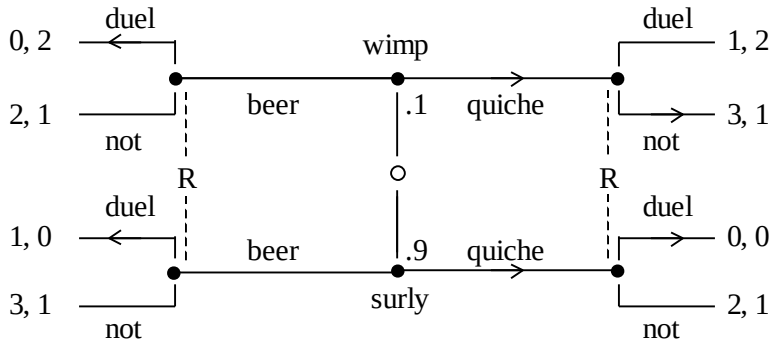
> I am t'. To prove this, I am sending m' instead of the equilibrium m. Note that if I were t I would not want to do this, no matter what you might infer from m'. And, as t', I have an incentive to do this provided it convinces you that I am not t.

Thus, given the two suppositions above, t' should deviate from the sequential equilibrium in which m is sent with probability one. On this ground, Cho & Kreps reject the equilibrium.

The paper proceeds to formalize these two suppositions and to study the consequences of rejecting the equilibria they characterize. To make things concrete, we will consider an example before getting formal.

Beer and Quiche:

The extensive-form game drawn below describes the entry-deterrence problem commonly analyzed in the industrial organization literature. The incumbent firm can be either "surly" or a "wimp." A surly firm prefers to have "beer" for breakfast, whereas a wimp prefers "quiche." However, it is worth enduring a distasteful breakfast if the potential entrant (the Receiver, R) can be deterred from entering the industry. More specifically, having the preferred breakfast is worth 1 to the incumbent, but avoiding a duel with the entrant is worth 2. The entrant's payoffs are independent of the incumbent's breakfast: the entrant prefers to duel with the wimp but not to duel with the surly incumbent. Finally, the prior probability that the incumbent is surly is 0.9.



There are two kinds of sequential equilibria here. In the first kind, both types of incumbent have beer for breakfast, and the entrant duels if quiche is observed but declines to duel if beer is observed. In such an equilibrium, the decision to duel following quiche is rationalized by any off-the-equilibrium-path belief that

puts sufficiently high probability (at least 1/2) on the incumbent being wimpy. [Note: If Pr(wimp) > 1/2, then the only equilibrium is {beer, beer; dual, dual}.]

In the second kind of equilibrium, both types of incumbent have quiche for breakfast, and the entrant duels if beer is observed but declines to duel if quiche is observed. Again, the beliefs that support the decision to duel are those that attach high probability to the wimp. But here such beliefs seem unnatural: the prior belief is .9 that the incumbent is surly, but when conditioned on the observation of beer—which is preferred if surly but not if wimpy—the posterior belief is at least .5 that the incumbent is wimpy.

This second kind of equilibrium is susceptible to a speech of the form outlined above. Rather than repeat the speech, we will argue in terms of the game tree. The objectionable equilibrium path is marked with arrowheads in the figure above. Note that in equilibrium the wimp gets 3 and could get at most 2 from deviating. Thus, the first supposition is met. The second supposition considers the payoff to the surly incumbent *if* the entrant concludes that only the surly type would deviate from the equilibrium by having beer for breakfast. Note that if the entrant concludes that the beer-drinker is surly, then declining to duel is the optimal decision. This leads to a payoff of 3 for the surly incumbent, which is better than the 2 earned in equilibrium. Thus, the second supposition also is met, so Cho & Kreps reject the second kind of equilibrium.

Formalizing the "Intuitive Criterion"

Cho & Kreps use these two suppositions to define an "intuitive criterion" for refining the set of sequential equilibria. Equilibria that do not satisfy the criterion will be rejected. Stating the criterion formally requires some notation.

After hearing $m \in M$, R's beliefs are $\mu(t|m)$. Sequential rationality requires that R's subsequent action a(m) maximize the expectation of $U^R(t,m,a)$ with respect to these beliefs. Define the set of such best responses as

$$BR(\mu, m) \equiv \underset{a \in A}{\operatorname{argmax}} \sum_{t \in T} \mu(t|m) U^R(t, m, a).$$

Then R's (behavior) strategy $p^R(a|m)$ is greater than zero only if $a \in BR(\mu,m)$. For subsets I of T, let BR(I,m) denote the set of best responses for R to beliefs concentrated on I:

$$BR(I, m) \equiv \bigcup_{\{\mu:\mu(I)=1\}} BR(\mu, m).$$

Given the equilibrium strategies $p = \{p^S(m|t), p^R(a|m)\}$, the equilibrium payoff to an S of type t is

$$U^*(t) \equiv \sum_{a \in A} \sum_{m \in M} p^R(a|m) p^S(m|t) U^S(t, m, a).$$

Cho & Kreps say that an equilibrium *fails* to satisfy the intuitive criterion if there exist

(a) an unsent message $m' \in M$ (i.e., $p^S(m'|t)=0$ for all $t \in T$),

(b) a subset J of T, and

(c) a type $t' \in T\sim J$

such that

(1) for all $t \in J$, for all $a \in BR(T,m')$, $U^*(t) > U^S(t,m',a)$, and

(2) for all $a \in BR(T\sim J,m')$, $U^*(t') < U^S(t',m',a)$.

In words, (1) says that types $t \in J$ are better off receiving their equilibrium payoff $U^*(t)$ than they conceivable could be from deviating- - no matter what inference R draws from m'. In (2), however, R uses (1) to conclude that no type $t \in J$ would send m' so $\mu(t|m')$ should be concentrated on T~J. Finally, S cannot be sure which belief concentrated on T~J will be held by R, but type t' doesn't care: t' is better off than in equilibrium no matter which of these beliefs R holds.

A quick check confirms that (1) and (2) formalize the two suppositions that led to the speech described above. One can imagine such speeches to be implicit: R is smart enough to look at a proposed equilibrium and discern that S would make such a speech if communication were allowed.

Cho & Kreps proceed to apply this intuitive criterion to the plethora of equilibria in Spence's well known model of signaling in job markets.

Spence's Signaling Model

In this model, a worker privately observes her productive ability and then chooses an amount of education to acquire. The market observes the worker's education choice and then offers a wage.

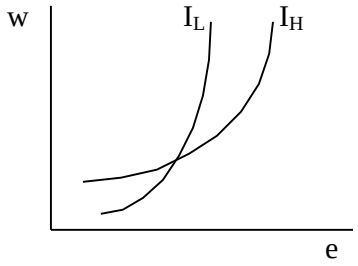
Since the variables (t,m,a) are not particularly suggestive in this context, let the worker's choice of education be $e \in [0,\infty)$, and let the market's wage offer be $w \in [0,\infty)$. (While this formulation violates the earlier assumption that M and A are finite, that assumption is much more important for relating the intuitive criterion to the Kohlberg-Mertens paper than it is for understanding the criterion for its own sake, so we dispense with it here.)

Let the worker's preferences be additively separable:

$$U^S(t,e,w) = w - c(t,e)$$

for $t \in \{H,L\}$, where c(t,e) is the (psychic) cost for worker t of acquiring eduction e. A crucial assumption is that the low-ability worker has higher marginal cost of education than does the high-ability worker. In terms of indifference curves in education $\times$ wage space, this means that the low-ability worker's curve (I_L) is everywhere steeper than that of the high-ability worker (I_H). It is easy to show that the Single Crossing Property implies that there cannot be pooling with just two types, but with more than two types D1 is

needed to get uniqueness.



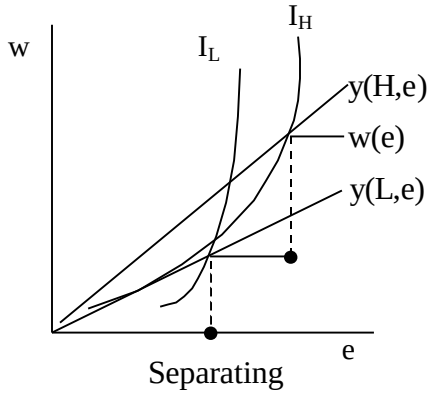
On the production side, let a worker of ability t and education e produce output y(t,e), where

$y(H,e) > y(L,e)$ for all e and

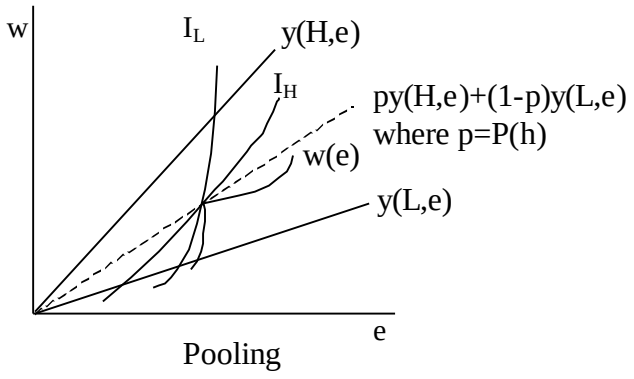
$\partial y(t,e)/\partial e > 0$ for all t,e.

Spence (1973) argues that competition among firms will drive profits to zero. In terms of a Bayesian Nash equilibrium, this means that given a conjecture e(t) about the worker's education choice, the market wage will satisfy w(e) = y(t,e). Cho & Kreps model this zero-profit assumption as a Bertrand game between two firms: both firms offer wages equal to the worker's expected productivity, conditioned on the (common) conjecture e(t).

Spence shows that there are *lots* of equilibria in this game. They can be organized into three categories: separating (in which the two types choose different levels of education), pooling (in which they choose the same education), and hybrid (in which at least one type randomizes between pooling with the other type and distinguishing itself). Examples of separating and pooling equilibria are displayed in the two figures below; from these it is straightforward to describe a hybrid equilibrium.



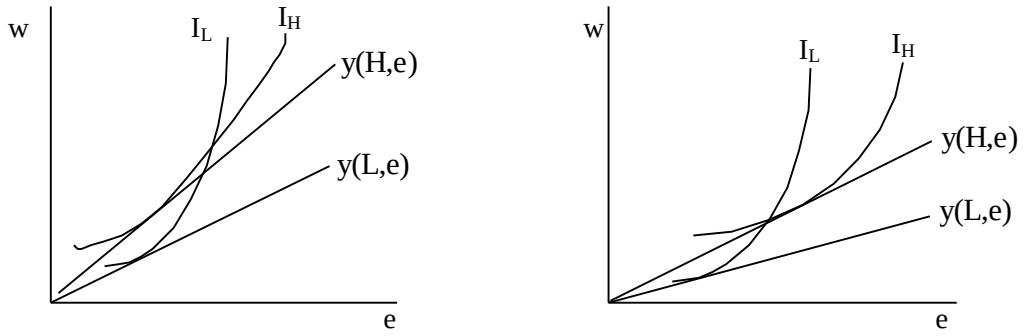
Separating



Pooling

(An aside: In what follows it will become clear that two cases can usefully be distinguished here: either the low-ability worker envies the high-ability worker's full-information education choice, or she does not. The

two are drawn below.



We will consider only the "envy" case, because it is more interesting and arguable more natural. The arguments for the "no envy" case are analogous to those presented below—in fact, they are somewhat simpler.)

Cho & Kreps test this plethora of equilibria with the intuitive criterion. (actually, this is like shooting a mouse with an elephant gun: much weaker tests will rule out many of the equilibria.) The argument proceeds in three steps. First, some of the Nash equilibria (including both of those drawn above can be rejected because they are not sequential equilibria. Second, some sequential equilibria can be rejected because they do not remain equilibria after weakly dominated strategies have been eliminated. And finally, other sequential equilibria can be rejected because they do not conform to the intuitive criterion.

Strikingly, exactly *one* equilibrium survives this process. Moreover, it is *the* equilibrium that was singled out from the crowd as being most reasonable a decade before the intuitive criterion existed. (Although there now exist other arguments against this equilibrium, or against this modeling of Spence's idea.)

To begin the first step of the argument, recall that in a sequential equilibrium the market must have beliefs $\mu(t|e)$ following any signal e. Since the wage is then the expected productivity given these beliefs,

$$y(L,e) \le w(e) \le y(H,e)$$

for each e. Therefore, the wage schedules in the separating and pooling equilibria drawn above are Nash but not sequential.

The rest of the argument refers to Figures 1-4 below.

Figure 1

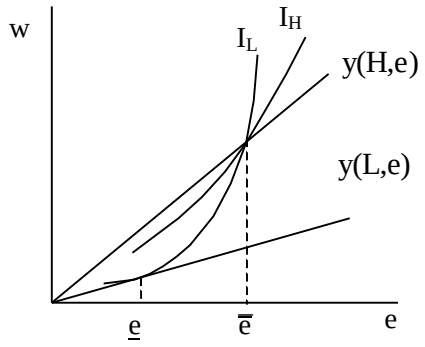
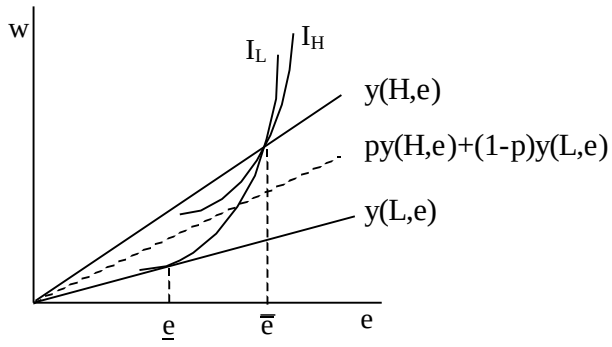


Figure 2



Suppose worker L separates with positive probability. Then (see Figure 1)

(a) it happens at the tangency of y(L,e) and I_L, hereafter$\underline{e}$ (because $w(\underline{e}) \geq y(L,\underline{e}) \Rightarrow w(\underline{e}) - c(L,\underline{e}) > y(L,e) - c(L,e)$ for all $e \neq \underline{e}$;

(b) L accepts no utility less than $y(L,\underline{e}) - c(L,\underline{e})$ by individual rationality (an implication of the equilibrium); and

(c) any hybrid equilibrium must do its pooling on the indifference curve I_L through $(\underline{e},y(L,\underline{e}))$.

Now the second step. Observe that (b) above holds for any equilibrium, whether or not L separates. Therefore, in any equilibrium

(d) education levels above $\bar{e}$ (determined by the intersection of the productivity curve y(H,e) and the indifference curve I_L through $(\underline{e},y(L,\underline{e}))$ are weakly dominated for L;

(e) market beliefs $\mu(t|e)$ for $e > \bar{e}$ in the pruned game tree must be degenerate on H;

(f) wages must be $w(e) = y(H,e)$ for $e > \bar{e}$; and

(g) H accepts no utility less than $y(H, \bar{e}) - c(H, \bar{e})$.

Returning to the assumption that L separates with positive probability yields

(h) the only possible hybrid is at $(\bar{e}, y(H, \bar{e}))$, but this wage earns negative profits unless the probability that L accepts is zero; and

(i) there are no hybrid equilibria in which L separates with probability less than one.

This establishes that there is a unique equilibrium in which L separates with positive probability. In it, both types separate with probability one, L at $(\underline{e},y(L,\underline{e}))$ and H at $(\bar{e}, y(H, \bar{e}))$.

Any alternative equilibrium must have L separating with probability zero. As it happens, no such equilibria conform to the intuitive criterion. There are two cases: pooling and hybrid equilibria. In such equilibria,

(*) $\qquad w(e) \leq p(H)y(H,e) + p(L)y(L,e),$

with equality for pooling equilibria and inequality for hybrids. If this prevents H from achieving the utility $y(H, \bar{e}) - c(H, \bar{e})$, then by (g) above these kinds of equilibria do not exist. (See Figure 2.) And if H can achieve the requisite utility then such equilibria exist but are rejected by the intuitive criterion, as follows. (See Figures 3 and 4.)

Figure 3

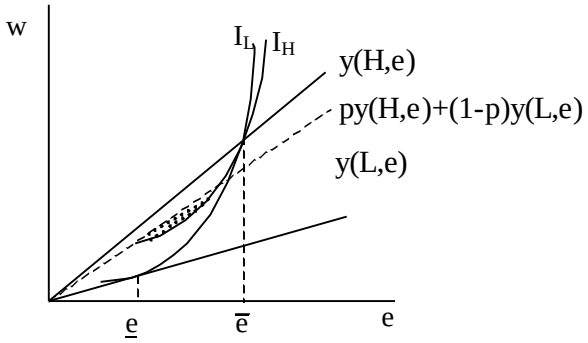
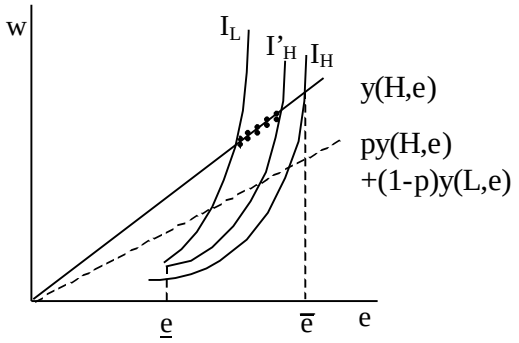


Figure 4



Pick a point satisfying (*) and H's utility constraint. Consider the indifference curves I_L and I_H through this point. By assumption, I_L is steeper, so the intersection of I_L and y(H,e) is to the left of the intersection of I_H and y(H,e). Any education level between these two points of intersection is an unsent message that fulfills the requirements of the intuitive criterion: the market should infer that the worker is H because such signals are worse than the equilibrium payoff for L, but if it is sure to be H then the wage must be w(e)=y(H,e), which makes H better off than in the equilibrium.

2. Forward Induction Equilibrium (Cho, 1987)

Recall the signaling game of Cho & Kreps (1987). The timing is:

1. nature draws a type $t \in T$ for the Sender, S;
2. S learns t and sends a message $m \in M$ to the Receiver, R; and
3. R observes m and takes an action $a \in A$.

The payoffs are $U^S(t,m,a)$ and $U^R(t,m,a)$. Everything is common knowledge except nature's choice of t for S.

In this setting, Cho & Kreps' intuitive criterion argued that we should reject any sequential equilibrium satisfying the following conditions: there exists an unsent message m' and a subset of types J such that

(1) for all $t \in J$, for all $a \in BR(T,m')$, $U^*(t) > U^S(t,m',a)$, and

(2) there exists $t' \in T \sim J$ such that for all $a \in BR(T \sim J,m')$, $U^*(t') < U^S(t',m',a)$,

where $U^*(t)$ is t's expected payoff in the equilibrium under consideration.

The informal argument for rejecting such equilibria has two steps. First, condition (1) suggests that R's

belief $\mu(t|m')$ should put no probability on types $t \in J$: reasonable μ's should be concentrated on T~J. And second, if there is a type t' satisfying (2), then surely this type would deviate from the proposed equilibrium, since t' is better off deviating no matter what reasonable belief R will hold.

One could imagine a weaker criterion for rejecting equilibria (i.e., one that would reject more equilibria). Following Cho (1987), let

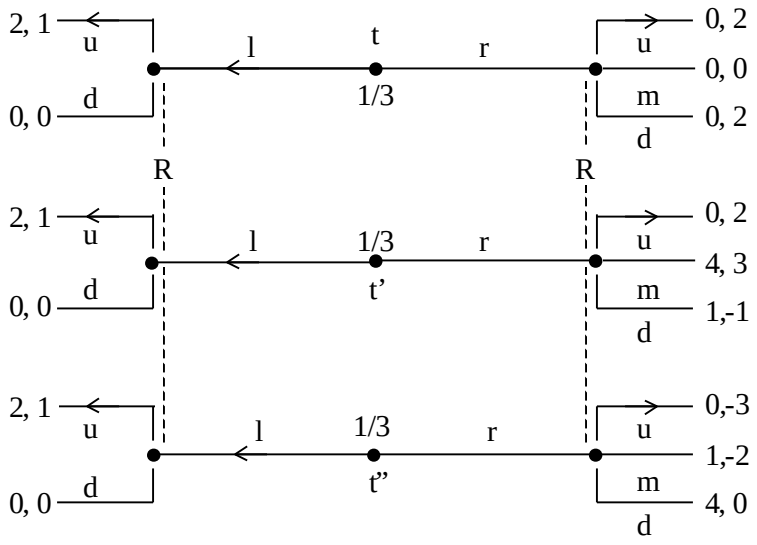
$J(m'|\rho) \equiv \{t \in T \mid U^*(t) > U^S(t,m',a) \text{ for all } a \in BR(T,m')\}$,

where ρ is the sequential-equilibrium strategy in question. This is the largest set J that satisfies (1) above. Reasonable beliefs following the deviation m' are then those that assign zero probability to $t \in J(m'|\rho)$:

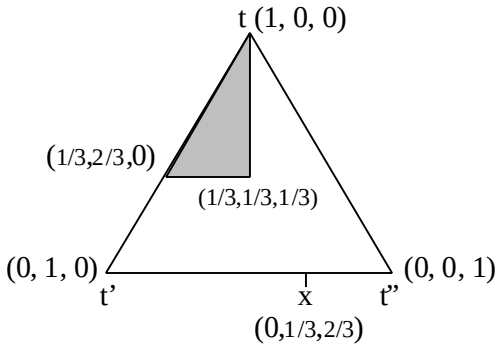
$\mu(J(m'|\rho) \mid m') = 0$,

provided $J(m'|\rho)$ is a proper subset of T. Cho says that such beliefs satisfy *introspective consistency*. Further, a sequential equilibrium is a *forward induction equilibrium* if it is supported by beliefs satisfying introspective consistency.

To understand these definitions, consider the following example. It shows that there are sequential equilibria that are not rejected by Cho & Kreps' intuitive criterion but that fail to be forward induction equilibria.



A pooling sequential equilibrium is described using arrowheads in the figure: all S-types play l, and R plays u both on and off the equilibrium path. R's choice off the equilibrium path is rationalized by beliefs in the shaded portion of the simplex below.



To apply the intuitive criterion to this equilibrium, let J={t} and m'=r. Then (1) holds, but neither t' nor t" satisfies (2). On the other hand, the beliefs that support the equilibrium do not satisfy introspective consistency: beliefs on the t'-t" axis to the left of x cause R to play m, while beliefs to the right cause d; there is no belief over {t',t"} that rationalizes u for R.

In fact, forward induction equilibria are a subset of the sequential equilibria that conform to the intuitive criterion. To show that this inclusion relationship holds, we turn now to the *extended* intuitive criterion, which rejects an equilibrium if and only if it is not a forward induction equilibrium.

The extended intuitive criterion consists of condition (1) above and a new condition (2') that replaces (2).

(2') for all $a \in BR(T\sim J,m')$ there exists $t' \in T\sim J$ such that $U^*(t') < U^S(t',m',a)$.

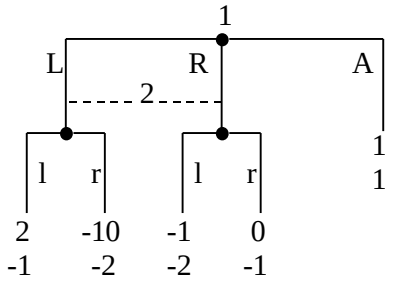
As before, a sequential equilibrium is to be rejected if it satisfies condition (1) and (2'). The extended intuitive criterion rejects more equilibria than does its counterpart because condition (2') is easier to meet than is (2). (This is so because the only difference between the two is the order of the quantifiers. A type t' that satisfies (2) will satisfy (2'), but the reverse is not true because (2') allows a different t' to profit from deviating for each best response in BR(T~J,m').)

Proposition. A strategy profile p is a forward induction equilibrium if and only if p is a sequential equilibrium satisfying the extended intuitive criterion.

The proof is left as an exercise. It remains to explain the relation of Cho's work to the notion of forward induction discussed in the notes on sequential equilibria.

Much of Cho's paper concerns forward induction equilibria in general (finite) extensive-form games, rather than the special class of signaling games. This involves generalizing the definition of $J(m'|p)$, the set of types that have no conceivable reason to deviate from p by sending m'. For a general game, a deviation from a specified equilibrium is said to be "bad" if it *always* yields the deviator less than her equilibrium payoff in *every* circumstance.

This restriction on beliefs off the equilibrium path has the desired effect in the simple example discussed by Kreps and Wilson. In this game (drawn below), the sequential equilibrium (A,r) seems unreasonable because it requires player 2 to believe with high probability that player 1 has made a ridiculous deviation from the equilibrium. More precisely, R would be a bad deviation for 1. Thus, (A,r) is not a forward induction equilibrium because it can be supported only by beliefs that assess positive probability that a bad deviation has occurred.



Unfortunately, in more complex games the set of bad deviations often is empty, in which case every sequential equilibrium is a forward induction equilibrium, and we must resort once again to ad hoc arguments to capture forward induction.

3. Divinity, D1, D2, Universal Divinity

Consider again the sender-receiver game: S learns her type $t \in T$ and sends a message $m(t) \in M$ to R, who takes an action $a(m) \in A$. The question is: "What should R infer from the message m? The intuitive criterion and the forward induction equilibrium are based on the following dominance argument.

Dominance: Eliminate t if m is sent and m is dominated by m' for t:

$$\min_a U(t, m', a) > \max_a U(t, m, a).$$

This is too weak a requirement. At the very least, we should require that R's action a is a best response for some beliefs: $a \in BR(T,m)$. This known as equilibrium dominance.

Stronger notions of dominance can be constructed as follows. Fix an equilibrium with payoff to S of $U^*(t)$. For each (t,m) find the set of best responses by R that cause S to defect. Define D_t to be the set of best responses by R that make S strictly prefer defection:

$D_t = \{a \in BR(T(m),m): U^*(t) < U(t,m,a)\}$.

And define D°_t to be the set of best responses by R that make S indifferent between defection and the equilibrium:

$D^\circ_t = \{a \in BR(T(m),m): U^*(t) = U(t,m,a)\}$.

The size of the set D_t relative to $D_{t'}$ in some sense measures how likely it is that type t benefits from the deviation relative to type t'. Our next set of refinements are all based on the sets D_t and $D°_t$.

The "D1" refinement requires that zero weight be put on the type t if m is sent if there is another type t' such that t' always strictly benefits from the deviation whenever t benefits from the deviation:

D1. If $\exists$ t' with $D_t \cup D°_t \subseteq D_{t'}$ then prune (t,m).

This refinement works well (i.e., identifies a unique equilibrium) in an important class of signalling games. Cho and Sobel (1988) demonstrate that, for monotonic signaling games, the set of D1 equilibria is the same as the set of stable equilibria. Moreover, if the single crossing property is satisfied, then D1 yields a unique equilibrium.

The "D2" refinement goes further (perhaps too far) by requiring that zero weight be put on t when m is sent if for every best response that causes t to deviate there is a t' that strictly benefits from the deviation:

D2. If $D_t \cup D°_t \subseteq \cup_{t' \neq t} D_{t'}$ then prune (t,m).

Banks and Sobel (*Econometrica*, 1988) define two other refinements based on D1 and D2. The first, *divinity*, is a weakening of D1. Rather than put zero weight on types t satisfying D1, divinity simply requires that the posterior belief after m is sent cannot increase the likelihood ratio of t to t'. The second, *universal divinity*, is a strengthening of D2. It requires that t be eliminated, using an iterative application of D2. With universal divinity the updated beliefs do not depend on the prior; whereas, divine beliefs do depend on the prior.

The last refinement is a slight strengthening of D2.

Never Weak Best Response: Prune (t,m) if $D°_t \subseteq \cup_{t' \neq t} D_{t'}$.

t is given zero weight if whenever t is indifferent between deviating and following the equilibrium, there is a t' that strictly benefits from the deviation.

Which refinement do we need in the Spence signaling game to get a unique equilibrium? As we saw if there are just two possible types, then the intuitive criterion is enough to guarantee that the separating equilibrium is unique. The high type is able to separate from the low type by adopting a level of education that is a bad deviation for the low type. But with more than two types pooling can occur. For example, with three types it is possible for the low and medium types to pool together (the high type can separate by adopting a level of education that would be a bad deviation for the two lower types). The pooling cannot be eliminated by the intuitive criterion, since no deviation that the medium type prefers to the pooling equilibrium is bad for the low type because the low type can think that the firm will infer that it is a high type as a result of the deviation. Applying D1, however, gives us a unique equilibrium. For any level of eduction that the medium type prefers to the pooling equilibrium, the set of best response wages that make

the medium type want to deviate is larger than the set of best response wages that make the low type want to deviate, we must therefore put all the weight on the medium type by D1. D1 in fact gives us uniqueness, even with a continuum of types.

Ordering of Refinements

It is possible to order the equilibrium refinements in terms of the set of equilibria they produce. For general games, we have

$NE \supset SE \supset PE \supset ProperE \supset IC \supset EIC.$

For signalling games, we have

$EIC \supset Div \supset D1 \supset D2 \supset UniDiv \supset NWBR \supset Stable \neq \varnothing.$

B. Cheap Talk

1. Strategic Information Transmission (Crawford and Sobel, 1982)

This paper presents a formal analysis of the informal advice that "talk is cheap" and "actions speak louder than words." The main result is that even if "talk" is the only available mode of communication, some information can be credibly transmitted. How much information is communicated depends on how similar the parties' preferences are.

There are two parties, a Sender (S) and a Receiver (R) of information. The timing is as follows:

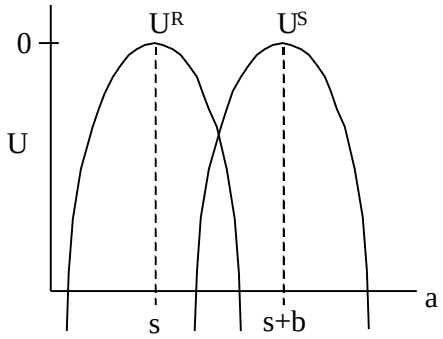
1) S privately observes the state of the world, $s \in [0,1]$;
2) S sends a message $m \in M$ to R; and
3) R takes an action $a \in (-\infty,\infty)$.

R does not observe the state s, but holds the prior belief that s has distribution F(s) on [0,1].

The *payoffs* are $U^S(a,s,b)$ and $U^R(a,s)$, where b measures how nearly the agents' interests coincide. A convenient example is

$U^S(a,s,b) = -[a - (s+b)]^2$ and $U^R(a,s) = -[a - s]^2$,

as shown in the figure below, which is drawn for fixed s.



Notice that the signal $m \in M$ is irrelevant to the payoff functions. This is the sense in which "talk is cheap" here, as distinctly opposed to Spence's signaling model in which education is costly.

Formally, the assumptions on the payoffs are: for i = R,S

(i) $U^i_1 = 0$ for some a,

(ii) $U^i_{11} < 0$, and

(iii) $U^i_{12} > 0$.

These assumptions imply that given s there is a unique action, a, that maximizes U^i. Moreover, these optimal actions (denoted $a^R(s)$ and $a^S(s,b)$, respectively) are continuous and strictly increasing functions of s.

In sum, the Sender and Receiver play the Bayesian game

$$\Gamma = \{A_S, A_R; T_S, T_R; p_S, p_R; U_S, U_R\},$$

where $A_S = M$, $A_R = (-\infty,\infty)$, $T_S = [0,1]$, $T_R = \{0\}$, $p_S = 1$ (no uncertainty), $p_R = f(s)$ (the density of F), and U^S and U^R are as defined above.

Recall that a strategy is a plan of action covering every contingency that might arise. For S, about whom R has incomplete information, a strategy is a function from types to actions (or mixtures over actions): let q(m|s) be the density of S's choice of m when the state is s. For R, whose type is known, a strategy must specify an action a(m) for each signal $m \in M$ that R might receive.

Note well that even though S's message necessarily precedes R's action in time, S and R choose strategies simultaneously: after observing a message m, R's optimal action depends on updated beliefs about the state s, p(s|m); to compute these beliefs R needs a conjecture about S's behavior for each of S's types, not just the one S knows arose. Specifically, if R conjectures that S chooses m according to the density q(m|s) when the state is s, then Bayes' rule yields

$$p(s|m) = \frac{q(m|s)f(s)}{\int_0^1 q(m|t)f(t)dt}.$$

When S announces the message $m \in M$, S is *not* announcing a strategy, so this is not the same as one player

moving first as in a Stackelberg game. For S to be moving first as in a Stackelberg game, R would have to observe S's entire signaling rule, q(m|s). Another perspective on the timing of this game emerges from a comparison of this game and the standard principal-agent model. Unlike the agency model, here R cannot commit to an action rule and communicate it to S before S moves.

The strategies {q(m|s),a(m)} form a Bayesian equilibrium if the usual Nash conditions hold:

(1) for each $s \in [0,1]$, $\int_M q(m|s)dm = 1$ and if $m^* \in M$ is in the support $\int$ $q(\cdot|s)$ then m^* solves $\max_{m \in M} U^S(a(m), s, b)$; and

(2) for each $m \in M$, a(m) solves $\max_a \int_0^1 U^R(a,s)p(s|m)ds$,

where p(s|m) is given by Bayes' rule as described above.

Note that because U^R is concave in the action a, it is not necessary to consider mixed strategies for R.

Proving the existence of and characterizing the Bayesian equilibria of Γ requires some new notation. Given an action rule a(m) and an arbitrary action $\bar{a} \in (-\infty, \infty)$, define the (possibly empty) set $\overline{M} \equiv \{m \in M : a(m) = \bar{a}\}$. Say that an S-type $\bar{s}$ induces the action $\bar{a}$ in the equilibrium {q(m|s),a(m)} if

$$\int_{\overline{M}} q(m|\bar{s})dm > 0.$$

Finally, recall that $a^S(s,b)$ and $a^R(s)$ maximize U^S and U^R, respectively, and that these are continuous and increasing functions of s.

Lemma: Suppose that b is such that no s Î [0,1] satisfies $a^S(s,b) = a^R(s)$. Then there exists $\varepsilon > 0$ such that if u and v are actions induced in equilibrium then $|u - v| \geq \varepsilon$. Further, the set of actions induced in equilibrium is finite.

Proof: Take u < v. Let s_u and s_v be S-types that induce u and v, respectively. Then revealed preference yields

$U^S(u,s_u,b) \geq U^S(v,s_u,b)$ and $U^S(v,s_v,b) \geq U^S(u,s_v,b)$.

By the continuity of U^S in s there existss such that $U^S(u,s,b) = U^S(v,s,b)$. Therefore, by the concavity of U^S in a,

(i) $u < a^S(s,b) < v$.

Because $U^S_{12} > 0$, any S-type $s > (<) \bar{s}$ strictly prefers v(u) to u(v) and would choose to induce v(u) rather than u(v) in equilibrium. That is,

(ii) u is not induced by any $s > \bar{s}$, and

(iii) v is not induced by any $s < \bar{s}$.

These last two statements, together with the assumption that $U^R_{12} > 0$, imply

(iv) $u \le a^R(\bar{s}) \le v$.

This is because: $a^R(s)$ is increasing; $a^R(\bar{s})$ would be R's action if S's type were certain to be $\bar{s}$; and u(v) signals to R that $s <(>) \bar{s}$. So, for instance,

$$\max_a E_s[U^R(a,s)|s < \bar{s}] \le a^R(\bar{s}).$$

Since $a^R(s)$ and $a^S(s,b)$ are continuous functions of s, $|a^R(s) - a^S(s,b)|$ attains its minimum over $s \in [0,1]$. By hypothesis, this minimum is positive. Therefore, there exists ε such that $v - u \ge |a^R(\bar{s}) - a^S(\bar{s},b)| \ge \varepsilon > 0$, where the first inequality follows from (i) and (iv).

Finally, since $a^R(s)$ is continuous and increasing, the set A of actions induced in equilibrium is bounded by $a^R(0)$ and $a^R(1)$, both of which must be finite. Therefore A is a finite set. ¦

Note that the Lemma applies to any message space M. The Lemma shows that in cheap-talk games with imperfectly aligned preferences there cannot be a separating equilibrium: unlike in Spence's signaling model, equilibrium communication is necessarily imperfect. This happens because there are no exogenous signaling costs. The only costs (or benefits) from signaling arise endogenously because different signals induce different actions.

The Theorem below characterizes these imperfect-communication equilibria in terms of some new notation. Let $s(M) \equiv (s_0(M),..., s_M(M))$ denote a partition of [0, 1] into M steps, where $0= s_0(M)< s_1(M)<...< s_M(M) = 1$. Where there is no possibility of confusion, denote $s_i(M)$ simply by s_i. For $0 \le \underline{s}, \bar{s} \le 1$, define

$$\bar{a}(\underline{s}, \bar{s}) \equiv \begin{cases} \operatorname{argmax}_a \int_{\underline{s}}^{\bar{s}} U^R(a,s) f(s) ds & \text{if } \underline{s} < \bar{s}, \\ a^R(\underline{s}) & \text{if } \underline{s} = \bar{s}. \end{cases}$$

Theorem. Suppose b is such that no $s \in [0,1]$ satisfies $a^S(s,b) = a^R(s)$. Then there exists a positive integer M(b) such that for every integer $M \in [1,M(b)]$ there exists an equilibrium {q(m|s),a(m)}, where for all $i \in \{1,...,M-1\}$

(i) q(m|s) is uniform on $[s_i, s_{i+1}]$ when $s \in (s_i, s_{i+1}]$,

(ii) $U^S(a(s_i, s_{i+1}), s_i, b) = U^S(\bar{a}(s_{i-1}, s_i), s_i, b)$,

(iii) $a(m) = \bar{a}(s_i, s_{i+1})$ when $m_i \in (s_i, s_{i+1})$,

and $s_0 = 0$ and $s_M = 1$. Further, every equilibrium of Γ is economically equivalent to one in this class, in the sense that the two equilibria induce the same outcome map from $s \in [0,1]$ to $a \in (-\infty,\infty)$.

Note the use of M as both the signal set and the number of elements in a partition. This is intentional. The last sentence of the Theorem implies that S can implement an equilibrium with M steps using any signal set with at least M elements: the information content of the equilibrium can be translated into any

(sufficiently rich) language.

Sketch of Proof: Given (i), if R hears the message $m \in (s_i, s_{i+1})$, the posterior belief is simply

$$p(m|s) = \frac{f(s)}{\int_{s_i}^{s_{i+1}} f(t)dt}.$$

Therefore R's strategy specified in (iii) is a best response to S's strategy given in (i). As for S, (ii) guarantees that the S-type $s = s_i$ is indifferent between the actions $\bar{a}(s_i, s_{i+1})$ and $\bar{a}(s_{i-1}, s_i)$. Types $s >(<) s_i$ strictly prefer the latter (former) to the former (latter). Moreover, S-types $s \in (s_i, s_{i+1})$ strictly prefer $\bar{a}(s_i, s_{i+1})$ to any of the other actions $\bar{a}(s_j, s_{j+1})$ induced by (iii). That is, S's strategy is a best response to R's.

To complete the proof that equilibria of the form specified by (i)-(iii) exist, it remains to determine M(b) and show that a solution to the difference equation (ii) exists for each integer $M \in [1,M(b)]$. There are two important steps. First, because $\bar{a}(\underline{s}, \bar{s})$ is increasing in both arguments and U^S is concave, given $s_{i-1} < s_i$ there is at most one $s_{i+1} \in (s_i,1]$ that solves (ii). And second, because the solution s_{i+1} moves smoothly in the initial conditions s_i and s_{i-1}, solutions to the difference equation can be constructed iteratively. These steps are illustrated in the example below.

The last part of the proof shows that all other equilibria are economically equivalent to those described here. The Lemma shows that only a finite number of actions are induced in equilibrium. Let the set of these actions be $A = \{a_j\}_{j=1}^{J}$, where $a_j < a_{j+1}$ for all $j < J$. As shown in the proof of the Lemma, for each pair (a_j,a_{j+1}) there exists an S-type s_j satisfying

(*) $\quad U^S(a_j,s_j,b) = U^S(a_{j+1},s_j,b).$

The concavity of U^S implies that s_j strictly prefers either of a_j or a_{j+1} to any other $a_k \in A$, and that S-types $s \in (s_{j-1},s_j)$ strictly prefer a_j to any other $a_k \in A$, including a_{j+1}.

Given the conjecture a(m) about R's behavior, each S-type $s \in (s_{j-1},s_j)$ will send the signal $m_j \in M$ that induces the action a_j via $a_j = a(m_j)$. In equilibrium, R holds a correct conjecture about which S-types send m_j, so $a(m_j)$ must be R's best response to this belief, namely $\bar{a}(s_{j-1},s_j)$. Thus (*) is exactly (ii), and (i) and (iii) are rephrased in terms of the general signal space M as

(i) S sends m_j when $s \in (s_{j-1},s_j)$, and

(ii) $a(m_j) = \bar{a}(s_{j-1},s_j)$.

This completes the sketch of proof.∎

The determination of M(b) and the construction of solutions to the difference equation (ii) for integers $M \in [1,M(b)]$ are illustrated by the following example: $U^S(a,s,b) = -[a - (s+b)]^2$, $U^R(a,s) = -[a - s]^2$, and F(s) uniform on [0,1]. Then $\bar{a}(s_i, s_{i+1}) = (s_i + s_{i+1})/2$, so (ii) becomes

$$s_{i+1} = 2s_i - s_{i-1} + 4b.$$

It is straightforward to check that

$$s_i = i s_1 + 2i(i - 1)b$$

is a solution to this difference equation for any s_i. Substituting i = M and $s_M = 1$ yields

$$1 = M s_1 + 2M(M - 1)b.$$

Since $s_1 \in (0,1)$, M(b) is the largest integer satisfying $2M(M - 1)b < 1$. Some algebra shows that M(b) is the largest integer less than $[1 + (1+2/b)^{1/2}]/2$. Note that $M(b) \to \infty$ as $b \to 0$, but that M(b) = 1 for $b \geq 1/4$: at least in this example, more communication is possible when preferences are more similar.

To continue with the example, suppose b = 1/20. Then M(b) = 3. Simple calculations show that the two-step equilibrium is {0, 2/5, 1} and the three-step equilibrium is {0, 2/15, 7/15, 1}.

When there exist multiple equilibria in this way, the players have a problem knowing which equilibrium to play. If one of the equilibria yields expected payoffs that Pareto-dominate the payoffs associated with the other equilibria, then it would seem a natural choice. This does not happen here: R's expected utility is higher in equilibria with more steps, but for some S-types this is not so. Consider, for instance, s = 3/20 in the two-step equilibrium above. R's actiona (0, 2/5) = 1/5 is the best possible action for this S-type, because in this example $a^S(s,b) = s + b$, so $a^S(3/20, 1/20) = 1/5$. It is true, however, that ex ante (i.e., before S learns s), S's expected payoff is higher in equilibria with more steps, but this far from proves that the equilibrium with M(b) steps should be played once S has learned s.

2. Neologisms (Farrell, 1985)

Consider Cho's forward induction equilibria in signaling games where talk is cheap. Mathematically, this means that $U^S(t,m,a)$ and $U^R(t,m,a)$ are independent of m. In this case, the set of types for whom m' is a bad deviation from the equilibrium strategy p is empty, so communicational consistency has no cutting power. To see this, recall that

$$J(m'|p) \equiv \{t \in T \mid U^*(t) > U^S(t,m',a) \text{ for all } a \in BR(T,m')\}.$$

Since $U^R(t,m,a)$ is independent of m, BR(T,m') becomes BR(T). Among the actions in BR(T) is the action a(t) that type t induces in equilibrium by sending the message m(t). Since $U^S(t,m,a)$ is independent of m,

$$U^S(t,m',a(t)) = U^S(t,m(t),a(t)) \equiv U^*(t).$$

So $J(m'|p) = \varnothing$ for all m' and p. Therefore, in a signaling game where talk is cheap, every sequential equilibrium is a forward induction equilibrium.

On the other hand, Crawford and Sobel show that there typically are multiple sequential equilibria in cheap-talk games. This implies that efforts to refine this set of equilibria will require new techniques. One such is studied by Farrell (1985). None of the other refinements work for cheap-talk games.

Farrell studies "credible neologisms" in communication games where talk is cheap (i.e., free), such as the game analyzed by Crawford and Sobel. A *neologism* is a new word, usage, or phrase. More formally, it is an unsent message in a signaling game. Roughly speaking, a neologism is *credible* if those S-types that might send this unexpected message can make a persuasive speech to R, along the lines envisioned by Cho & Kreps. The game Farrell analyzes is nearly identical to the signaling game discussed by Cho & Kreps and by Cho. The timing is:

1. nature draws a type t from a finite set T for the Sender, S;
2. S learns t and sends a message $m \in M^*$ to the Receiver, R; and
3. R observes m and takes an action $a \in A$, where A is finite.

Because talk is cheap, the payoffs are $U^S(t,a)$ and $U^R(t,a)$, independent of m.

The important difference is that the message space M^* is not finite. Rather, Farrell says that M^* is "infinite but discrete—for instance, the set of all (arbitrarily long) utterances in English." Among the intended consequences of this definition is that messages from S to R of the form "my type is in a subset X of T" are *always* available as neologisms. (One might wonder what happens if one of these messages is sent in equilibrium. Farrell's intent is that the relevant neologism is then "my type is in X, and this is a neologism, not the equilibrium message you were prepared for." For cheap-talk games, it seems reasonable (but slippery) to define the Sender's strategy space as being infinite in this way.)

Like Cho & Kreps, Farrell proposes to refine the set of sequential equilibria in this signaling game by administering a test. In words, a sequential equilibrium is reasonable if and only if it is neologism-proof. Equivalently, a sequential equilibrium should be rejected if and only if there is a credible neologism S could sent to R. It remains to define a credible neologism.

Let X be a non-empty subset of T, and let $\mu(t|X)$ be the distribution over types $t \in X$ that results from restricting the prior distribution of types $\mu(t)$ to X:

$$\mu(t|X) = \begin{cases} \dfrac{\mu(t)}{\sum_{t \in X} \mu(t)} & \text{if } t \in X \\ 0 & \text{if } t \notin X. \end{cases}$$

Let $a^*(X)$ solve

$$\max_{a \in A} \sum_{t \in X} \mu(t|X) U^R(t,a),$$

and assume $a^*(X)$ is unique. (This uniqueness can be justified via concavity of U^R in the action a, as in Crawford and Sobel, or by observing that the space of payoff functions $U^R \in R^{|T||A|}$ such that $a^*(X)$ is not unique has Lebesgue measure zero.)

So if R's beliefs are $\mu(t|X)$ then $a^*(X)$ will follow. S's payoff would then be $V(X,t) \equiv U^S(a^*(X),t)$. We now want to compare this payoff to t's equilibrium payoff. Let R's behavioral strategy be $p^R(a|m)$, which specifies a distribution over $a \in A$ for each $m \in M^*$ that might be observed. Then t's best response m(t) yields the payoff

$$\max_{m \in M^*} \sum_{a \in A} p^R(a|m) \cdot U^S(a,t) \equiv U^*(t)$$

(Farrell points out that in an equilibrium it is not a problem that M^* is not compact: in an equilibrium, S must play a best response to R's strategy, so a best response must exist, so sup and max are equivalent.)

Define the set $K(X|p)$ as, loosely speaking, the complement of Cho's set $J(m'|p)$:

$$K(X|p) \equiv \{t \in T \mid U^*(t) < V(X,t)\},$$

where as

$$J(m'|p) \equiv \{t \in T \mid U^*(t) > U^S(t,m',a) \text{ for all } a \in BR(T,m')\}.$$

So $K(X|p)$ is the set of types who would deviate from the equilibrium p if in so doing they led R to hold the belief $p(t|X)$. $J(m'|p)$, in contrast, is the set of types who would *not* deviate from p by sending m', no matter what belief this induced R to hold.

Farrell says that a subset X of T is *self-signaling* given the equilibrium p if $K(X|p) = X$, and that the neologism (unsent message) "t is in X" is *credible* if X is self-signaling. The reasoning behind this definition parallels the speech that Cho & Kreps envision S making to R when an equilibrium fails to satisfy the intuitive criterion. Here if $t \in X$ then S says (or R reasons):

> "My type is in X. Moreover, every other type in X and no type outside X has an incentive to make this speech. For if you believe it then your belief should be $\mu(t|X)$, so your action should by $a^*(X)$, which only we types in X would prefer to our equilibrium payoff."

Given a sequential equilibrium p, if there exists a credible neologism then Farrell rejects the equilibrium. If, on the other hand, there does not exist a credible neologism then Farrell says that p is *neologism-proof* and accepts it. This definition is analogous to the definition of *perfect sequential equilibria* presented in Grossman and Perry [1986].

One problem with this definition is that there may not exist a neologism-proof equilibrium, as the following example shows. Let $T = \{t_1,t_2\}$, $\mu(t_1) = \mu(t_2) = 1/2$, $A = \{a_1,a_2,a_3\}$, and let the payoffs be as given below.

U^S

	t_1	t_2
a_1	2	-1
a_2	-1	-2
a_3	0	0

U^R

	t_1	t_2
a_1	3	0
a_2	0	3
a_3	2	2

In any pooling equilibrium, R will play a_3. Farrell states that this is the only sequential equilibrium outcome (i.e., there are no separating or hybrid equilibria). By assumption, there always exists an unsent message, so equilibrium outcomes are all that matter—equilibrium strategies are irrelevant.

Given the pooling equilibrium outcome, the set $X = \{t_1\}$ is self-signaling: if R believes the neologism "$t \in X$" then a_1 will replace a_3 as a best response; this yields a payoff of 2 for t_1 (-1 for t_2), which is better (worse) than the equilibrium payoff of 0. Thus, Farrell rejects all the sequential equilibria in this game. Grossman and Perry [1986] also get nonexistence in a bargaining game for a range of parameter values.

When Farrell or Grossman and Perry is applied to the Crawford and Sobel model, the equilibrium with the finest partition is picked out.

3. Perfect Sequential Equilibria

Grossman and Perry (*Journal of Economic Theory*, 1986)

Grossman and Perry propose a refinement for signaling games that is similar to Farrell's "neologism proof" concept for cheap talk games. One difficulty with the refinements based on equilibrium dominance is that the beliefs following a deviation do not rationalize the deviation in an equilibrium sense. Grossman and Perry take the view that once a deviation has occurred, the other should try to rationalize the deviation by trying to find a set of types $K \subseteq T$ that benefit from the deviation if it is thought K deviated, but $t \notin K$ lose from the deviation. If such a K exists, then the beliefs following the deviation should require that the receiver infer that K deviated.

Their refinement, perfect sequential equilibrium (PSE), can be motivated from the Nash equilibrium and sequential equilibrium concepts:

NE requires best responses along the equilibrium path.

SE requires best responses at every information set given beliefs.

PSE requires best responses at every information set *for all* beliefs.

NE supports too many equilibria, because a player can threaten to take (non-credible) actions that hurt the other if the equilibrium is not followed. SE, by requiring best responses at every information set, does not allow a player to threaten with actions, but a player can threaten with beliefs: "If you do not follow the

equilibrium, I will infer that you are the type of player I enjoy beating up." PSE, by requiring best responses at every information set for all beliefs, attempts to limit a player's ability to threaten with beliefs. Inferences following a deviation are required to be one that rationalizes the deviation if such an inference exists.

To define a PSE we must extend the definition of a strategy (a prescribed action to take given each history $s^R(m) \in A$). A *metastrategy* is an action to take at each information set and all beliefs ($s^R(m,\mu) \in A$). An *updating rule* maps the message m and prior p into a posterior belief $\mu = g(m,p)$. The heart of the PSE refinement is to place a restriction on the updating rule.

Definition. A strategy profile and updating rule (s,g) is a *PSE* if $\forall$ information sets and $\forall$ beliefs, s is a best response and g is *credible*.

Definition. The updating rule g is *credible* if:

(a) the support of the posterior is contained in the support of the prior,

(b) if $\exists$ K such that $U^S(t,m,s^R(m,p_K)) \geq U^*(t)\ \forall\ t \in K$

$U^S(t,m,s^R(m,p_K)) \leq U^*(t)\ \forall\ t \notin K$

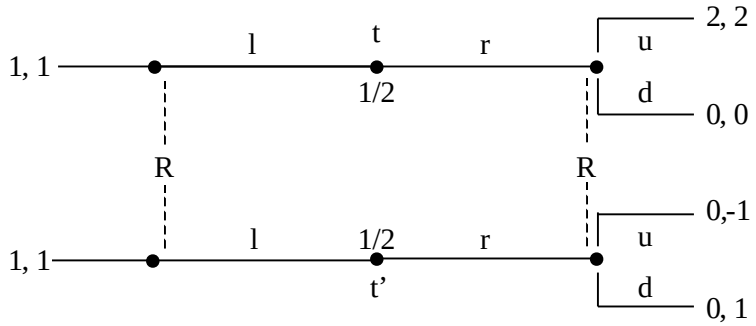
then $g(m,p) = p_K = p(t)/(\Sigma_{t\in K} p(t))$.

(c) use Bayes rule when possible.

Part (b) is the essence of the refinement. In words it says that if there is a set of types K that benefit from deviating if it is thought that K deviated, and those not in K prefer not to deviate, then the receiver must infer that the deviation came from the set K. A credible updating rule attempts to rationalize the deviation: Does $\exists$ K such that K benefits from deviating and the others do not?

We now present a number of examples that serve to illustrate the PSE idea.

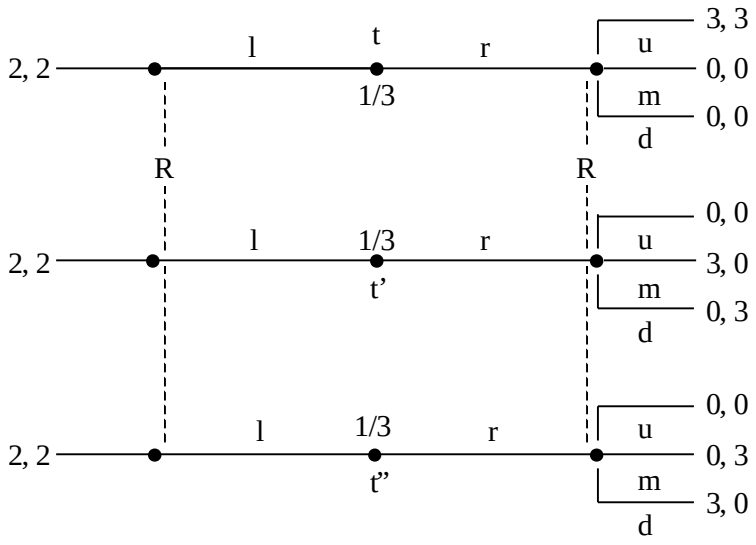
Example 1.



Here there are two SE: (1) ll,d, and (2) rl,u. Only (2) is a PSE. (1) requires that R put sufficiently high weight on t' if r is played, so that d is a best response. But K={t} rationalizes the deviation r, since t benefits if thought to be t by playing r (R responds with u yielding 2 rather than the equilibrium payoff of 1) and t'

would not want to deviate if thought to be t' (t' gets 0 from the deviation vs. 1 from the equilibrium). Notice that (2) is the only equilibrium satisfying the intuitive criterion as well, since r is a bad deviation for t': r yields 0 for t' regardless of R's response vs. 1 in the equilibrium.

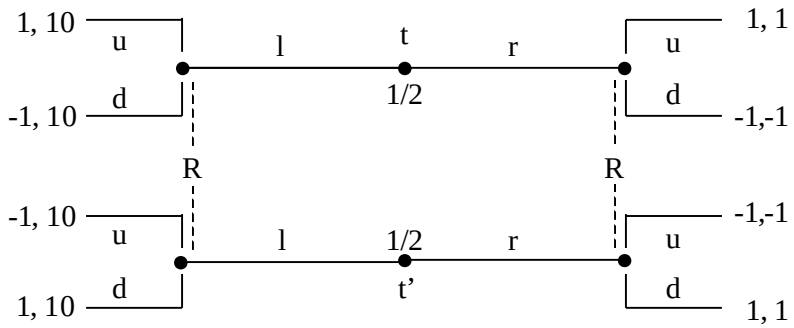
Example 2.



Again there are two SE: (1) lll, and (2) rll,u. Only (2) is a PSE. (1) is not a PSE, because there is a unique rationalization of the deviation r. K={t} rationalizes r, since if R infers t from r then R's best response is u, yielding 3 rather than the equilibrium payoff of 2 to S. Neither t' nor t" benefit from the deviation, since they would get 0 rather than 2. But here, lll is not rejected by the intuitive criterion, since r is not a bad deviation for either t' or t": if R puts sufficient weight on t" (so m is R's best response), then t' gains from the deviation r; whereas, if R puts sufficient weight on t' (so d is R's best response), then t" gains from the deviation r. [Can you see that applying D1 does not eliminate lll either?]

Our next example illustrates that multiple credible updating rules can lead to multiple PSE.

Example 3.



Here there are three SE: (1) (rl; ud), (2) (lr,ud), and (3) (ll, 1/2(ud)). All three are PSE. The two separating equilibria, (1) and (2), are PSE, since they are SE and all messages are sent in equilibrium, so beliefs are uniquely defined from Bayes' rule. The pooling equilibrium is a PSE as well, since the deviation can be rationalized by the inference K = {t,t'}. With K = {t,t'}, R maintains the prior belief if r is played, but then randomizing 50-50 between u and d is a best response for R, so both t and t' get 0 from deviating vs. 0 from the equilibrium. Hence, given the inference K = {t,t'}, not deviating is a (weak) best response for S, and so the pooling equilibrium stands. The three different credible updating rules lead to three different PSE.

Farrell would reject the pooling equilibrium, since the updating rules that lead to the separating equilibria *strictly* rationalize the deviation in the sense that the deviator strictly gains from the deviation with an inference of either K = {t} or K = {t'}, whereas no type strictly gains from deviating with K = {t,t'}. Farrell's concept in some sense lets the sender pick the credible updating rule by sending the neologism "my type is in K." This may be reasonable if the setting allows the sender to communicate in this way.