Refinements of Nash Equilibrium

1 Overview

In game theory, “refinement” refers to the selection of a subset of equilibria, typically on the grounds that the selected equilibria are more plausible than other equilibria. These notes are a brief, largely informal, survey of some of the most heavily used refinements. Throughout, “equilibria” means Nash equilibria (NE), unless I state otherwise explicitly. And throughout, I assume that the game is finite. Extending some of these concepts to more general settings can be non-trivial; see, for example, Harris, Reny and Robson (1995).

My focus is on refinements that make an appeal to rationality arguments. This is the traditional approach to refinement. An important alternative approach is based on dynamic equilibration: players start out of equilibrium and in one sense or another learn (or fail to learn) to play an equilibrium over time. For example, in 2x2 strategic form games with two pure strategy equilibria and one mixed, the mixed equilibrium is ruled out by almost any dynamic story, even though it survives a host of traditional refinements. I make some comments about dynamic refinements at various points, but I do not attempt to be systematic.

For more thorough treatments of refinements, consult Fudenberg and Tirole (1991a), van Damme (2002), and Govindan and Wilson (2008). For surveys on the epistemic foundations of some key refinements (formalizations of arguments along the lines, “it is optimal for me to play $s^i$ because I think that my opponent will play $s^j$ because I think that he thinks ...”), see Brandenburger (2008) and the introduction to Keisler and Lee (2011).

One word of warning: while it would be convenient if refinements lined up cleanly in a hierarchy from least to most restrictive, the actual relationship between most of them is complicated, as some of the examples below illustrate.

2 Strategic Form Refinements

2.1 Admissibility

Definition 1. Given a strategic form game $G$, a NE $\sigma$ is admissible iff $\sigma^i(s^i) > 0$ implies that $s^i$ is not weakly dominated.
Thus, an equilibrium is admissible iff no player plays a weakly dominated strategy with positive probability. It is not hard to see that admissible equilibria always exist in finite games. The informal rationale for admissibility is that weakly dominated strategies are imprudent and that therefore players should avoid them.

In the game of Figure 1, \((I, A)\) and \((O, F)\) are both pure strategy equilibria, but \((O, F)\) is not admissible since \(A\) weakly dominates \(F\).

\[
\begin{array}{c|cc}
 & A & F \\
\hline
I & 15,15 & -1,-1 \\
O & 0,35 & 0,35 \\
\end{array}
\]

Figure 1: Admissibility.

\((O, F)\) is not admissible since \(A\) weakly dominates \(F\).

### 2.2 Iterated Admissibility

Iterated admissibility requires that players play only those strategies that survive the iterated deletion of weakly dominated strategies. This can be an extremely powerful refinement in some games. See, Ben-Porath and Dekel (1992) for a striking example.

On the epistemic foundation of iterated admissibility, see Brandenburger (2008) and Keisler and Lee (2011).

With iterated strict dominance, it does not matter for predicted play whether players delete strategies in turn or simultaneously, or whether all weakly dominated strategies are deleted at each round, or just some of them. With weak dominance, procedural issues of this sort become substantive. Consider the game in Figure 2. There are many equilibria here, among which are (a) an equilibrium where both players randomize 50:50 between \(A\) and \(B\), (b) an equilibrium where player 1 randomizes 50:50 between \(A\) and \(B\) and player 2 plays \(C\), and (c) an equilibrium where player 1 plays \(C\) and player 2 randomizes 50:50 between \(A\) and \(B\). Note that all these equilibria have different expected payoffs. Of these, only the (a) equilibrium is admissible, because \(C\) is weakly dominated by a 50:50 randomization between \(A\) and \(B\). But equilibrium (b) survives iterated admissibility if players delete strategies in turn, with player 1 deleting first, and similarly for the (c) equilibrium. The standard response to this issue is to consider maximal deletion, by all players, at each round. Otherwise, as this example illustrates, iterated admissibility does not even imply admissibility.

\[
\begin{array}{c|ccc}
 & A & B & C \\
\hline
A & 10,0 & 0,10 & 3,5 \\
B & 0,10 & 10,0 & 3,5 \\
C & 5,3 & 5,3 & 2,2 \\
\end{array}
\]

Figure 2: Iterated Admissibility: Order Matters.
Even when deletion order is unambiguous, it is not always obvious that iterated admissibility picks out the correct equilibrium. Consider the following game. \((M, R)\)

\[
\begin{array}{c|cc}
 & L & R \\
\hline
T & 1,1 & 0,0 \\
M & 0,100 & 100,100 \\
B & 0,0 & 99,100 \\
\end{array}
\]

Figure 3: Iterated Admissibility: Other Considerations.

and \((T, L)\) are both admissible equilibria, but \((T, L)\) survives iterated admissibility while \((M, R)\) does not. \((T, L)\) is arguably implausible. First, it is Pareto dominated by \((M, R)\). Second, from the perspective of many forms of learning dynamics, \((M, R)\) is at least as robust as \((T, L)\).

### 2.3 Trembling Hand Perfection

The idea behind trembling hand perfection (THP), introduced in Selten (1975), is that your opponents might deviate from their intended strategy (their hands might tremble when pushing a button or setting a dial and thereby select the wrong strategy) and you should prepare for that in choosing your strategy. THP makes explicit the idea, implicit in much of the refinement literature, that the game is not a complete description of the strategic situation but rather is an approximation. I return to this issue briefly at the end of this subsection.

Formally, for any \(\varepsilon > 0\), replace the original game with a new game in which the set of mixed strategies is constrained to contain only those strategies that give weight at least \(\varepsilon\) to every pure strategy. A NE of this constrained game is called an \(\varepsilon\)-perfect equilibrium. For \(\varepsilon\) sufficiently small, an \(\varepsilon\)-perfect equilibrium exists (the proof is nearly identical to the proof of existence of Nash equilibrium).

**Definition 2.** A strategy profile \(\sigma\) is a trembling hand perfect (THP) equilibrium iff there is a sequence \(\{\sigma_\varepsilon\}\) such that each \(\sigma_\varepsilon\) is an \(\varepsilon\)-perfect equilibrium and \(\lim_{\varepsilon \to 0} \sigma_\varepsilon = \sigma\).

Note the quantifier: one does not need to check every sequence \(\sigma_\varepsilon\); one must merely show that there exists one such sequence. A compactness argument establishes that a THP equilibrium exists and continuity implies that any THP is a NE. Hence THP is a refinement of NE. If \(\sigma\) is already fully mixed, then it is trivially THP: for \(\varepsilon\) sufficiently small, take \(\sigma_\varepsilon = \sigma\).

THP implies admissibility and in two-player games, admissibility implies THP. Thus, in two-player games, THP and admissibility are equivalent. In games with three or more players, there are examples in which admissible (or even iterated admissible) equilibria fail to be THP. Thus, in games with three or more players, THP is stronger than admissibility.
To understand why, recall that while a strictly dominated strategy \( s^i \) is never a best response to any distribution over opposing strategy profiles, the converse is sometimes false in games with three or more players unless the distribution exhibits correlation. For essentially the same reason, while a weakly dominated strategy \( s^i \) is never a best response to any fully mixed distribution over opposing strategy profiles, the converse is sometimes false unless the fully mixed distribution exhibits correlation. THP implicitly assumes independent randomization (the strategy distribution generated by \( \sigma_\varepsilon \) is independent); as a consequence, THP can be more restrictive than admissibility.

THP does not imply iterated admissibility. For example, the \((M, R)\) equilibrium in Figure 3 is THP but not iterated admissible. Conversely, iterated admissibility implies THP in two player games but not in games with three or more players, for the reason discussed in the preceding paragraph.

In applying THP, the game is fixed but players tremble away from their best responses. Suppose that instead players best respond but that the true game is nearby in payoff terms. One can then ask which equilibria of the original game are limits of sequences of equilibria of nearby games. The answer is: all of the equilibria. See Fudenberg, Levine and Kreps (1988) and also, for recent related work, Jackson, Rodriguez-Barraquer and Tan (2011). The basic point is that refinements motivated by approximation arguments require careful thought.

### 2.4 Properness

Properness, introduced in Myerson (1978), is motivated by the idea that if \( i \) does tremble, he is more likely to tremble in directions that are least harmful to him.

Formally, say that a strategy profile \( \sigma \) is an \( \varepsilon \)-proper equilibrium iff (a) it is fully mixed and (b) for any \( i \) and any pure strategies \( s^i \) and \( \tilde{s}^i \), if the payoff to \( s^i \) is greater than that to \( \tilde{s}^i \) then \( \sigma^i(s^i) \geq \sigma(\tilde{s}^i)/\varepsilon \). For \( \varepsilon \) small, \( 1/\varepsilon \) is large, and hence this says that \( \sigma^i(s^i) \) must be much larger than \( \sigma(\tilde{s}^i) \). One can show that \( \varepsilon \)-proper equilibria exist, for \( \varepsilon \) sufficiently small (this requires more work than showing existence of \( \varepsilon \)-perfect equilibria).

**Definition 3.** A strategy profile \( \sigma \) is a proper equilibrium iff there is a sequence \( \{\sigma_\varepsilon\} \) such that each \( \sigma_\varepsilon \) is an \( \varepsilon \)-proper equilibrium and \( \lim_{\varepsilon \to 0} \sigma_\varepsilon = \sigma \).

Again, proper equilibrium exist and any proper equilibrium is a NE. Hence proper equilibrium is a refinement of Nash equilibrium. It is also almost immediate that any proper equilibrium is THP. Hence, in particular, any proper equilibrium is admissible, although not necessarily iteratively admissible.

As an example, consider the game in Figure 4. No strategy is weakly dominated, hence, in particular, the NE \((M, C)\) is THP and even survives the iterated deletion of weakly dominated strategies (trivially, as no strategies are weakly dominated). However, \((M, C)\) is not proper. As long as most of player 2’s probability weight is on \( C \), \( T \) is a better response for player 1 than \( B \). But if player 1 puts much more
weight on $T$ than on $B$, then $L$ is a better response than $C$ for player 2. Thus $(M, C)$ is not proper.

As a second example, consider again the game of Figure 3. In that game, the NE $(M, R)$ is proper even though it does not satisfy iterated admissibility. $(T, L)$ is proper as well.

Properness is seldom invoked explicitly in applied work, possibly because it can be difficult to verify. But it is important conceptually because it is relatively cleanly linked to extensive form refinements. In particular, for a given strategic form game, van Damme (1984) established that a proper equilibrium induces a sequential equilibrium (Section 3.4) in every extensive form game consistent with the original strategic form game.

2.5 Other Strategic Form Refinements

There is a bestiary of refinements and I will not be exhaustive. But let me briefly mention some issues.

First, I would be remiss not to cite the literature on Kohlberg-Mertens Stability (K-M Stability). The original paper, Kohlberg and Mertens (1986), is a trove of interesting examples on refinements in general and on the connection between strategic and extensive form refinements. The most restrictive K-M Stability refinement is the one in Mertens (1989); its definition is extremely technical.

Second, $\sigma$ is a strict NE iff, for every $i$, $\sigma^i$ is a strict best response to $\sigma^{-i}$. A strict NE is necessarily pure. A strict NE need not exist (none exists in matching pennies, for example), in contrast to most of the other refinements discussed here. When it does exist, however, it is iterated admissible, proper, and passes every version of K-M Stability.

Third, another obvious refinement is to restrict attention to the NE that are Pareto efficient relative to the other NE (they need not be Pareto efficient in an overall sense; in games that model free riding in the provision of public goods, for example, the NE are typically inefficient). One motivation is that if players can engage in pre-game bargaining over which NE to play, then they will not choose a NE that is Pareto dominated by another NE.

Application of the last two refinements is not always straightforward. For example, in the game of Figure 3, the Pareto efficient pure NE is $(M, R)$, which is not strict, while the strict equilibrium is $(T, L)$, which is Pareto dominated.
As another example, consider the game in Figure 5, which belongs to a class of games sometimes called “Stag Hunt.” This game has two pure NE, \((A, A)\) and \((B, B)\). Both are strict. \((A, A)\) Pareto dominates \((B, B)\), but \((B, B)\) is arguably more plausible. \(A\) is optimal if and only if the probability that your opponent plays \(A\) is at least \(99/101\): you have to be almost certain that your opponent is playing \(A\) in order for \(A\) to be optimal. Under learning dynamics, this translates into \((B, B)\) having a larger basin of attraction and under some (but not all) learning dynamics there is a sense in which play converges to \((B, B)\) no matter how players start out initially; see Kandori, Mailath and Rob (1993).

\[ \begin{array}{c|cc}
A & B \\
\hline
A & 100, 100 & 0, 98 \\
B & 98, 0 & 99, 99 \\
\end{array} \]

Figure 5: Pareto Dominance

3 Extensive Form Refinements

3.1 Subgame Perfection

Subgame perfection is a generalization of backward induction, one of the oldest solution concepts in game theory. Whereas backward induction is typically restricted to finite games of perfect information, subgame perfection applies to arbitrary extensive form games. Backward induction dates at least to Zermelo (1913); yes, the article is in German. Subgame perfection was introduced in Selten (1967); yes, once again, the article is in German. Rather than develop backward induction explicitly, I focus on subgame perfection.

Informally, given a game \(\Gamma\) in extensive form, a subgame of \(\Gamma\) consists of a decision node, all successor decision nodes and terminal nodes, and all associated information sets, actions, and payoffs, provided that these all comprise a well defined game. In particular, a node that is not in a singleton information set (an information set containing exactly one node) cannot serve as the initial node of a subgame. For a formal definition of subgames, see a game theory text such as Fudenberg and Tirole (1991a).

A game always contains itself as a subgame, just as a set always contains itself as a subset. A subgame that excludes at least one node of the original game is called a proper subgame. Any (behavior) strategy profile induces, in a natural way, a strategy profile in any subgame. Given a strategy profile \(\sigma\) and a subgame, the subgame is on the play path iff the probability, under \(\sigma\), of reaching the initial node of the subgame is positive.

**Theorem 1.** \(\sigma\) is a NE iff it induces a NE in every subgame on the play path.
Proof. Since any game is a subgame of itself, the “if” direction is immediate. As for “only if,” let σ be a strategy profile and let x be the initial node of a proper subgame along the play path. If σ does not induce a NE in the proper subgame, then there is some player i who can gain some amount ∆ > 0 within the subgame by deviating. If the probability of reaching x under σ is p > 0, then i can gain pΔ > 0 in the overall game by deviating within the subgame, hence σi was not optimal for i, hence σ was not a NE. The proof then follows by contraposition. ■

Definition 4. σ is a subgame perfect equilibrium (SPE) iff it induces a NE in every subgame.

In view of Theorem 1, SPE strengthens NE by requiring that players play a NE in every subgame, and not just in every subgame on the play path. One implication is that there is a difference between NE and SPE only if there is some subgame (necessarily a proper subgame) that is not on the play path. In particular, if the game has no proper subgames, which is frequently the case in Bayesian games, then SPE has no “bite”: there is no difference between NE and SPE.

A canonical way to illustrate SPE is the entry deterrence game, a version of which is given in Figure 6. Player 1 is a potential entrant into a market where player 2 is a monopolist. If player 1 stays out, then player 2 earns her monopoly profit of 35. If player 1 enters, player 2 can either acquiesce (A), in which case both firms earn 15, or start a price war (F), in which case both get a payoff of -1. This game has one proper subgame, the trivial game rooted at player 2’s decision node. The equilibrium of the proper subgame is for player 2 to choose A, and if player 1 anticipates A then player 1 will enter: the overall SPE is (I, A).

But (O, F) is also a pure strategy NE. Informally, under (O, F), player 2 threatens a price war upon entry and player 1, believing this, stays out: player 2 is able to deter entry. SPE captures the idea that this threat is not credible, because F is
not optimal should entry actually occur. This is not to say that entry deterrence is impossible. Rather, the point is that a convincing model of entry deterrence will have to have additional structure.

Another well known application of SPE is ultimatum bargaining. Player 1 can propose an integer amount \( a \in \{0, \ldots, 100\} \). Player 2 can Accept or Reject. If Player 2 Accepts then the payoffs are \((a, 100 - a)\). If Player 2 Rejects then the payoffs are \((0, 0)\).

There is a different proper subgame for each of Player 1’s possible proposals. For every proposal of 99 or less, the NE of that subgame has player 2 Accept, since getting something is better than getting nothing. Thus, in any SPE, player 2’s strategy must Accept any proposal of 99 or less. If the proposal is 100, then either Accept or Reject is a NE of that subgame, since either way player 2 gets nothing. There are, therefore, two pure SPE. In one, player 1 proposes \( a = 99 \) and player 2 plays the strategy \((\text{Accept, Accept, \ldots, Accept, Reject})\); that is, Reject 100, Accept otherwise. In the other pure SPE, player 1 proposes \( a = 100 \) and player 2 plays the strategy \((\text{Accept, \ldots, Accept})\); that is, Accept every proposal. There are also SPE that are mixtures over these pure SPE.

But there are many other NE. In particular, every \( a \in \{0, \ldots, 100\} \) can be supported in NE. For example, it is an equilibrium for player 1 to propose \( a = 50 \) and player 2 to play the strategy \((\text{Accept, \ldots, Accept, Reject, \ldots, Reject})\), with the first Reject at 51; that is, player 2 rejects proposals above 50 but accepts otherwise. In effect, player 2 says, give me half or else. And it is also an equilibrium for player 1 to propose \( a = 50 \) and player to Accept only 50, and Reject everything else. And so on.

The ultimatum bargaining game has been heavily studied in laboratory experiments. See, for example, Roth et al. (1991). The results are roughly as follows. For subjects in the role of player 1, the modal proposal is 50, with some subjects making higher proposals. Subjects in the role of player 2 accept 50 but frequently reject more aggressive proposals. The behavior of subjects in the role of player 1 is broadly consistent with maximizing monetary payoffs, given the behavior of subjects in the role of player 2. The experimental data indicate, however, that subjects in the role of player 2 are not maximizing monetary payoffs. For more discussion along these lines, see Levine and Zheng (2010).

Returning to the entry deterrence game of Figure 6, the strategic form was already given in Figure 1. In the strategic form, the NE that was not SPE, namely \((O,F)\), is eliminated by admissibility. Similarly, if we were to write out the strategic form for the ultimatum bargaining game, we would find that only the SPE were admissible. These examples suggest a connection between SPE and admissibility or iterated admissibility. A relationship exists but it is complicated. There are classes of games in which, as in the examples above, SPE and admissibility or iterated admissibility are equivalent. But there are also simple examples where they are not.

To see that SPE does not imply admissibility, consider the game in Figure 7.
The associated strategic form is in Figure 8. \( L \) weakly dominates \( R \) and hence the unique admissible equilibrium is \((D, L)\). But \((U, R)\) is SPE, because there are no proper subgames.

Perhaps more surprisingly, SPE does not imply admissibility even in games with perfect information. Consider the game of Figure 9, which appears in van Damme (2002). In effect, player 1 can either implement \( A \) herself or delegate the choice to player 2. The (full) strategic form is in Figure 10. There are two pure SPE: \((UA, A)\) and \((DA, A)\). But only \((UA, A)\) is admissible: it is weakly dominant to choose \( A \) oneself.

Conversely, admissibility, or even iterated admissibility, does not imply SPE. Consider, for instance, the game of Figure 11. The unique SPE of this game has player 1 choose \( I \) and both player 2 and 3 randomize 50:50. But \((O, D, L)\) is also a NE, and it is (iterated) admissible, for the simple reason that in this game no strategy for any player is weakly dominated.

### 3.2 Weak Perfect Bayesian Equilibrium.

As already mentioned, SPE has no bite when there are no proper subgames. This happens routinely in Bayesian games but the issue is not restricted to Bayesian games. This issue motivates a class of refinements of which the seminal paper is Kreps and Wilson (1982); see also Fudenberg and Tirole (1991b).
Given an extensive form game, define a function $\mu$ from information sets to the interval $[0, 1]$ with the property that for any information set $h$,

$$\sum_{x \in h} \mu(x) = 1.$$ 

If $i$ is the player at information set $h$, and $x \in h$, then $\mu(x)$ is player $i$’s belief, conditional on being at $h$, that she is at $x$. The pair $(\sigma, \mu)$ is an assessment.

Thus far, I have not said anything about how $\sigma$ and $\mu$ are related. Let $\text{Prob}[x|\sigma]$ denote the probability of reaching decision node $x$ given the strategy profile $\sigma$. Similarly, let $\text{Prob}[h|\sigma]$ denote the probability of reaching information set $h$ given the strategy profile $\sigma$. Thus, the decision node $x$ is on the play path iff $\text{Prob}[x|\sigma] > 0$. Similarly, the information set $h$ is on the play path iff $\text{Prob}[h|\sigma] > 0$.

**Definition 5.** $(\sigma, \mu)$ is Bayes Consistent (BC) iff for any information set $h$ on the play path and any decision node $x \in h$,

$$\mu(x) = \frac{\text{Prob}[x|\sigma]}{\text{Prob}[h|\sigma]}.$$ 

Off the play path, Bayes Consistency imposes no restrictions on $\mu(x)$. 

---

**Figure 9:** An SPE that is not admissible

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>1,1</td>
<td>1,1</td>
</tr>
<tr>
<td>D</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>A</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>B</td>
<td>1,1</td>
<td>1,1</td>
</tr>
</tbody>
</table>

**Figure 10:** The strategic form for the game in Figure 9.
Turning now to $\sigma$, say that $(\sigma, \mu)$ is sequentially rational (SR) at information set $h$, if, in the game fragment comprising $h$ and all successor nodes and terminal nodes, and all associated information sets, actions, and payoffs, no player can get higher expected payoff by deviating at $h$ or at any successor information set, given the strategies of the other players. If there are two or more decision nodes in an initial information set, then we will have to use $\mu$ to compute expected payoffs from that point in the game looking forward. This is why $\mu$ is critical. A formal definition of SR is slightly fussy; see Fudenberg and Tirole (1991a).

**Theorem 2.** $\sigma$ is a NE iff there is a $\mu$ such that $(\sigma, \mu)$ is (a) BC and (b) SR at every information set on the play path.

I omit the proof, which is similar to that of Theorem 1.

**Definition 6.** $(\sigma, \mu)$ is a Weak Perfect Bayesian Equilibrium (WPBE) iff it is (a) BC and (b) SR at every information set.

In view of Theorem 2, WPBE strengthens NE by requiring that $\sigma$ be SR at every information set and not just at every information set on the play path. Given a NE $\sigma$, say that $\sigma$ is supported as a WPBE iff there exists a $\mu$ such that $(\sigma, \mu)$ is WPBE. It will typically be the case that if some information sets are not on the play path then the supporting $\mu$ is not unique. This is illustrated in some of the examples below.

Think of WPBE as follows. Given a game in extensive form, compute the set of NE. For a given NE $\sigma$, check whether it can be rationalized in the sense that there is a $\mu$, giving player beliefs at the various information sets, such that $\sigma$ makes sense (is sequentially rational) given $\mu$. Bayes Consistency is a minimal condition for $\mu$ to makes sense. The spirit of WPBE is that if $\sigma$ cannot be supported as a
WPBE, then it is implausible. If $\sigma$ can be supported as a WPBE, then it may still be implausible; in particular, it may be that $\mu$ is strange, as some of the examples below will illustrate.

The relationship between SPE and WPBE is messy. Both generalize backward induction to games without perfect information, and both coincide with backward induction in games with perfect information. But in games of imperfect information, they are different.

Consider first the game in Figure 12. Here, there is a pure NE $(O, R)$. This is SPE for the irritating reason that there are no proper subgames. But it is clearly odd because player 2 always does better playing $L$ rather than $R$ at her information set, regardless of whether she is at $x$ or $x'$. Because of this, for any $\mu$, $(O, R)$ is not SR at 2’s information set. Only the NE $(U, L)$ is supported as WPBE in this game. In this WPBE, BC requires that since 2’s information set is on the play path, $\mu(x) = 1$.

Note that in the game in Figure 12, strategy $L$ is weakly dominated by $R$. So it is tempting to conjecture that WPBE is related to admissibility. In some sense it is, but the game of Figure 9 serves as a reminder that the situation is subtle. In that game, $(DA, A)$ can be supported as a WPBE (trivially, since the game is one of perfect information), but it is not admissible.

Now consider the game in Figure 13. The unique SPE is $(IU, L)$. In particular, in the proper subgame, $U$ strictly dominates $D$, so the unique equilibrium in the subgame is $(U, L)$. But if player 1 chooses $O$, then BC does not require $\mu(x) = 1$, as strange as this may seem. $(OU, R)$ is supported as a WPBE in which $\mu(x) = 0$ (or, more generally, $\mu(x) \leq 1/2$). In effect, player 2 is threatening player 1 with strange beliefs: she is saying, “play $O$ (which gives me my preferred payoff of 10) or else if you enter, I will assume that you have played $D$, even though this contradicts the
stated strategy $OU$, and respond with $R$.”

3.3 Perfect Bayesian Equilibrium

As illustrated by the $(OU, R)$ NE of Figure 13, implausible NE can be supported as WPBE with implausible $\mu$. This motivates further restrictions on $\mu$. The restrictions required for PBE have the flavor of, “$\mu$ should reflect Bayesian updating as much as possible (even at information sets off the play path).” Rather than formalize this, which is somewhat messy, let me note that if $(\sigma, \mu)$ is a PBE, then it induces a WPBE in every proper subgame. In particular, if $\sigma$ can be supported as a PBE, then it is an SPE. This is enough to knock out the $(OU, R)$ equilibrium in Figure 13.

3.4 Sequential Equilibrium

Historically, sequential equilibrium, introduced in Kreps and Wilson (1982), predates WPBE and PBE. Conceptually, however, it is a strengthening of PBE.

**Definition 7.** Given an assessment $(\sigma, \mu)$, $\mu$ is consistent iff there exists a sequence of assessments $\{(\sigma_t, \mu_t)\}$ such that (a) $\lim_{t \to \infty}(\sigma_t, \mu_t) = (\sigma, \mu)$, (b) for every $t$, $\sigma_t$ is fully mixed, and (c) for every information set $h$ and every $x \in h$,

$$\mu_t(x) = \frac{\text{Prob}[x|\sigma_t]}{\text{Prob}[h|\sigma_t]}.$$

Note that the definition of $\mu_t(x)$ makes sense because $\sigma_t$ is fully mixed and hence $\text{Prob}[h|\sigma_t] > 0$: every information set is on the play path. Note also that consistency does not require that one consider every possible $\{\sigma_t\}$; it only requires that there be some such sequence $\{\sigma_t\}$. 

Figure 13: WPBE does not imply SPE.
For motivation, recall that, given $\sigma$, ordinary probability calculations pin down $\mu$ at information sets on the play path. The question is how to specify $\mu$ at information sets off the play path. Under Bayes consistency, the answer is: anyway you want. Under consistency, one must justify $\mu$ off the play path as being the limit of the conditional probabilities that would be generated if players trembled.

**Definition 8.** $(\sigma, \mu)$ is a sequential equilibrium (SE) iff it is (a) consistent and (b) SR at every information set.

An SE always exists. One can show that if $(\sigma, \mu)$ is an SE, then it is a PBE, and hence is a WPBE and that $\sigma$ is an SPE. Let me make a few comments.

Consistency implies Bayes consistency and hence a consistent $\mu$ is pinned down by $\sigma$ at information sets on the play path; for such information sets, you don’t need to check sequences of fully mixed $\sigma_t$. It is only for information sets off the play path that consistency needs to be checked. One implication is that if an NE has the property that all information sets are on the play path, then that NE can be supported as an SE. In particular, in a simultaneous move game, any NE can be supported as an SE.

The definition of consistency looks a bit like the definition of THP, and I even mentioned trembling by way of motivation, but consistency does not assume or imply that $\sigma_i^t$ is an approximate best response to $\sigma_{-i}^t$: there is no claim that $\sigma$ can be approximated by a sequence of $\varepsilon$-perfect equilibria. Because of this, SE does not imply admissibility. For example, consider again the game of Figure 7. $(D, R)$ can be supported as an SE (trivially, since player 2’s information set is on the play path), but it is not admissible. Similarly, in the game of Figure 9, $(DA, A)$ can be supported as an SE, but it is not admissible. Conversely, since THP does not imply SPE, it does not imply SE either.

On the other hand, properness does imply SE. As noted in Section 2.4, if $\sigma$ is proper in a strategic form then it is supported as an SE in every extensive form that generates that (reduced) strategic form; see van Damme (1984). The converse is not true, since properness implies admissibility, while SE does not.

Last, SE is stronger than PBE in general games, but the precise way in which this is true is subtle and I will not pursue it; see Fudenberg and Tirole (1991a) for a discussion. In practice it is often the case that either WPBE or SPE is strong enough or that something even stronger than SE is needed to eliminate implausible equilibria.

### 3.5 Trembling Hand Perfection in the Agent Normal Form

SE is closely related to another solution concept, trembling hand perfection in the agent normal form (THPANF), introduced, along with THP, in Selten (1975). I mention it here briefly for completeness; think of this section as parenthetical. In applied work, game theorists almost invariably use SE or PBE rather than THPANF.
Under THPANF, THP is applied to a modified strategic form in which, in effect, there is a different player for each information set. One can show that if an equilibrium $\sigma$ is THPANF, then it can be supported as a SE. Moreover, THPANF rejects $(U, R)$ in the game of Figure 7. Thus, THPANF is stronger than SE. But SE and THPANF are equivalent for generic terminal node payoffs, a condition violated in Figure 7. And SE is easier to compute (even though checking consistency can be difficult), which is one of the reasons it, or PBE, is much more widely used.

Based on the game of Figure 7, it is tempting to conjecture that THPANF implies admissibility. This is false. The game of Figure 9 is once again a counterexample: the inadmissible equilibrium $(DA, A)$ is THPANF. Informally, the reason is that, under THPANF, player 1 is just as worried about his own trembling at his second information set as he is about player 2 trembling. In summary, THP does not imply THPANF (since a THP does not have to be subgame perfect) and THPANF does not imply THP (since a THPANF does not have to be admissible).

For given $\varepsilon > 0$, a strategy profile that is an $\varepsilon$-perfect equilibrium of the modified strategic form can be very far from any equilibrium of the original game. For example, consider an extensive form game of perfect information with only one player, who can choose either $O$ immediately, for a payoff of 1000, or $I$, in which case the player faces a sequence of $T$ choices between actions $A$ and $B$. If he chooses correctly at every node, then he gets a payoff of 1001; otherwise he gets 0. There is perfect information, so the player knows which sequence of $A$ and $B$ to play. The unique Nash equilibrium, and hence the unique THPANF equilibrium, is for the player to choose $I$ followed by the correct sequence of $A$ and $B$. But for any given $\varepsilon > 0$, no matter how small, if $T$ is sufficiently large, the $\varepsilon$-perfect equilibrium in the modified strategic form has the player choose $O$. Similar phenomena can also arise for THP in strategic form games.

3.6 The Cho-Kreps Intuitive Criterion

The SE belief restriction called consistency depends on the strategy profile $\sigma$ and on the extensive form of the game, but not on the game’s payoffs. Consistency does not ask whether, given the game’s payoffs, beliefs make sense off the play path. As illustration of why payoffs might affect beliefs off the play path, consider the game of Figure 14, taken from Cho and Kreps (1987). Player 1 can either be Weak or Strong. Her action is her choice of breakfast, Beer or Quiche. Player 2 observes player 1’s breakfast, but not her type, and chooses whether to Fight player 1 or Run away. Other things being equal, player 1 would prefer Quiche if Weak and Beer if Strong, but she would also like player 2 to Run. She thus potentially faces a tradeoff between having her preferred breakfast and inducing player 2 to Run. Player 2 would like to Fight a Weak player 1 but would like to Run from a Strong player 1. Less whimsically, one can think of this as a model of entry deterrence with asymmetric information. Player 1 is the incumbent firm in some market and has either low production costs or high production costs. Player 2, the potential entrant,
cannot observe player 1’s production costs directly, but he can observe player 1’s market behavior. Beer corresponds to player 1 charging a low price while Quiche corresponds to player 1 charging a high price.

This game has two pure strategy pooling NE: ((Beer, Beer), (Run, Fight)) and ((Quiche, Quiche), (Fight, Run)), where the first strategy profile, for example, is read, Beer if Weak, Beer if Strong, Run if Beer, Fight if Quiche. There is no separating NE in this particular game. The first pooling NE can be supported as a WPBE provided \(\mu(y) \geq 1/2\) (by Bayes Consistency, \(\mu(x) = 1\)). The second can be supported as a WPBE provided \(\mu(x) \geq 1/2\) (by Bayes Consistency, \(\mu(y) = 1\)). One can verify easily, although I will not do so, that all of these WPBE are, in fact, SE.

Of these two classes of WPBE, the (Quiche, Quiche) WPBE are arguably implausible for the following reason. The Weak type player 1 has no possible reason to deviate from Quiche: she is making 3 in equilibrium by eating Quiche and could get at most 2 by drinking Beer, even if she somehow convinced player 2 to Run in response to Beer. Quiche is said to be equilibrium dominated for the Weak type in the (Quiche, Quiche) NE. This suggests that in any (Quiche, Quiche) WPBE, player 2 should assign zero probability to player 1 being Weak if she sees player 1 drink Beer: \(\mu(x) = 0\). But \(\mu(x) = 0\) is not compatible with any of the (Quiche, Quiche) WPBE: if \(\mu(x) = 0\), then sequential rationality forces player 2 to Run after seeing Beer, and if player 2 Runs after seeing player 1 drink Beer, then a Strong player 1 will deviate to drinking Beer.
In contrast, the (Beer, Beer) equilibrium is free from this problem. Quiche is equilibrium dominated for Strong in the (Beer, Beer) equilibrium, suggesting that player 2 should assign zero probability to player 1 being Strong after seeing player 1 eat Quiche: $\mu(y) = 1$. And there is, in fact, a (Beer, Beer) WPBE with $\mu(y) = 1$.

The Intuitive Criterion of Cho and Kreps (1987) (CKIC) formalizes the ideas just illustrated. In contrast to the other refinements discussed here, CKIC applies only to a special class of games, of which the game of Figure 14 was an example. In these games, there are two players. Nature moves first and chooses a type for player 1 (e.g., Weak or Strong). Player 1 observes her type and chooses an action $a \in A$ as a function of her type (e.g., Beer or Quiche). For simplicity, assume that, as in Figure 14, the set of allowed actions $A$ is the same for all types; Cho and Kreps (1987) allows the set of allowed actions to vary with type. Player 2 observes player 1’s action, but not her type, and chooses a response. The game is then over and players get their payoffs. These games all have the property that the set of SE and the set of WPBE coincide.

**Definition 9.** Given a NE $\sigma$ of a game of the above form and an action $a$ for player 1 that is not played under $\sigma$, $a$ is equilibrium dominated for type $\theta$ iff $\theta$’s equilibrium payoff is higher than her payoff from deviating to action $a$, for any response to action $a$ that is sequentially rational for player 2 given beliefs $\mu$, for any $\mu$.

**Definition 10.** A NE $\sigma$ of a game of the above form fails CKIC iff there is an action $a$ for player 1 that is not played under $\sigma$ and for which the following is true.

1. There is at least one type for whom $a$ is equilibrium dominated.

2. There is at least one other type whose payoff from deviating to action $a$ is higher than her equilibrium payoff, for any response by player 2 to action $a$ that is sequentially rational given $\mu$, for any $\mu$ that, at the $a$ information set, puts probability zero on all types for which action $a$ is equilibrium dominated.

Note that for CKIC to eliminate equilibria, it must that there is an unused action that is equilibrium dominated for some types but not for others. If no such unused action exists, because all unused actions are either equilibrium dominated for all types or for no types (or because no action is unused), then the equilibrium satisfies CKIC vacuously.

In the (Quiche, Quiche) Nash equilibrium of the example, action $a$ was Beer. As already noted, Beer is equilibrium dominated for Weak. If $\mu(x) = 0$, then sequential rationality requires that player 2 Run after seeing player 1 drink Beer, which gives a Strong player 1 a payoff of 3 from drinking Beer, which is more than the 2 she gets in equilibrium by eating Quiche. Thus, the (Quiche, Quiche) NE fails CKIC. Conversely, the (Beer, Beer) Nash equilibrium passes CKIC. Quiche is equilibrium dominated for Strong but if zero probability is put on Strong after seeing Quiche, hence $\mu(y) = 1$, then sequential rationality forces player 2 to Fight on seeing Quiche,
which means that a Weak player 2 has a lower payoff from deviating to Quiche than from sticking to the equilibrium strategy of Beer.

As defined, the property of satisfying CKIC is a refinement of NE rather than of SE. In practice, however, CKIC is used as a refinement of SE. Thus, one computes (even if only informally) the set of SE and then checks which of these SE satisfy CKIC. In particular, given a set of SE characterized by a strategy profile \( \sigma \) and a set of associated \( \mu \), one checks whether any of these \( \mu \) have the property that, conditional on any unused action, zero probability is assigned to types for whom that action was equilibrium dominated. If such an \( \mu \) exist, then \( \sigma \) satisfies CKIC, if not, then \( \sigma \) fails CKIC. I illustrated this in the example above.

CKIC is the best known, and most heavily used, of a large class of related refinements in games of this general type. For more on this, see Cho and Kreps (1987) and some of the other surveys cited. For another classic paper in this genre, see Banks and Sobel (1987).

References


