John Nachbar
Washington University in St. Louis
This Version: February March 3, 2022

Refinements of Nash Equilibrium[1]

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 et al. (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 2×2 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 (1991), 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 σ 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. An admissible equilibrium always exists in finite games: delete all weakly dominated pure strategies and take the equilibria of the remaining game. 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.

	A	F
I	$15, 15$	$-1, -1$
O	$0, 35$	$0, 35$

Figure 1: Admissibility.

As a second example, which will be useful later, consider the game in Figure 2. This game has two NE: (T, L) and (M, C): there are no mixed NE. (M, C) is not admissible.

	L	C
T	$1, 1$	$0, 0$
M	$0, 0$	$0, 0$

Figure 2: A game with exactly two NE.

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 strictly 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 3. 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 A 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

	A	B	C
A	$10, 0$	$0, 10$	$3, 5$
B	$0, 10$	$10, 0$	$3, 5$
C	$5, 3$	$5, 3$	$2, 2$

Figure 3: Iterated Admissibility: Order Matters.

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.

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) and (T, L) are both admissible equilibria, but (T, L) survives iterated admissibility while (M, R) does not. (B is strictly dominated by a mixture of T and M. Deleting B, R is weakly dominated.) (M, R), however, is arguably as plausible as (T, L). (M, R) Pareto dominates (T, L). (M, R) is proper (see Section 2.4). And, from the perspective of many forms of learning dynamics, (M, R) is at least as robust as (T, L).

	L	R
T	$1, 1$	$0, 0$
M	$0, 100$	$100, 100$
B	$0, 0$	$99, 100$

Figure 4: Iterated Admissibility: Other Considerations.

2.3 Trembling Hand Perfection

The standard motivation for trembling hand perfection (THP), introduced in Selten (1975), is that players might deviate from their intended strategy (their hands might tremble when pushing a button or setting a dial and thereby select the wrong strategy).[2] 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 motivation briefly at the end of this subsection.

There are several equivalent characterizations of THP; the characterization that I use here first appeared in Myerson (1978). For any $\varepsilon > 0$ say that σ is an *ε-perfect equilibrium* iff σ is fully mixed and gives weight at most ε to any strategy that is not a best response.

[2] Selten (1975) refers to the solution concept simply as "perfection" but "trembling hand perfection" has become standard.

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

Continuity of expected payoffs implies that any THP equilibrium is a NE. Hence THP equilibrium is a refinement of NE. If σ is already fully mixed, then it is trivially THP: for ε sufficiently small, take $\sigma_\varepsilon = \sigma$ for every ε. Existence of ε-perfect equilibrium follows from an argument similar to the one used to establish existence of Nash equilibrium.[3] Existence of a THP equilibrium then follows by compactness.

THP implies admissibility in any game. In two-player games, admissibility implies THP (van Damme (1991) contains a proof). Thus, in two-player games, THP and admissibility are equivalent. In games with three or more players, this equivalence can break down. Figure 10, which appears later in these notes, provides one example. Thus, in games with three or more players, THP is stronger than admissibility.

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 et al. (1988) and also, in a similar vein, Jackson et al. (2012). 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, they are (much) more likely to tremble in directions that are least harmful to them.

Formally, say that a strategy profile σ is an ε-proper equilibrium iff (a) it is fully mixed and (b) for any i and any pure strategies s^i and $\hat{s}^i$, if the payoff to s^i is greater than that to $\hat{s}^i$ then $\sigma^i(s^i) \geq \sigma(\hat{s}^i)/\varepsilon$. For ε small, $1/\varepsilon$ is large, and hence this says that $\sigma^i(s^i)$ must be much larger than $\sigma(\hat{s}^i)$.

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

For ε sufficiently small, an ε-proper equilibrium exists; the proof is similar to the proof of existence of ε-perfect equilibrium given in Footnote 3. Existence of a

[3] For any $\varepsilon > 0$ sufficiently small (smaller than the reciprocal of the cardinality of any player's strategy set), compute the best response correspondence for the modified game in which the probability weight on every pure strategy is constrained to be at least ε. For any player, any best response in the modified game gives weight ε to any pure strategy that is not a best response in the original game. The best response correspondence for the modified game is non-empty, convex-valued and has a closed graph. Existence of ε-perfect equilibrium then follows by the Kakutani fixed point theorem.

	L	C	R
T	$1, 1$	$0, 0$	$-2, -2$
M	$0, 0$	$0, 0$	$-1, -1$
B	$-2, -2$	$-1, -1$	$-3 - 3,$

Figure 5: Properness

proper equilibrium then follows by compactness. Any proper equilibrium is a THP. Hence properness is a refinement of NE and any proper equilibrium is admissible, although not necessarily iteratively admissible.

As an example, recall the game of Figure 2, which has two NE, one of which is not THP (it is not admissible, and since this is a two-player game, THP and admissibility are equivalent). Now suppose that we add a strictly dominated strategy for each player, as in Figure 5, which is a variation of an example in Myerson (1978). Since B and R are strictly dominated, the NE of this modified game are still (T, L) and (M, C). But, because of the modification, M and C are not weakly dominated, and so (M, C) is admissible, hence THP. (M, C) is not proper, however, as one can verify. As this example illustrates, THP is not robust to the introduction of strictly dominated strategies, but properness is.

As a second example, consider again the game of Figure 4. In that game, the NE (M, R) is proper even though it does not satisfy iterated admissibility, reinforcing the argument made earlier that (M, R) is a reasonable NE. ((T, L) is proper as well.)

Properness is seldom invoked explicitly in applied work. But it is important conceptually because it is relatively cleanly linked to extensive form refinements. In particular, for a given strategic form game, Kohlberg and Mertens (1986) 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 Kohlberg-Mertens Stability

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. Section 3.6 discuss some implications of K-M Stability.

2.6 Other Strategic Form Refinements

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

First, σ is a *strict* NE iff, for every i, σ^i is a strict best response to σ^{-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.

Second, 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 4, 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 6, 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 et al. (1993) and Young (1993).

	A	B
A	$100, 100$	$0, 98$
B	$98, 0$	$99, 99$

Figure 6: Pareto Dominance

3 Extensive Form Refinements

I assume that the extensive form game is finite and satisfies perfect recall, so that there is an equivalence between behavior strategies and mixtures over pure strategies. Accordingly, assume that any extensive form mixed strategy is, in fact, given in its behavior strategy form (which is standard in the literature).

3.1 Subgame Perfection

Subgame perfection was introduced in Selten (1967) as a generalization of backward induction, one of the oldest solution concepts in game theory; yes, the article is in German, and no, I do not read German. Whereas the term backward induction is typically restricted to finite games of perfect information, subgame perfection applies to arbitrary extensive form games. Rather than develop backward induction explicitly, I focus on subgame perfection.

Informally, given a game Γ in extensive form, a *subgame* of Γ 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 an advanced 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 induces a strategy in any subgame, often called the *continuation strategy*: for each information set the subgame, the associated continuation strategy chooses the same probability distribution over actions as the original strategy did. Given a strategy profile σ and a subgame, the subgame is on the play path iff the probability, under σ, of reaching the initial node of the subgame is strictly positive.

Theorem 1. *σ 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 $\Delta > 0$ within the subgame by deviating. If the probability of reaching x under σ is $p > 0$, then i can gain $p\Delta > 0$ in the overall game by deviating within the subgame, hence σ^i 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 example for illustrating SPE is the entry deterrence game, a version of which is given in Figure 7. 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.

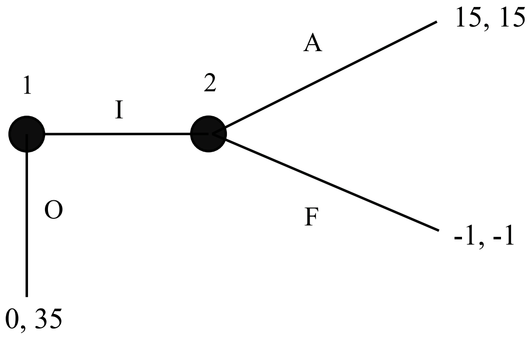

Figure 7: A Basic Entry Deterrence Game.

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 (Accept, Accept, ..., Accept, Reject); that is, Reject 100, Accept otherwise. In the other pure SPE, player 1 proposes $a = 100$ and player 2 plays the strategy (Accept, ..., Accept); that is, Accept every proposal. There are also SPE that are 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 (Accept, ..., Accept, Reject, ..., 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 7, the strategic form was already given in Figure 1. In the strategic form, the NE that was not an SPE, namely (O,F), is eliminated by THP/admissibility. This suggests a connection between SPE and THP/admissibility or iterated admissibility. While there are classes of games in which, as in the examples above, SPE and THP/admissibility or iterated admissibility are equivalent, there are also simple examples where they are not.

Figure 8 shows that THP need not imply SPE. The unique SPE is (IB, I), and this is indeed also THP. But the NE (OA, O) is THP even though it is not an SPE. The problem is that the strategies OA and OB look identical in the strategic form; see Figure 9. This motivates the concept of THP in the Agent Normal Form, discussed in Section 3.5.

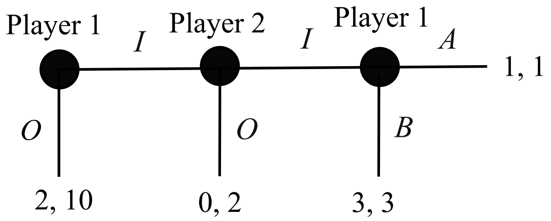

Figure 8: A game with a THP NE that is not an SPE

	O	I
OA	$2, 10$	$2, 10$
OB	$2, 10$	$2, 10$
IA	$0, 2$	$1, 1$
IB	$0, 2$	$3, 3$

Figure 9: The strategic form for the game in Figure 8.

Remark 1. To make the game of Figure 8 somewhat more interesting, I have chosen payoffs so that the equilibrium (OA, O) has the following interpretation. Player 2 threatens to play O in order to induce player 1 to play O as well, giving player 2

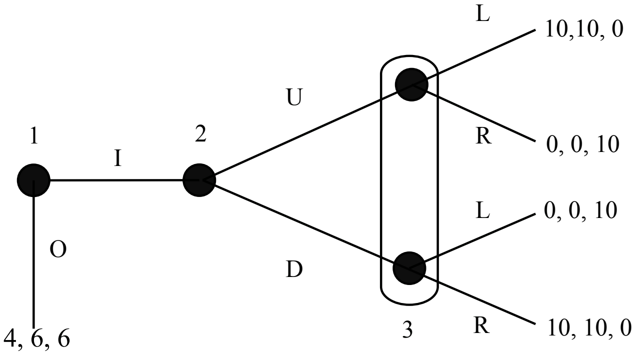

Figure 10: An game with an admissible NE that is not subgame perfect.

their preferred payoff, namely 10. Moreover, this threat by player 2 is credible in the (admittedly feeble) sense that O would be optimal for player 2 *if* player 2 believed player 1 would choose A rather than B. Do not read too much into this, however, since the example continues to work even if I changed the 10 to, say, -1. A similar remark about many of the examples below. □

In addition, admissibility, or even iterated admissibility, may not imply subgame perfection even when THP does, as illustrated by the game of Figure 10. The unique SPE of this game has player 1 choose I and both player 2 and 3 randomize 50:50. But (O, D, L), although not an SPE or THP, is (iterated) admissible, because no strategy for any player is weakly dominated.

Conversely, SPE does not imply THP/admissibility. Consider the game in Figure 11. The associated strategic form is in Figure 12. L weakly dominates R and hence

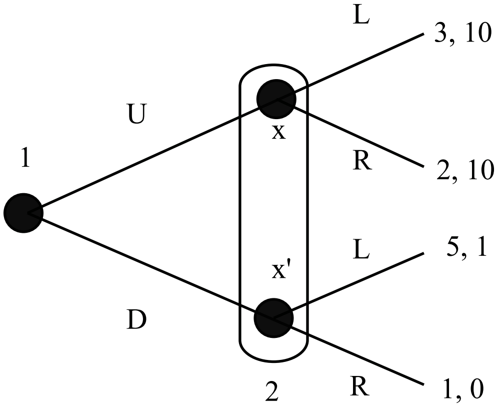

Figure 11: A game with an SPE that is not admissible.

	L	R
U	$3, 10$	$2, 10$
D	$5, 1$	$1, 0$

Figure 12: The strategic form for the game in Figure 11.

the unique THP/admissible equilibrium is (D, L). But (U, R) is SPE, because there are no proper subgames.

Finally, SPE does not imply THP/admissibility even in games with perfect information. Consider the game of Figure 13, taken from Kohlberg and Mertens (1986). In effect, player 1 can either implement A herself or delegate the choice to player

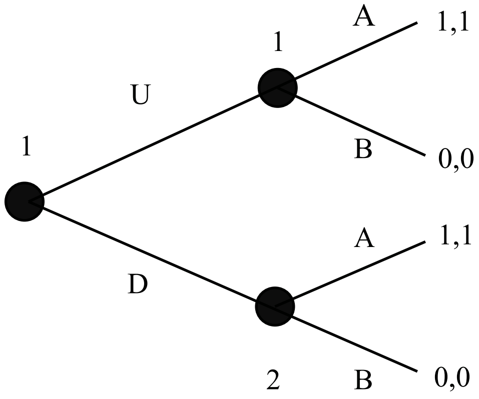

Figure 13: An example with an SPE that is not THP/admissible.

2. The (full) strategic form is in Figure 14. There are two pure SPE: (UA, A) and

	A	B
UA	$1, 1$	$1, 1$
UB	$0, 0$	$0, 0$
DA	$1, 1$	$0, 0$
DB	$1, 1$	$0, 0$

Figure 14: The strategic form for the game in Figure 13.

(DA, A). But only (UA, A) is THP/admissible: it is weakly dominant to choose A oneself.

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 weakness of SPE motivates a class of refinements of which the seminal paper is Kreps and Wilson (1982).

Given an extensive form game, define a function μ from information sets to the interval $[0, 1]$ with the property that for any information set h,

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

The intended interpretation is that if i is the player acting at h, and $x \in h$, then $\mu(x)$ is player i's *belief*, conditional on being at information set h, that she is at decision node x. The pair (σ, μ) is called an *assessment*.

Thus far, I have not said anything about how σ and μ are related. Let $\text{Prob}[x|\sigma]$ denote the probability of reaching decision node x given the strategy profile σ. Similarly, let $\text{Prob}[h|\sigma]$ denote the probability of reaching information set h given the strategy profile σ. 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. *(σ, μ) is* Bayes consistent (BC) *iff for any information set h on the play path and any decision node $x \in h$,*

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

Thus, given a strategy profile σ, if x is a decision node for player i at information set h, and if h is on σ's play path, then Baye's consistency requires that i's belief that she is at decision node x conditional on being at information set h equals the actual probability that she is at x conditional on h. If, however, h is off the play path, then Bayes consistency imposes no restrictions on μ.

As for σ, say that (σ, μ) is *sequentially rational* (SR) at information set h (the following characterization is informal), iff, in the game fragment comprising h and all successor decision 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 the game fragment is not a subgame (if the initial information set contains more than one decision node, for example), then one will have to use μ compute expected payoffs.

Theorem 2. *σ is a NE iff there is a μ such that (σ, μ) 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. BC is essential for this result. For example, suppose that the game is Matching Pennies, formalized as an extensive form game in which player 2 moves second without seeing player 1's action. If we drop BC, we could have an assessment in which both players choose T, and T is sequentially rational for player 2 because player 2 believes (violating BC), that player 1 has actually chosen H.

Theorem 2 motivates the following definition, which is roughly analogous to that for subgame perfection.

Definition 6. (σ, μ) *is a* weak perfect Bayesian equilibrium *(WPBE) iff it is (a) Bayes consistent and (b) sequentially rational at every information set.*

Remark 2. The name "weak perfect Bayesian equilibrium" originates, I think, in Mas-Colell et al. (1995). The name is a reference to "perfect Bayesian equilibrium," introduced in Fudenberg and Tirole (1991b) and discussed in Section 3.3. The same concept also appears in Myerson (1991), where it is called *weak sequential equilibrium.* □

Existence of WPBE follows from existence of sequential equilibrium (Section 3.4) which in turn follows from existence of THP-ANF equilibrium (Section 3.5).

In view of Theorem 2, if (σ, μ) is a WPBE then σ is a NE. WPBE strengthens NE by requiring that σ be SR at every information set and not just at every information set on the play path. Given a NE σ, say that σ is *supported* as a WPBE iff there exists a μ such that (σ, μ) is WPBE. It will typically be the case that if some information sets are not on the play path then the supporting μ 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 σ, check whether σ can be rationalized in the sense that there are beliefs μ such that σ makes sense (is sequentially rational at every information). Bayes consistency is a minimal condition for μ to makes sense. The spirit of WPBE is that if σ *cannot* be supported as a WPBE, then it is implausible. If σ *can* be supported as a WPBE, then it may *still* be implausible; in particular, it may be that μ is strange, as some of the examples below will illustrate.

SPE and WPBE ccoincide in games with perfect information. But in games of imperfect information, they can be different.

Consider first the game in Figure 15, which appeared originally in Selten (1975). There are no proper subgames, hence all NE are SPE. In particular, the NE (D, a, L) is an SPE. This NE looks odd, however, because if player 3 plays L then player 2 "should" chose d instead of a (and get 4 instead of 1 should player 1, for whatever reason, play A instead of D). In the language just developed above, the action a is not sequentially rational given the strategies of the other players, hence this NE is not a WPBE. The belief function μ plays no role in this particular argument. The NE (A, a, R) can be supported as a WPBE provided $\mu(x) \leq 1/3$ (making R sequentially rational).

Consider next the game in Figure 16. Once again, there are no proper subgames and hence all NE are SPE. The NE (O, R), however, looks odd because player 2

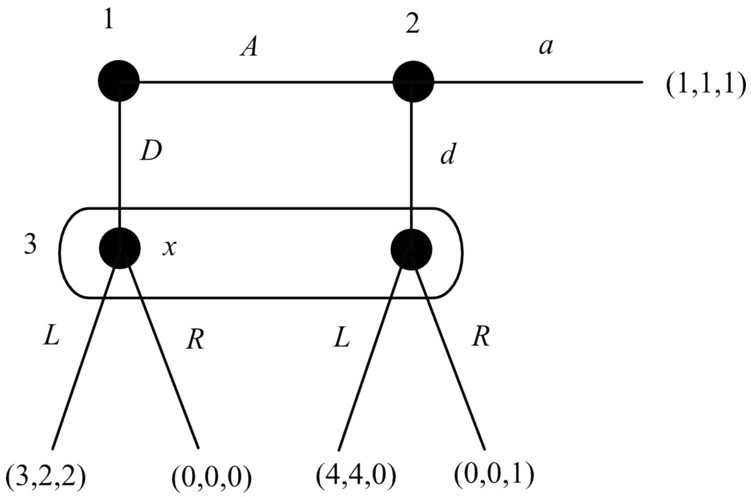

Figure 15: A game with an SPE that cannot be supported as a WPBE.

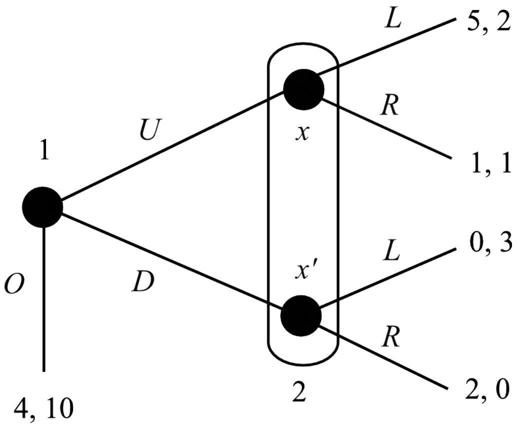

Figure 16: Another game with an SPE that cannot be supported as a WPBE.

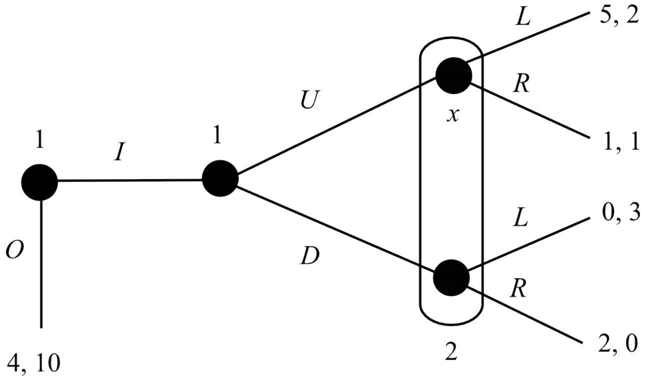

Figure 17: A modification of Figure 16: SPE and WPBE now coincide.

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 μ, R is not sequentially rational at 2's information set. Only the NE (U, L) is supported as WPBE in this game. In this WPBE, 2's information set is on the play path, and hence Bayes consistency requires $\mu(x) = 1$. In this example, in contrast to that of Figure 15, μ plays a role: we used it to check that R was not sequentially rational.

Consider next the game in Figure 17. This is the same game as that in Figure 16, except that in order to play U or D, player 1 now has first to select I. Arguably, this should not affect the outcome of the game. However, with this modification, there is now a proper subgame and the unique equilibrium of that subgame is (U, L). Therefore, the unique SPE of the game overall is (IU, L), which yields the same outcome as the WPBE

In the game in Figure 16, 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 13 serves as a reminder that the situation is subtle. In that game, (DA, A) can be supported as a WPBE (trivally, since the game is one of perfect information), but it is not admissible.

Finally, consider the game in Figure 3.2, which is similar to Figure 17 but differs from it in an important respect. 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). Since all information sets are reached under (IU, L), it can be supported as WPBE; in particular, Bayes consistency requires $\mu(x) = 1$.

But the NE (OU, R), which is also subgame perfect, can also be be supported as a WPBE. If player 1 chooses O, then Bayes consistency does not require $\mu(x) = 1$, as strange as this may seem; because player 2's information set is not reached, μ is not restricted at x. (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

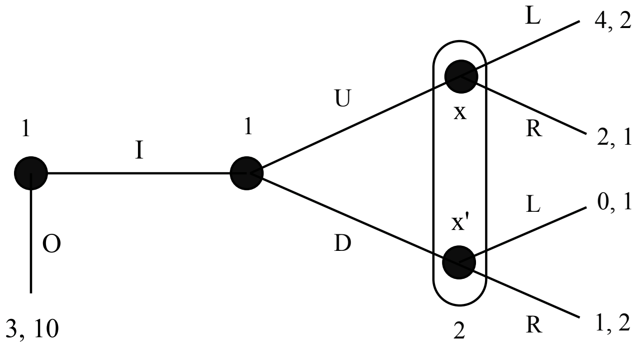

Figure 18: A game with a WPBE that is not subgame perfect.

you enter, I will assume that you have played D (even though this contradicts the stated strategy OU) and respond with R."

Remark 3. Do not to read too much into the intuition just given. I provided payoffs to make player 2's "belief threat" in the (OU, R) equilibrium seem more plausible. But even if I change the payoffs from O to $(3, -10)$, (OU, R) is still supported as a WPBE, even though player 2 now no longer has any incentive to bring it about. □

3.3 Perfect Bayesian Equilibrium

As illustrated by the (OU, R) Nash equilibrium for the game in Figure 3.2, implausible NE can be supported as WPBE with implausible μ. This motivates further restrictions on μ. Informally, an assessment is a *Perfect Bayesian Equilibrium (PBE)* iff it is a WPBE and μ reflects Bayesian updating "as much as possible" (even at information sets off the play path). The term PBE was coined in Fudenberg and Tirole (1991b), which provides a formalization for a specific class of games. In application, PBE tends to be invoked informally, in a "this equilibrium looks unreasonable and must therefore not be a PBE" kind of way. See Watson (2017) for one formalization of a general definition of PBE.

Rather than go further down this rabbit hole, let me note that every version of PBE in the literature (that I know of) implies that the following condition, which I call "subgame perfect WPBE (or SP-WPBE)." I am not aware of a standard name.

Definition 7. *(σ, μ) is a* subgame WPBE (SP-WPBE) *iff it induces a WPBE in every subgame.*

In particular, if σ can be supported as a SP-WPBE, then it is an SPE. This is enough to knock out the (OU, R) equilibrium in Figure 3.2.

3.4 Sequential Equilibrium

Sequential equilibrium, introduced in Kreps and Wilson (1982), predates WPBE and PBE/SP-WPBE. Conceptually, however, it is a strengthening of PBE.

Say that an assessment (σ, μ) is fully mixed if σ is fully mixed. If an assessment is fully mixed then every information set is on the play path and hence Bayes consistency implies that

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

for every information set h and every decision node $x \in h$.

Definition 8. *Given an assessment (σ, μ), μ is* consistent *iff there exists a sequence of fully mixed, Bayes consistent assements $((\sigma_t, \mu_t))$ such that $\lim_{t\to\infty}(\sigma_t, \mu_t) = (\sigma, \mu)$.*

For motivation, recall that, given σ, ordinary probability calculations pin down μ at information sets on the play path. The question is how to specify μ at information sets off the play path. Under Bayes consistency, the answer is: anyway you want. Under consistency, one must justify μ off the play path as being the limit of the conditional probabilities that would be generated if players fully mixed because they "trembled" (deviated slightly from equilibrium play).

Consistency is weak, by design, in that it requires that μ be the limit of *some* sequence (μ_t). Another sequences (μ_t), generated by a different sequence of "trembles," could converge to a different μ or might fail to converge altogether. The point of view is that if μ fails to be consistent then it is implausible. But μ might still be implausible even if it is consistent. I discuss the plausibility of μ further, but from a somewhat different point of view, in Section 3.6.

Definition 9. *(σ, μ) is a* sequential equilibrium (SE) *iff it is (a) consistent and (b) sequentially rational at every information set.*

It is easy to see that if (σ, μ) is an SE, then it is an SP-WPBE and hence WPBE and SPE. Consistency implies Bayes consistency and hence a consistent μ is pinned down by σ at information sets on the play path; for such information sets, you don't need to check sequences of fully mixed σ_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.

As illustration, consider the game of Figure 3.2. As already noted, (OU, R) is a NE that can be supported as a WPBE provided $\mu(x) \leq 1/2$. It cannot be supported, however, as an SE. Consider any sequence of fully mixed (behavior) strategies for player 1, σ_{1t}, converging to σ_1. Then σ_{1t} puts probability, say, ε_t on I

and probability η_t on D, with $\varepsilon_t, \eta_t \to 0$. The probability of being at x conditional on being in player 2's information set is

$$\mu_t(x) = \frac{\varepsilon_t(1-\eta_t)}{\varepsilon_t(1-\eta_t) + \varepsilon_t \eta_t} = 1 - \eta_t$$

which converges to 1. Thus, under the (OU, R) strategy profile, consistency requires $\mu(x) = 1$. But if $\mu(x) > 1/2$, then R is not sequentially rational for player 2, implying that (OU, R) cannot be supported as a SE.

In the last example, (OU, R) is also eliminated by subgame perfection, and hence by SP-WPBE. For an example in which SE is stronger than SP-WPBE, consider the game in Figure 19.

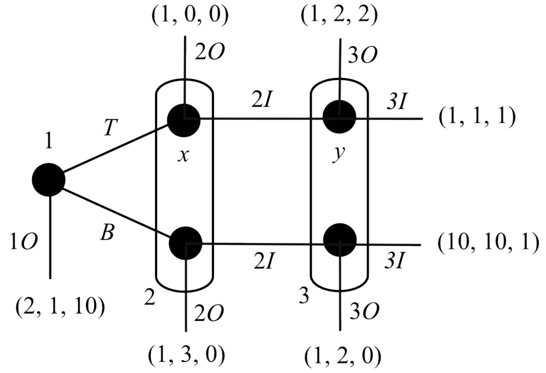

Figure 19: A game with an SP-WPBE that is not an SE.

There is a NE in which player 1 chooses action B, player 2 goes in (action $2I$) and player 3 goes in (action $3I$). Since all information sets are reached under this profile, this NE can automatically be supported as a SE, and hence a SP-WPBE etc.

However, there is also a NE in which player 1 goes out (action $1O$), player 2 goes out, and player 3 goes out. This NE can be supported as an SP-WPBE provided $\mu(x) \leq 1/3$ while $\mu(y) \geq 1/2$. Any such μ is not consistent; consistency requires that $\mu(x) = \mu(y)$. Therefore, this NE cannot be supported as an SE. Informally, the issue is that consistency requires that not only do players 2 and 3 have a coherent theory (the σ_{1t}) about how player 1 might have deviated, but they must have the *same* theory. SE is, in this respect, quite demanding.

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 σ_t^i is an approximate best response to σ_t^{-i}: there is no claim that σ can be approximated by a sequence of ε-perfect equilibria. Because of this, SE does *not* imply admissibility.

For example, consider again the game of Figure 11. (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 13, (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 σ is proper in a strategic form then it is supported as an SE in every extensive form that generates that (reduced) strategic form; the result was first proved in a working paper version of Kohlberg and Mertens (1986). The converse is not true, since properness implies admissibility, while SE does not.

3.5 Trembling Hand Perfection in the Agent Normal Form

Sequential equilibrium (SE) is closely related to another refinement, *trembling hand perfection in the agent normal form* (THP-ANF), introduced, along with THP, in Selten (1975). I discuss it for completeness and because it sheds some additional light on sequential equilibrium; think of this section as parenthetical. In applied work, game theorists almost invariably use SE or PBE rather than THP-ANF.

Given a game in extensive form, construct the *agent normal form* as the normal (e.g., strategic) form for a modified game in which each player is split into different "agents," one agent for each of the player's information sets; all of a player's agents receive the same payoff. A strategy profile σ in the original extensive form game is THP-ANF iff it corresponds to a strategy profile that is THP in the agent normal form.[4]

THP-ANF is stronger than SE in the sense that if an equilibrium σ is THP-ANF, then it can be supported as an SE. Explicitly, a sequence of fully mixed ε-perfect equilibria (σ_ε), with $\sigma_\varepsilon \to \sigma$, induces a sequence of fully mixed, Bayes consistent assessments $((\sigma_\varepsilon, \mu_\varepsilon))$. By compactness, μ_ε has a convergent subsequence with limit, say, μ. μ, in turn, is consistent in the SE sense. By the definition of ε-perfection and continuity of expected payoff, σ is sequentially rational with respect of μ.

Moreover, THP-ANF rejects (U, R) in the game of Figure 11 even though this NE can be supported as an SE. Thus, THP-ANF is stronger than SE. But SE and THP-ANF are equivalent for generic terminal node payoffs (Kreps and Wilson (1982)), a condition violated in Figure 11. 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 11, it is tempting to conjecture that THP-ANF implies THP/admissibility. This is false. The game of Figure 13 is once again a counterexample: the inadmissible equilibrium (DA, A) is THP-ANF. Informally, the reason is that, under THP-ANF, player 1 can be just as worried about his "agent"

[4] An equivalent characterization, which appears in Selten (1975) section 12, uses perturbations of behavior strategies.

trembling at player 1's second information set as he is about player 2 trembling.

In summary, THP does not imply THP-ANF (since a THP does not have to be subgame perfect; see Figure 8) and THP-ANF does not imply THP (since a THP-ANF does not have to be admissible).

3.6 The Cho-Kreps Intuitive Criterion

In the definition of sequential equilibrium, payoffs do not play a direct role in the restriction on beliefs called consistency. A subsequent literature, pioneered by McLennan (1985), discusses how payoff considerations can lead to additional belief restrictions, and potentially more powerful equilibrium refinements. In this section, I discuss a well known refinement along these lines, the Intuitive Criterion of Cho and Kreps (1987).

The Intuitive Criterion applies to a special class of games called signaling games. By a **basic signaling game**, I mean a game of the following form.

1. Nature chooses a **type** $t \in T$ for player 1, typically called the **sender**.

2. The sender sees their type and chooses an action, typically called a **message**. The set of available messages can depend on type.

3. The game is either over at this point and players receive their payoffs, or player 2, the **receiver**, can respond with an action, typically called a **response**.

4. If player 2 can respond, then player 2 can condition on player 1's message but not on player 1's type.

5. After player 2's response, the game is over and players receive their payoffs.

As a simple example, consider the signaling game of Figure 20. There are two Nash equilibrium *outcomes*.

One NE outcome corresponds to the NE $((I, O), T)$, where (I, O) is read I if t_1 and O if t_2. This is called a **separating** NE/outcome: the two types of player 1 choose different messages, implying that player 2 can deduce player 1's type from player 1's message.

The other NE outcome is (O, O) and is generated by NE of the form $((O, O), q)$, where q is the probability that player 2 plays T and satisfies $q \leq 7/10$ (so that O is sequentially rational for player 1). This is called a **pooling** outcome: the two types of player 1 are "pooling" on the single action/message O.

All of these NE can be supported as WPBE and, in fact, sequential equilibria. In the case of the separating NE, this is trivial since the only information set for the receiver is on the play path. In the case of the pooling NE, we require that the receiver play B with probability at least $3/10$. B is sequentially rational if $\mu(x) \leq 1/3$. Any such μ can be generated by a sequence of fully mixed strategies in which t_2 is sufficiently more likely to play I than t_1.

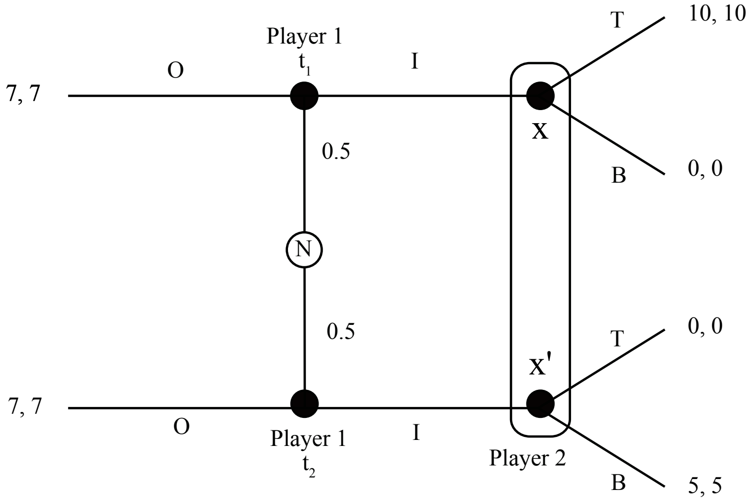

Figure 20: A Simple Signaling Game.

Nevertheless, a case can be made that the pooling outcome is implausible. The reason, informally, is that the only type with a possible motivation to play I is t_1, which contradicts the story just given for supporting the pooling outcome. More explicitly, type t_2 never wants to play I because they are getting 7 from O and could get only 5 at best from I. On the other hand, type t_1 *might* want to play I if the receiver were to respond with T. Moreover, T *would* be sequentially rational for the receiver if the receiver thought that type t_2 would indeed never play I, hence $\mu(x) = 1$.

The Intuitive Criterion of Cho and Kreps (1987) formalizes and generalizes the ideas just illustrated. To define the Intuitive Criterion, I first introduce some auxiliary concepts.

Definition 10. *Fix a NE outcome of a basic signaling game. Message m is* **equilibrium dominated** *for type t iff m is feasible for t and t's payoff under the NE outcome is strictly greater than t's maximum payoff from playing m, where the maximum is over all possible responses by the receiver.*

At an information set h for receiver, call a response r **never sequentially rational** iff, for every belief μ, r is not sequentially rational at h.

Definition 11. *Fix a NE outcome. Modify the game by deleting, at each information set, any responses that are never sequentially rational at that information set. The NE outcome* **fails the Intuitive Criterion** *iff, in the modified game, the following is true. There is an unused message m and a type t such that*

1. *message m is* **not** *equilibrium dominated for type t,*

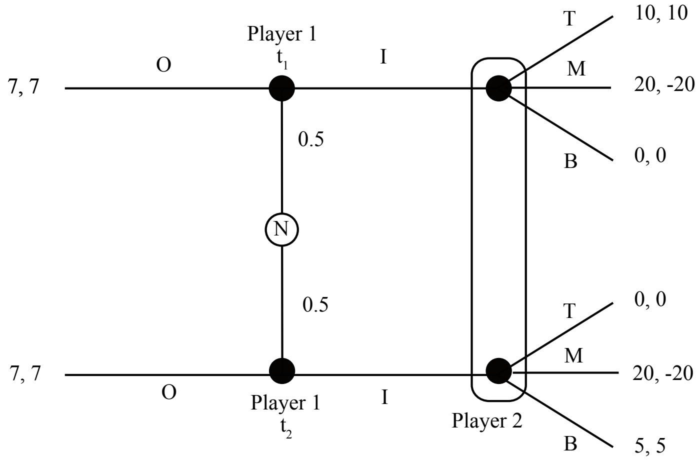

Figure 21: Why the Intuitive Criterion uses a Modified Game.

2. *type t's equilibrium payoff is less than the minimum payoff that t can get from playing m, where the minimum is across all sequentially rational responses to m, for all beliefs μ that, at the information set generated by message m, assign zero probability to (the decision node corresponding to) any type for whom message m* **is** *equilibrium dominated.*

An equilibrium outcome satisfies the Intuitive Criterion (i.e., does not fail it) vacuously iff, in the modified game, any unused message is either equilibrium dominated for every type or not equilibrium dominated for any type.

Consider again the pooling NE outcome of the example.

- There is no response that is never sequentially rational. Therefore, the modified game is the original game.
- In the pooling NE outcome, message I is unused. Consider t_1.
  - I is not equilibrium dominated for t_1.
  - I is equilibrium dominated for the only other type, t_2. If μ assigns probability zero to t_2, then $\mu(x) = 1$. The only sequentially rational response to this μ is T. For this response, t_1 gets a higher payoff from I than from the NE outcome.
- Therefore, the pooling NE outcome fails the Intuitive Criterion.

Figure 21 shows why the game modification step is used in the definition of the Intuitive Criterion. The game is essentially the same as Figure 20 but there is an added response, M, for player 2. M is never sequentially rational but if it is not deleted then I is *not* equilibrium dominated for t_2 (since $20 > 7$), in which case, without the game modification, the pooling outcome *would* satisfy the Intuitive Criterion.

The focus on NE *outcomes* rather than on strategy profiles or assessments mirrors, by design, the approach in Kohlberg and Mertens (1986), mentioned briefly in Section 2.6. One of the main motivations for Cho and Kreps (1987) is to understand what K-M Stability implies for signaling games. Loosely, in the context of two-player signaling games, if an outcome is K-M Stable then it satisfies the Intuitive Criterion. Cho and Kreps (1987) goes on to consider other refinements that are also implications of K-M Stability.

References

Ben-Porath, E. and E. Dekel (1992): "Signaling Future Actions and the Potential for Self-Sacrifice," *Journal of Economic Theory*, 57, 36–51.

Brandenburger, A. (2008): "Epistemic Game Theory: An Overview," in *The New Palgrave Dictionary of Economics*, ed. by L. E. B. Steven N. Durlauf, New York: Mcmillan, second ed.

Cho, I.-K. and D. Kreps (1987): "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102, 179–221.

Fudenberg, D., D. K. Levine, and D. Kreps (1988): "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, 44, 351–380.

Fudenberg, D. and J. Tirole (1991a): *Game Theory*, Cambridge, MA: MIT Press.

——— (1991b): "Perfect Bayesian and Sequential Equilibrium," *Journal of Economic Theory*, 53, 236–260.

Govindan, S. and R. Wilson (2008): "Refinements of Nash Equilibria," in *The New Palgrave Dictionary of Economics*, Palgrave.

Harris, C., P. Reny, and A. Robson (1995): "The Existence of Subgame-Perfect Equilibrium in Continuous Games with Almost Perfect Information," *Econometrica*, 63, 507–544.

Jackson, M., T. Rodriguez-Barraquer, and X. Tan (2012): "Epsilon-Equilibria of Perturbed Games," *Games and Economic Behavior*, 75, 198–216.

Kandori, M., G. Mailath, and R. Rob (1993): "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 61, 29–56.

Keisler, H. J. and B. S. Lee (2011): "Common Assumption of Rationality Preliminary Report," Department of Mathematics, University of Toronto.

Kohlberg, E. and J.-F. Mertens (1986): "On the Strategic Stability of Equilibria," *Econometrica*, 54, 1003–1037.

Kreps, D. and R. Wilson (1982): "Sequential Equilibria," *Econometrica*, 50, 863–894.

Levine, D. and J. Zheng (2010): "The Relationship of Economic Theory to Experiments," in *The Methods of Modern Experimental Economics*, ed. by G. Frechette and A. Schotter, Oxford University Press, washington University, St. Louis.

Mas-Colell, A., M. D. Whinston, and J. R. Green (1995): *Microeconomic Theory*, New York, NY: Oxford University Press.

McLennan, A. (1985): "Justifiable Beliefs in Sequential Equilibrium," *Econometrica*, 53, 889–904.

Mertens, J.-F. (1989): "Stable Equilibria - A Refomulation, Part I: Definition and Basic Properties," *Mathematics of Operations Research*, 14, 575–624.

Myerson, R. (1978): "Refinements of the Nash Equilibrium Concept," *The International Journal of Game Theory*, 7, 73–80.

——— (1991): *Game Theory*, Cambridge, MA: Harvard University Press.

Roth, A., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991): "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, 81, 1068–1095.

Selten, R. (1967): "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit," *Zeitschrift für die gesamte Staatswissenschaft*, 12, 301–324.

——— (1975): "A Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25–55.

van Damme, E. (1991): *Stability and Perfection of Nash Equilibria*, Springer, 2 ed.

Watson, J. (2017): "A General, Practicable Definition of Perfect Bayesian Equilibrium," UCSD.

Young, P. (1993): "The Evolution of Conventions," *Econometrica*, 61, 57–84.