Nonequilibrium Solution Concepts: Iterated Dominance and Rationalizability

Introduction

We began our preparation for the study of nonequilibrium solution concepts by introducing the notion of strategic dominance. We defined what it means for a strategy of one player to be dominated by another of her strategies. Because a rational player would never play a dominated strategy, we can sometimes use a dominance analysis to rule out some outcomes as possibilities when the game is played by rational players. In some games, e.g. the prisoners' dilemma, a dominance analysis leads to a unique prediction of the outcome when players are rational; we say that these games are dominance solvable. In other games, e.g. matching pennies, a dominance analysis results in no refinement of the set of possible outcomes. Other games lie between these two extremes: dominance analysis rejects some outcomes as impossible when the game is played by rational players but still leaves a multiplicity of outcomes.

Closely related to the concept of a strategy being dominated for a player is the idea that this strategy is "never a best response" for that player: No matter what beliefs she has about the actions of her opponents, she could not rationally choose to play that strategy. If a strategy is dominated, it can never be a best response. However, it is not obvious that the implication holds in the reverse direction. I.e. it's not obvious that a strategy which is never a best response is also a dominated strategy. Therefore the set of strategies which are never best responses is weakly larger than the set of dominated strategies. An analysis based on whether strategies are possibly best responses *does* exhaust the implications of all players being rational: A strategy cannot be plausibly chosen by a rational player if and only if it is never a best response.

We have seen that in two-player games a strategy is never a best response if and only if it is dominated. For two-player games, then, a dominance analysis fully exploits the assumption that all players are rational. However, for games with three or more players, it is possible that an undominated strategy will yet never be a best response. Therefore we can sometimes rule out as a plausible choice a strategy even when it is undominated. For more-than-two-player games, then, a dominance argument need not fully exploit the assumption that all players are rational.

We can often make sharper predictions about the possible outcomes of a game if we are willing to make stronger assumptions. Up until now we have assumed that the players are rational but we haven't even assumed that each knows that the others are rational. Beyond that we could further assume that each player knows that the other players know that the others are all rational. We could continue adding additional layers of such assumptions *ad nauseam*. Fortunately we can summarize the entire infinite hierarchy of such assumptions by simply saying that the rationality of the players is *common knowledge*. Rationality constrains players to choose best responses to their beliefs but does not restrict those beliefs. Common knowledge of rationality imposes a consistency requirement upon players' beliefs about others' actions.

By assuming that the players' rationality is common knowledge, we can justify an iterative process of outcome rejection—the *iterated elimination of strictly dominated strategies*—which can often sharpen our predictions. Outcomes which do not survive this process of elimination cannot plausibly be played when the rationality of the players is common knowledge. A similar, and weakly stronger, process—the *iterated elimination of strategies which are never best responses*—leads to the solution concept of rationalizability.[1] The surviving outcomes of this process constitute the set of *rationalizable* outcomes. Each such outcome is a plausible result—and these are the only plausible results—when the players' rationality is common knowledge. In two-player games the set of rationalizable outcomes is exactly the set of outcomes which survive the iterated elimination of strictly dominated strategies. In three-or-more-player games, the set of rationalizable outcomes can be strictly smaller than the set of outcomes which survives the iterated elimination of strictly dominated strategies. In a rationalizable outcome players' beliefs about the same question can differ—and hence some are incorrect; and a player can find—after the others' choices are revealed—that she would have preferred to have made a different choice.

Recapitulation

Let's briefly review the standard paradigm and notation. We have a finite set I of n players, $I = \{1, \ldots, n\}$. The finite pure-strategy space for player i is S_i; her mixed-strategy space is $\Sigma_i \equiv \Delta^{\#S_i - 1}$; typical elements are $s_i \in S_i$ and $\sigma_i \in \Sigma_i$. When player i chooses the mixed strategy σ_i, the probability with which she plays the pure strategy $s_i \in S_i$ is $\sigma_i(s_i)$. When we omit the subscript on a set defined for each player, we mean the Cartesian product of all the player sets: e.g., $S \equiv \mathsf{X}_{i \in I}$. A subscript "$-i$" means "$I \backslash \{i\}$". An element $\boldsymbol{s}_{-i} \in S_{-i}$ is called a deleted pure-strategy profile. Player i's von Neumann-Morgenstern utility function is u_i: $S \to \mathbb{R}$. (We often abuse notation and consider this utility function to

[1] "Weakly stronger" seems a little oxymoronic!

take a mixed-strategy profile as its argument instead of a pure-strategy profile. In such a case it represents the player's expected utility when the players randomize independently according to their component mixed strategies in the mixed-strategy profile.)

We began our preparation for the study of nonequilibrium solution concepts by introducing the notion of strategic dominance. Let $\sigma_i, \sigma_i' \in \Sigma_i$ be two mixed strategies for player i. We say that σ_i' strictly dominates σ_i if σ_i' gives player i a strictly higher expected utility than does σ_i for every possible deleted pure-strategy profile s_{-i} which her opponents could play, i.e. if[2]

$$\forall \boldsymbol{s}_{-i} \in S_{-i}, \square\, u_i(\sigma_i', \boldsymbol{s}_{-i}) > u_i(\sigma_i, \boldsymbol{s}_{-i}). \tag{1}$$

Iterated strict dominance

We saw above that in some games, e.g. the Prisoners' Dilemma, each player has a dominant strategy and we could therefore make a very precise prediction about the outcome of the game. To achieve this conclusion we only needed to assume that each player was rational and knew her own payoffs. We also saw an example, viz. matching pennies, where dominance arguments got us nowhere—no player had any dominated strategies. There are games which lie between these two extremes in the degree to which and ease with which dominance arguments can refine the set of possible outcomes.

The technique we'll discuss now is called the *iterated elimination of strictly dominated strategies*.[3] In order to employ it we will need to make stronger informational assumptions than we have up until now. For example, we won't merely assume that each player is rational. We might need to assume as well, in a two-payer game for example, that player 1 *knows* that player 2 is rational; and player 2 knows that player 1 knows that player 2 is rational, etc. In some games application of the iterated elimination of strictly dominated strategies can require that these hierarchies of beliefs about beliefs be quite deep.

Consider a two-player game between Row and Column, whose pure-strategy spaces are S_R and S_C, respectively. Prior to a dominance analysis of a game, we know only that the outcome will be one of the strategy profiles from the space of strategy profiles $S = S_R \times S_C$. We reasoned above that a rational player would never play a dominated strategy. If Row has a dominated strategy, say $\tilde{s}_R$, but Column does not, then Row, being rational, would never play this strategy. We could therefore confidently predict that the outcome of the game must be drawn from the smaller space of strategy profiles

$$S' = (S_R \backslash \{\tilde{s}_R\}) \times S_C. \tag{8}$$

Here is the interesting point and the key to the utility of the iterative process we're developing: Although Column had no dominated strategy in the original game, he may well have a dominated

---

2 We showed that satisfaction of this condition was equivalent to satisfaction of the same condition but with the substitution of $\boldsymbol{\sigma}_{-i} \to \boldsymbol{s}_{-i}$ and $\Sigma_{-i} \to S_{-i}$. I.e. without loss of generality, in order to assess questions of dominance for player i, we can restrict attention to deleted pure-strategy profiles by her opponents.

3 See Fudenberg and Tirole [1991].

strategy $\tilde{s}_C$ in the new, smaller game S'.[4] If so, and if we make sufficient assumptions, we can rule out as possible outcomes all which involve such newly dominated strategies from Column's strategy space; this again results in a smaller space of strategy profiles. And in this smaller game additional row strategies may now be revealed to be dominated. In some cases this process can continue until a unique strategy for each player survives this elimination process. In this case we say—as we did when each player had a dominant strategy in the original game—that the game is *dominance solvable*.

Common knowledge of rationality

I just said that we had to make assumptions to justify the deletion of Column's dominated strategy $\tilde{s}_C$. What assumptions are necessary for this step? First, Column must be rational. Additionally, in order for Column to see that $\tilde{s}_C$ is dominated for him, he must see that Row will never play $\tilde{s}_R$. Row will never play $\tilde{s}_R$ if she is rational; therefore we must assume that Column *knows* that Row is rational. With these additional assumptions we can confidently predict that any outcome of the game must be drawn from

$$S'' = (S_R \backslash \{\tilde{s}_R\}) \times (S_C \backslash \{\tilde{s}_C\}). \tag{9}$$

Let's not get too tedious, but let's carry this out one more level. It may be the case that in the game defined by the strategy-profile space S'' there is now a strategy of Row's which is newly dominated, call it $\hat{s}_R$. However, we can't rule out that Row will play $\hat{s}_R$ unless we can assure that Row knows that the possible outcomes are indeed limited to S'', i.e. that Column will not choose $\tilde{s}_C$. Column won't choose $\tilde{s}_C$ if he is rational and knows that Row is rational. Therefore we must assume that Row knows that Column is rational and knows that Column knows that Row is rational.

In any finite game this chain of assumptions can only be usefully carried out to a finite depth. To ensure that we can make such assumptions to an arbitrary depth we often make a convenient assumption: that it is *common knowledge* that all players are rational.[5]

What does it mean for something to be common knowledge?[6] Let $\mathcal{P}$ be a proposition, e.g. that "player 1 is rational." If $\mathcal{P}$ is common knowledge, then

> Everyone knows $\mathcal{P}$;
> Everyone knows that (Everyone knows $\mathcal{P}$);
> Everyone knows that [Everyone knows that (Everyone knows $\mathcal{P}$)];
> Etc.

---

4 The new game is defined by the strategy spaces $S_R' = S_R \backslash \{\tilde{s}_R\}$ and $S_C' = S_C$ and by the utility functions $u_i'\colon S' \to \mathbb{R}, i \in \{R, C\}$, where each u_i' is the *restriction* of $u_i\colon S \to \mathbb{R}$ to the smaller domain S', viz. $\forall \boldsymbol{s} \in S'$, $u_i'(\boldsymbol{s}) = u_i(\boldsymbol{s})$.

5 When we predicted that Row would not play $\tilde{s}_R$ we implicitly assumed that Row knew her own payoffs but not that Row knew Column's payoffs. When we predicted that Column would never play $\tilde{s}_C$, we implicitly assumed that Column knew that Row knew Row's payoffs *and* that Column knew Row's payoffs. So if we wanted to be perfectly explicit about our assumptions about the players' knowledge of their payoffs, there's another whole hierarchy of beliefs to merge into the hierarchy of beliefs concerning rationality. We can summarize this hierarchy by assuming that "the players' payoffs are common knowledge."

6 Aumann [1976] gives the definition for two players; Myerson [1991] and Pearce [1984] provide it for many players.

In other words, if $\mathcal{P}$ is common knowledge, then every statement of the form

$$\text{(Everyone knows that)}^k \text{ everyone knows } \mathcal{P}, \tag{10}$$

is true for all $k \in \{0, 1, 2, \dots\}$.

Prima facie, it might seem unreasonable to expect real-life players to be able to keep track of such unbounded hierarchies of beliefs about beliefs. Actually, common knowledge can be achieved rather simply. Suppose you and I are both present when an event E (assumed so salient that it cannot be ignored) occurs. Suppose further that then we both immediately make eye contact with one another. It is then common knowledge (between the two of us) that E occurred.[7]

Example: Solving a game by the iterated elimination of strictly dominated strategies.

Consider the game in Figure 5(a). Note that Column has no strategies which are dominated by pure strategies. Note further that none of Column's pure strategies can be dominated by a mixture of the other two. However, for Row, Up dominates Middle, and Down is undominated. So a rational Row would never play Middle. Therefore if

$$\text{Row is rational}, \tag{11}$$

then we can strike Middle from Row's strategy space and collapse the bimatrix, resulting in the game in Figure 5(b).

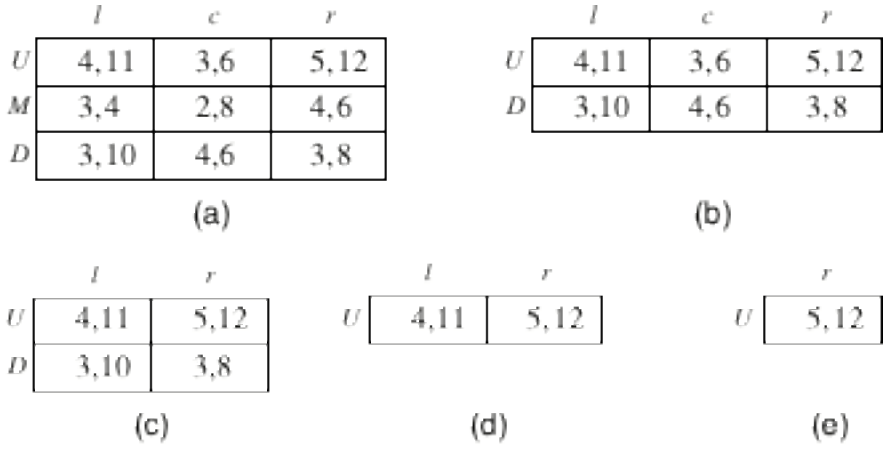

Figure 5: Solving a game by the iterated elimination of strictly dominated strategies.

If Column knows that the possible outcomes must belong to the game in Figure 5(b)—which requires that he know (11), then he would see that center is dominated by both left and right. (Left and right are both undominated.) Therefore if in addition to (11) both

$$\text{Column is rational}, \tag{12}$$

7 For example, let E be the event that two particular people are in the same room at a particular time.

Column knows Row is rational, (13)

we can strike center, which results in the game of Figure 5(c).

If Row sees that the possible outcomes must be drawn from the game in Figure 5(c) , which requires that she know (12) and (13), then she would see that Up dominates Down. Therefore if also

Row knows Column is rational, (14)

Row knows Column knows Row is rational, (15)

we could strike Down, which results in the game of Figure 5(d).

If Column knows that Figure 5(d) is the relevant game, which requires that he also knows (14) and (15), then Column would recognize that right dominates left. Hence, if also

Column knows Row knows Column is rational, (16)

Column knows Row knows Column knows Row is rational, (17)

the only strategies which survive the iterated elimination of strictly dominated strategies are Up for Row and right for Column. This strategy profile is shown in Figure 5(e). So we see that the *necessary* assumptions to solve this game were (13) $\rightarrow$ (17). (We always assume that the players are rational. What is new here are the assumptions about the players' higher-order beliefs about rationality.) All of these assumptions about beliefs are implied by the sufficient assumption that it is common knowledge that both players are rational.

Example: Iterated strict dominance can require mixed-strategy domination.

I wanted to make the logical reasoning in the previous example as transparent as possible, so I chose the game such that elimination required only domination by pure strategies. However, a more general mixed-strategy analysis can be necessary. Consider the game in Figure 6.

	l	r
U	6,4	0,2
M	0,3	6,1
D	2,1	2,4

Figure 6: A mixture of Up and Middle dominates Down; then left dominates right.

There are no pure-strategy dominance relationships in the original game. However, the mixed strategy $\frac{1}{2} \circ U \oplus \frac{1}{2} \circ M$ dominates Down. After deleting Down, left dominates right for Column. After deleting right, Up dominates Middle. Therefore the only possible outcome under common knowledge of rationality is (U, l).

Iterated strict dominance: formal definition

In order to support rigorous proofs of later claims regarding the surviving outcomes of the iterated elimination of strictly dominated strategies we'll now define this solution concept more formally. The process is iterative; we can think of it as an algorithm (which is depicted in flowchart form in Figure 7).[8]

Very loosely to begin with…. We start with the original game S. We delete all the dominated strategies for each player, which results in a smaller game S^1. More generally, we consider the game defined by some set of not-yet-rejected outcomes S^{t-1}. By rejecting any player's strategy which is dominated, we reach the weakly smaller game S^t.[9] (Therefore S_i^t is the set of player i's pure strategies which are undominated in the game S^{t-1}.) When we reach a point where the resulting game cannot be further shrunk by the elimination of strictly dominated strategies, then our process has concluded. We say that this set of outcomes, denoted S^∞, has survived the iterated elimination of strictly dominated strategies.

More formally now…. We use $t \in \mathbb{Z}_+ \equiv \{0, 1, 2, \dots\}$ as a counter. We denote by $S_i^t \subset S_i$ the set of player-i pure strategies which are unrejected after t rounds of this iterative procedure. Therefore the period-0 game is just the original game: $\forall i \in I, S_i^0 = S_i$. We denote by $\Sigma_i^t \subset \Sigma_i$ the set of player-i mixed strategies which are mixtures only over the pure strategies which are unrejected after t rounds; i.e. $\Sigma_i^t = \{\sigma_i \in \Sigma_i : \operatorname{supp} \sigma_i \subset S_i^t\}$.[10,11] In particular, $\Sigma_i^0 = \Sigma_i$.

Now consider the game resulting after $t-1$ rounds of elimination, viz. $S^{t-1} = \mathsf{X}_{i \in I} S_i^{t-1}$. Consider some player $i \in I$ and consider each of her not-yet-rejected pure strategies $s_i \in S_i^{t-1}$. The strategy s_i is dominated in the game S^{t-1} if there exists a mixed strategy σ_i over these not-yet-rejected pure strategies S_i^{t-1}, i.e. $\sigma_i \in \Sigma_i^{t-1}$, such that σ_i dominates s_i in the game S^{t-1} (i.e. for all deleted strategy profiles $\boldsymbol{s}_{-i} \in S_{-i}^{t-1}$ by her opponents). Therefore the set S_i^t of strategies which are undominated in S^{t-1} is

$$S_i^t = \{s_i \in S_i^{t-1} : \nexists \sigma_i \in \Sigma_i^{t-1}, \forall \hat{\boldsymbol{s}}_{-i} \in S_{-i}^{t-1}, u_i(\sigma_i, \hat{\boldsymbol{s}}_{-i}) > u_i(s_i, \hat{\boldsymbol{s}}_{-i})\}. \tag{18}$$

---

8 Just a note of explanation about the "$i \in I$ loop" → "end i loop" constructions in the flowchart for those who haven't programmed computers: When an "$i \in I$ loop" box is initially encountered from above, i is set to the first element of I and control passes downward. When the "end i loop" box is encountered, one of two things can happen. 1 If there are still elements $i \in I$ which haven't been processed in this loop, control returns to the preceding "$i \in I$ loop" and the i counter is incremented to the next element of I. 2 If this was the last element i in the set I, then control drops out of the loop to the next lower box.

9 The previous round of eliminations may have revealed newly dominated strategies for one or more players.

10 Note that we are not at this point requiring that the mixed strategies in Σ_i^t be undominated.

11 Note that Σ_i^t is not a mixed strategy for player i in her smaller strategy space S_i^t. Rather it is a mixed strategy over her original strategy space S_i but whose support includes only pure strategies in the smaller strategy space S_i^t.

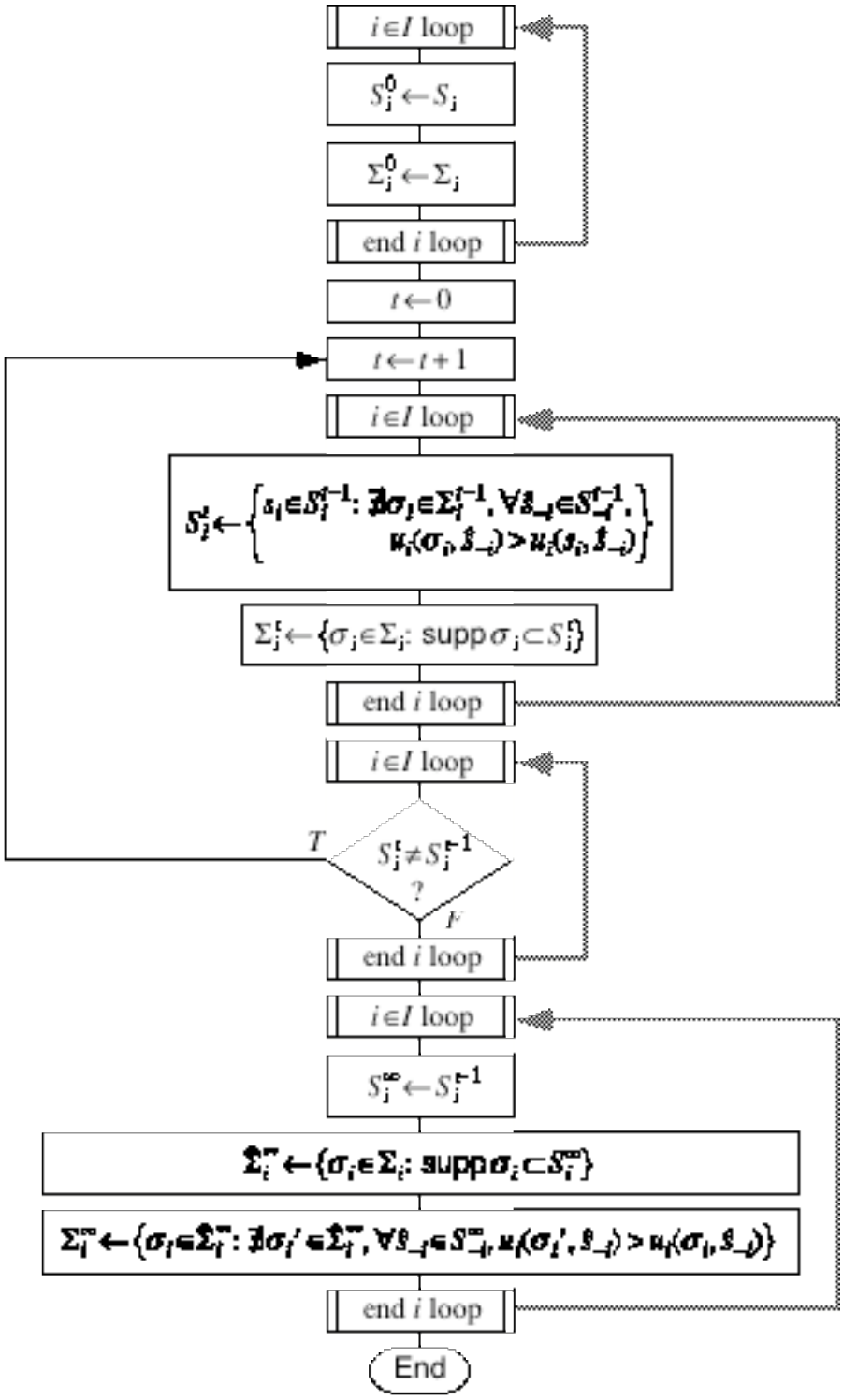

Figure 7: A flowchart definition of the iterated elimination of strictly dominated strategies.

We then form the set of mixed strategies

$$\Sigma_i^t = \{\sigma_i \in \Sigma_i : \operatorname{supp} \sigma_i \subset S_i^t\} \tag{19}$$

which put weight only upon the still-admissible pure strategies of S_i^t; these constitute our arsenal with which to try to dominate strategies in the new game S^t. We note that the sequence of player-i strategy spaces $\{S_i^t\}_{t \in \mathbb{Z}_+}$ is nested:

$$\forall t \in \mathbb{Z}_+,\ S_i^{t+1} \subset S_i^t. \tag{20}$$

We must eventually get to a stage in the iterations such that the surviving strategy set for each player is unchanged from the previous round, i.e. $\exists \tau \in \mathbb{N} \equiv \{1, 2, \ldots\}, \forall i \in I,\ S_i^\tau = S_i^{\tau-1}$. (Otherwise at least one pure strategy from at least one player must be eliminated in each round. The number of pure strategies for each player is finite. Therefore it's impossible to remove strategies forever.) If we were to continue the algorithm beyond this stage, we'd find that the strategy sets remained unchanged, i.e. for all $i \in I$, $S_i^\tau = S_i^{\tau+1} = S_i^{\tau+2} = \cdots$. [This is clear from examination of (18). Consider any $i \in I$ and $s_i \in S_i^\tau$. This s_i was undominated in $S^{\tau-1}$ (which is why it survived to belong to S_i^τ). But $S^\tau = S^{\tau-1}$, so s_i must be undominated in S^τ as well, and therefore deserves membership in $S^{\tau+1}$. And so on for $S^{\tau+k}$, $k \in \mathbb{N}$.]

Once we reach a stage in this iterative process at which the strategy sets are no longer shrinking, say at period τ as in the above paragraph, we have exhausted the implications of an iterative dominance analysis for behavior in the game. We set, for each $i \in I$,

$$S_i^\infty = S_i^\tau, \tag{21}$$

where S_i^∞ is the set of player-i pure strategies which survive the iterated elimination of strictly dominated strategies.[12] When the game is played under the conditions of common knowledge of rationality, every player i would choose some strategy $s_i \in S_i^\infty$. The Cartesian product of these player-strategy sets, viz. $S^\infty = \mathsf{X}_{i \in I} S_i^\infty$ is the set of strategy profiles which survive the iterated elimination of strictly dominated strategies. When the game is played under the conditions of common knowledge of rationality, any pure-strategy profile must be within the set S^∞, i.e. $\boldsymbol{s} \in S^\infty$.

To determine the set of mixed strategies Σ_i^∞ for player i which are compatible with an iterated dominance analysis based on the common knowledge of rationality, we first find all the mixed strategies $\hat{\Sigma}_i^\infty$ which put positive weight only upon the unrejected pure strategies S_i^∞,

$$\hat{\Sigma}_i^\infty = \{\sigma_i \in \Sigma_i : \operatorname{supp} \sigma_i \subset S_i^\infty\}, \tag{22}$$

However, we have seen before that a mixed strategy which spreads all its weight only among undominated pure strategies can still be itself dominated.[13] Therefore we must filter this set of mixed

---

12 We could alternatively write that S_i^∞ is the intersection of the infinite sequence of player-i strategy spaces, viz. $S_i^\infty = \cap_{t \in \mathbb{Z}_+} S_i^t$.

13 See the example on pages 10–13 in the "Dominance" handout of September 7, 1993.

strategies to remove any which are dominated by other mixed strategies in that set. This results in the set Σ_i^∞ of mixed strategies for player i which are not rejected by an iterated dominance argument,

$$\Sigma_i^\infty = \{\sigma_i \in \hat{\Sigma}_i^\infty : \nexists \sigma_i' \in \hat{\Sigma}_i^\infty, \forall \boldsymbol{s}_{-i} \in S_{-i}^\infty, u_i(\sigma_i', \boldsymbol{s}_{-i}) > u_i(\sigma_i, \boldsymbol{s}_{-i})\}. \quad (23)$$

Example: A dominance-solvable two-player game

Let's perform iterated elimination of strictly dominant strategies on the game in Figure 8. U dominates M. With M removed, l dominates r. With r removed, D dominates U. Therefore the iterated-dominance outcome is (D, l).

	l	r
U	10,5	10,4
M	8,4	8,5
D	12,5	6,4

Figure 8: A dominance-solvable two-player game.

In terms of our formalism, the nested pure-strategy sets for each player at each stage of the iterative process are

$$S_R^0 = \{U, M, D\}, \qquad S_C^0 = \{l, r\},$$

$$S_R^1 = \{U, D\}, \qquad S_C^1 = \{l, r\}$$

$$S_R^2 = \{U, D\}, \qquad S_C^2 = \{l\},$$

$$S_R^3 = \{D\}, \qquad S_C^3 = \{l\},$$

$$S_R^\infty = \{D\}, \qquad S_C^\infty = \{l\}.$$

The sets of mixed strategies which survive the iterated elimination of strictly dominated strategies, viz. Σ_R^∞ and Σ_C^∞, are trivial in this example. Each player has only one surviving pure strategy and therefore the only mixture over that strategy is the corresponding degenerate mixed strategy, viz.

$$\Sigma_R^\infty = \{(0 \circ U \oplus 0 \circ M \oplus 1 \circ D)\}, \qquad \Sigma_C^\infty = \{(1 \circ l \oplus 0 \circ r)\}.$$

Iterated strict dominance: looking more closely

Let's revisit the formal specification of the iterated elimination of strictly dominated strategies to hunt for and hopefully resolve possibly problematic issues. Consider again the example from Figure 8. Why did we reject r for Column? Because we had previously rejected M for Row on account of being dominated by U. But we later rejected U itself for Row. Since we used U to reject M, but later decided that Row would never actually play U, perhaps we should call into question our rejection of M and therefore of r for Column. How can we think about issue more clearly?

Perhaps one justification for reconsidering our rejection of M would be: if, starting with our final game $S_R^{\infty} \times S_C^{\infty} = \{D\} \times \{l\}$, we reintroduced M into Row's strategy set and found that M was no longer dominated, then we might conclude that our rejection of M was mistaken—that we were mislead by the later-to-be-rejected-itself strategy U. However, if we perform that experiment we have the game of Figure 9, and we see that M is still dominated, albeit by D now rather than U.

	l
M	8,4
D	12,5

Figure 9: The dominance-solved game of Figure 8 with M restored to Row's strategy space.

So we see that, in this example at least, a strategy which was rejected as dominated in an early stage of the iterative process was still dominated when reinjected into its player's strategy space at the end of the iterative process, even though the originally dominating strategy which justified its rejection had been later itself rejected as dominated. We will now see that this is a general result: any strategy which is dominated at some stage of the iterative process would still be dominated at any later stage if reintroduced into its player's strategy space.

Let's first establish a relevant fact. For some player $i \in I$ consider the game $S_i \times S_{-i}^{\infty}$. We'll now see that S_i^{∞} is the set of player i's pure strategies which are undominated in the game $S_i \times S_{-i}^{\infty}$. To prove this equality we need to show that 1 S_i^{∞} contains all the strategies which are undominated in $S_i \times S_{-i}^{\infty}$ and 2 any strategy which is dominated in $S_i \times S_{-i}^{\infty}$ does not belong to S_i^{∞}.

The set S_i^{∞} contains all of player i's strategies which are never rejected during the iterative elimination process. Therefore to show that S_i^{∞} contains all the undominated strategies in $S_i \times S_{-i}^{\infty}$ we need to show that all of the strategies which are rejected during the iterative elimination process are in fact dominated in $S_i \times S_{-i}^{\infty}$. So we consider a rejected strategy $s_i \in S_i \backslash S_i^{\infty}$ and let $t \in \mathbb{Z}_+$ be the stage of the process in which it is rejected. I.e.

$$\exists \sigma_i \in \Sigma_i^t,\ \forall \hat{\boldsymbol{s}}_{-i} \in S_{-i}{}^t,\ u_i(\sigma_i, \hat{\boldsymbol{s}}_{-i}) > u_i(s_i, \hat{\boldsymbol{s}}_{-i}). \tag{24}$$

If s_i is dominated in $S_i \times S_{-i}^{\infty}$, then

$$\exists \sigma_i' \in \Sigma_i,\ \forall \hat{\boldsymbol{s}}_{-i} \in S_{-i}^{\infty},\ u_i(\sigma_i', \hat{\boldsymbol{s}}_{-i}) > u_i(s_i, \hat{\boldsymbol{s}}_{-i}). \tag{25}$$

We need to show that satisfaction of (24) implies satisfaction of (25). But we note from (19) and (20) that $\Sigma_i^t \subset \Sigma_i$ and $S_{-i}^{\infty} \subset S_{-i}{}^t$. Therefore in attempting to satisfy (25), compared to satisfying (24), we can choose a mixed strategy from a larger set Σ_i and dominance need hold only in fewer cases (viz. for all $\hat{\boldsymbol{s}}_{-i} \in S_{-i}^{\infty}$ rather than for all $\hat{\boldsymbol{s}}_{-i} \in S_{-i}{}^t$). Therefore if (24) is satisfied, we can set $\sigma_i' \leftarrow \sigma_i$ in order to satisfy (25). This shows 1.

Before we show 2, I need to state (and you can establish!) a useful result:

Theorem Let $S_i' \subset S_i$ be the set of player i's undominated pure strategies. Let $\underline{S}_i$ be the set of player i's dominated pure strategies. Let $\sigma_i \in \Sigma_i$ be a mixed strategy which puts positive weight on at least one dominated pure strategy; i.e. $\text{supp}\,\sigma_i \cap \underline{S}_i \neq \varnothing$.

Then there exists a mixed strategy $\sigma_i' \in \Sigma_i$ such that 1 σ_i' dominates σ_i and 2 σ_i' puts no weight on dominated pure strategies; i.e. $\text{supp}\,\sigma_i' \subset S_i'$.

Proof It's up to you. This is extra-credit challenge #1.

Now we want to show 2, i.e. that any strategy which is dominated in $S_i \times S_{-i}^{\infty}$ does not belong to S_i^{∞}. Therefore we need to show that any strategy which is dominated in $S_i \times S_{-i}^{\infty}$ is rejected at some stage $t \in \mathbb{Z}_+$ in the iterative elimination process. Let $\tau \in \mathbb{Z}_+$ be the "final" period of the process (i.e. $S_i^{\infty} = S_i^{\tau}$). I'll show that if s_i survives until the last stage, viz. until stage τ, then it will be rejected there. So assume $s_i \in S_i^{\tau}$. In order to be rejected in period τ we must have

$$\exists \sigma_i \in \Sigma_i^{\tau},\ \forall \hat{s}_{-i} \in S_{-i}^{\tau},\ u_i(\sigma_i, \hat{s}_{-i}) > u_i(s_i, \hat{s}_{-i}). \tag{26}$$

Because s_i is dominated in $S_i \times S_{-i}^{\infty}$; i.e. there exists a $\sigma_i' \in \Sigma_i$ such that (25) holds. If it's already the case that $\sigma_i' \in \Sigma_i^{\tau}$, then we're done. [Just take $\sigma_i \leftarrow \sigma_i'$ and note (21) to see that (26) is satisfied.] So consider the remaining case where $\sigma_i' \notin \Sigma_i^{\tau}$. Therefore σ_i' must put positive weight on some pure strategy which is dominated in $S_i \times S_{-i}^{\infty}$ (because S_i^{∞} contains all the undominated strategies in this game and therefore $\Sigma_i^{\infty} = \Sigma_i^{\tau}$ contains all the mixtures over the undominated strategies). Therefore we can use the above theorem to assert the existence of another mixed strategy $\sigma_i \in \Sigma_i^{\tau}$ which dominates σ_i' and therefore satisfies (26).

So we have shown that, for each player $i \in I$, the set of her pure strategies which survive the iterated elimination of strictly dominated strategies, viz. S_i^{∞}, is exactly the set of pure strategies which are undominated when her opponents can only choose deleted strategy profiles in S_{-i}^{∞}. Now that this is established you can construct an argument to show that, if any combination of rejected strategies $s_i \in S_i \backslash S_i^{\infty}$ is reinjected into player i's strategy space at the last stage of the iterated elimination process, it would still be rejected at that stage.

Rationalizability

We concluded earlier that a rational player would never play a dominated strategy (because she could do strictly better in all cases by choosing instead a strategy which dominated that strategy). By assuming that the rationality of all players is common knowledge, we developed the technique of the iterated elimination of strictly dominated strategies in order to determine a set of strategies for each player which survive this elimination procedure. We concluded that no player would ever choose a strategy outside of her surviving set. The implication here was one way: common knowledge of rationality implies that the game's outcome must survive the iterated elimination of strictly dominated strategies. We did *not* show that every surviving strategy could be reasonably chosen by a rational player.

An alternative expression of the common knowledge of rationality is closely related to our domination discussion: A rational player would never choose a strategy which is not a best response to some beliefs about opponents' choices. Further, a player's beliefs about the choices of others are constrained in that the other players must be believed to also be playing strategies which are best responses to their own beliefs. The *rationalizable* outcomes are those which survive the iterated elimination of strategies which are never best responses.[14] However, this argument *does* go both ways. Bernheim [1984] argues not only that a rational player must choose her strategies from their rationalizable set, but also that every strategy in this set can be consistently justified as a rational choice.

A rational player must choose a best response to her beliefs about the actions of the other players. For example, in Figure 10, E is dominated by D; therefore there is no possible belief which Row could hold about Column's strategy to which E would be a best response.[15] Therefore a rational Row would never play E.

Although a player's rationality constrains her action to be a best response to her beliefs, it does not restrict what her beliefs about others' actions can reasonably be. (After all, an irrational opponent might do anything.) A rational Row player could play B because it is a best response to z. But a rational Column would never play z, because it is dominated by y. If Row knew that Column is rational, Row would realize that Column would never choose z, and Row would further deduce that B is not a best response to anything Column might rationally do. (B is dominated by C when Column's strategy space is reduced to $\{w, x, y\}$.) Row's knowledge of Column's rationality restricts Row's beliefs to put zero weight on Column choosing z.

	w	x	y	z
A	**7**,**5**	−8,4	0,4	99,3
B	5,0	4,1	15,**9**	**100**,8
C	6,0	5,**8**	**20**,4	10,2
D	2,6	**7**,−10	3,**9**	10,8
E	1,6	2,−10	1,**7**	8,6

Figure 10

In summary, if Row and Column are both rational and if Row knows that Column is rational, then we can restrict our attention to the smaller game $\{A, C, D\} \times \{w, x, y\}$ shown in Figure 11. You can convince yourself that no further elimination of strictly dominated strategies is possible; hence (because this is a two-player game) all of the outcomes in this smaller game are rationalizable.[16]

---

14 See Fudenberg and Tirole [1991].

15 I have indicated in boldface type the row payoffs which are maximal in each column and the column payoffs which are maximal in each row.

16 Recall that in two-player games the rationalizable outcomes are exactly those which survive the iterated elimination of strictly dominated strategies. In three-or-more–player games the set of rationalizable outcomes is a weakly smaller set than those survivors of iterated elimination of strictly dominated strategies.

	w	x	y
A	**7**,**5**	−8,4	0,4
C	6,0	5,**8**	**20**,4
D	2,6	**7**,−10	3,**9**

Figure 11: The rationalizable subset of the game from Figure 10.

Rationalizability as a consistent system of beliefs

We defined the rationalizable outcomes as those which survived the iterated elimination of strategies which were never best responses. In order to focus explicitly on the constraints which common knowledge of rationality imposes upon players' beliefs I will now discuss rationalizability from a different perspective. Consider the strategy profile (C, x) in the game of Figure 11. I will show that this profile is rationalizable by showing that C and x are rationalizable strategies for Row and Column, respectively. To do this I will show that there exists a *consistent system of beliefs* for the players which justifies their choices—i.e. which shows that these choices do not conflict with the common knowledge of rationality assumption. (See Bernheim [1984].)

Let's establish some notation so that we can tractably talk about beliefs about beliefs about beliefs about…. Let $\mathcal{R}$ and $\mathcal{C}$ stand for the Row and Column players, respectively. If Row chooses A, we write $\mathcal{R}(A)$, and similarly for other choices by either player. If Column believes that Row will choose A, we write $\mathcal{C}\mathcal{R}(A)$. If Column believes that Row believes that Column will choose y, we write $\mathcal{C}\mathcal{R}\mathcal{C}(y)$, etc.

A rational Row player would play C, i.e. $\mathcal{R}(C)$, if she believes that Column will play y; i.e. if $\mathcal{R}\mathcal{C}(y)$. Is this belief by Row reasonable? Column would play y if he thought that Row would play D; therefore we assume $\mathcal{R}\mathcal{C}\mathcal{R}(D)$. Would Row do this? Row would play D if she thought that Column would play x; therefore we assume $\mathcal{R}\mathcal{C}\mathcal{R}\mathcal{C}(x)$. Finally, Column would choose x if he believes that Row would choose C, viz. $\mathcal{R}\mathcal{C}\mathcal{R}\mathcal{C}\mathcal{R}(C)$. But C is justified by the sequence of beliefs we have just described. We summarize the beliefs of Row's which justify her playing C:[17]

$\mathcal{R}\ (C)$	$\mathcal{R}$ plays C,	(27a)
$\mathcal{R}\ \mathcal{C}\ (y)$	$\mathcal{R}$ believes $\mathcal{C}$ will play y,	(27b)
$\mathcal{R}\ \mathcal{C}\ \mathcal{R}\ (D)$	$\mathcal{R}$ believes $\mathcal{C}$ believes $\mathcal{R}$ will play D,	(27c)
$\mathcal{R}\ \mathcal{C}\ \mathcal{R}\ \mathcal{C}\ (x)$	$\mathcal{R}$ believes $\mathcal{C}$ believes $\mathcal{R}$ believes $\mathcal{C}$ will play x,	(27d)
$\mathcal{R}\ \mathcal{C}\ \mathcal{R}\ \mathcal{C}\ \mathcal{R}\ (C)$	$\mathcal{R}$ believes $\mathcal{C}$ believes $\mathcal{R}$ believes $\mathcal{C}$ believes $\mathcal{R}$ will play C.	(27e)

This hierarchy of beliefs establishes a cycle of strategies (C, y, D, x, C), all of which are thus shown to be rationalizable. To see how the above argument is sufficient to show that x is rationalizable, let's look explicitly at Column's beliefs which would make his choice of x rationalizable. Column would play x, $\mathcal{C}(x)$, if he believed that Row would play C, i.e. if $\mathcal{C}\mathcal{R}(C)$. Row would play C if she believed that Column would play y; so we assume the belief for Column that $\mathcal{C}\mathcal{R}\mathcal{C}(y)$. Column would play y if Row

[17] $\mathcal{R}(C)$ is not a belief; but the $\mathcal{R}\mathcal{C}\cdots(\cdot)$ expressions below it are beliefs.

were playing D; therefore we assume $\mathscr{C}\mathscr{R}\mathscr{C}\mathscr{R}(D)$. Row would play D if Column were choosing x. So we have the same cycle (but shifted) of rationalizable strategies $\{x, C, y, D, x\}$. We summarize Column's beliefs:

$$\mathscr{C}\ (x) \tag{28a}$$
$$\mathscr{C}\ \mathscr{R}\ (C) \tag{28b}$$
$$\mathscr{C}\ \mathscr{R}\ \mathscr{C}\ (y) \tag{28c}$$
$$\mathscr{C}\ \mathscr{R}\ \mathscr{C}\ \mathscr{R}\ (D) \tag{28d}$$
$$\mathscr{C}\ \mathscr{R}\ \mathscr{C}\ \mathscr{R}\ \mathscr{C}\ (x) \tag{28e}$$

Comparing notes

We have shown that the strategy profile (C, x) is rationalizable and have explicitly determined the beliefs each player must hold in order to justify her strategy choice. Let's examine the properties of this outcome from two different perspectives. First, we ask what the players would find if they got together to compare notes about their belief systems. Second, we'll ask what their assessments of the wisdom of their strategy choices would be after the game was played; we'll ask them: "would you do it over again the same way?"

If Row and Column, planning to play C and x, respectively, met to discuss truthfully and candidly their perspectives on the game being played, Row would find that her beliefs were all wrong but Column would have correctly anticipated everything.

We see from (28b) that Column correctly conjectured (27a)—that Row would play C. Further, we see from (28c) → (28e) that Column also correctly intuited Row's higher-order beliefs (27b) → (27d). For example, Column correctly believed that Row believed that Column believed that Row would choose D.

Row was not so clairvoyant (or lucky). From (27b) we see that Row believed that Column would play y. In fact, Column played x instead. Not surprisingly, Row also got Column's higher-order beliefs all wrong. For example, from (27c), we see that Row thought that Column thought that Row would play D; from (28b) we see that Column actually thought that Row would play C.

After the game is played, and the actual strategy choices revealed, would the players be happy with the choices they had made? Would they do it over again the same way?[18] Column correctly forecast that Row would choose C, and Column played his best response to C, viz., x. So Column would be satisfied with the choice he made.

---

[18] There is a subtlety here. In general the players choose mixed strategies. I am *not* asking whether, after the game is played, a player is happy with the pure-strategy realization of her mixed strategy given the pure-strategy realizations of others' mixed strategies. I am asking whether she would be happy with her choice of mixed strategy given her opponents' mixed-strategy choices. To "do it over again the same way" means to once again choose the same lottery but have the roulette wheel spun again. The way I frame this scenario, it is implicit that the players would observe their opponents' mixed strategies, not just the pure-strategy realizations. This issue will arise again.

Row on the other hand incorrectly forecast Column's choice, thinking that Column would choose y instead of x. Row's best response to y, viz. C, was not a best response to the actually played x. Therefore Row could have received a higher payoff, 7 instead of 5, by playing the best response to x, viz. D, instead.

When beliefs are held in common

Let's contrast the properties we identified above for the outcome (C, x) with those of another rationalizable outcome: (A, w). The demonstration that A and w are rationalizable requires a much less circuitous hierarchy of beliefs. Row would choose A if she thought that Column would choose w, i.e. $\mathcal{RC}(w)$. Column would choose w if he thought that Row would choose A; i.e. we assume that $\mathcal{RCR}(A)$. This yields our desired cycle of strategy profiles (A, w, A):

$\mathcal{R}\ (A)$ (29a)
$\mathcal{R}\ \mathcal{C}\ \ (w)$ (29b)
$\mathcal{R}\ \mathcal{C}\ \ \mathcal{R}\ \ (A)$ (29c)

Similarly, Column's choice of w can be justified by the beliefs:

$\mathcal{C}\ (w)$ (30a)
$\mathcal{C}\ \mathcal{R}\ \ (A)$ (30b)
$\mathcal{C}\ \mathcal{R}\ \ \mathcal{C}\ \ (w)$ (30c)

We can easily summarize the belief system generated by (29a) $\rightarrow$ (30c) by saying: It is common knowledge that Row will play A and Column will play w.

Let's ask the same questions for this rationalizable strategy profile that we asked for (C, x). First, what misconceptions would Row and Column discover if they met at Gentle Ben's to trade their deepest secrets? Absolutely none. Both players not only correctly anticipated the other's action but also correctly divined the other's beliefs. For example, from (30c) and (29b) we see that Column correctly believed that Row believed that Column would indeed choose w.

After Row and Column played the game and observed each other's strategy choice, would either want to change her strategy? No. Because each player correctly anticipated her opponent's action, each played a best response to the opponent's actual choice. Neither player could improve upon her payoff given the choice of her opponent.

So we see a striking qualitative difference between the profile (C, x) and (A, w). The key lies in the following observation: Consider (C, x). Although x is Column's best response to C, C is not Row's best response to x. However, consider (A, w). A is Row's best response to w, *and* w is Column's best response to A. This is evident immediately from observing the boldface type in Figure 11. The payoff vector $(5, \mathbf{8})$ corresponding to the strategy profile (C, x) had only one element in boldface, indicating that only one player was picking a best response to the other's choice. On the other hand, both of the elements in the

payoff vector (**7**, **5**) corresponding to (A, w) appear in boldface, indicating that both players were picking a best response to the other's choice.

References

Aumann, Robert J. [1976] "Agreeing to Disagree," *Annals of Statistics* **4** 6: 1236–1239.

Bernheim, B. Douglas [1984] "Rationalizable Strategic Behavior," *Econometrica* **52** 4 (July): 1007–1028.

Fudenberg, Drew and Jean Tirole [1991] *Game Theory*, MIT Press.

Myerson, Roger B. [1991] *Game Theory: Analysis of Conflict*, Harvard University Press.

Pearce, David G. [1984] "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* **52** 4 (July): 1029–1050.