TI 2010-033/1
Tinbergen Institute Discussion Paper

But Some Neutrally Stable Strategies are More Neutrally Stable than Others

Matthijs van Veelen

Department of Economics and Econometrics, University of Amsterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.:   +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at
http://www.tinbergen.nl.

But some neutrally stable strategies are more neutrally stable than others

Matthijs van Veelen
Department of Economics and Econometrics
Universiteit van Amsterdam
Roetersstraat 11, 1018 WB Amsterdam
the Netherlands
C.M.vanVeelen@uva.nl

March 18, 2010

Abstract

For games in which there is no evolutionarily stable strategy, it can be useful to look for neutrally stable ones. In extensive form games for instance there is typically no evolutionary stable strategy, while there may very well be a neutrally stable one. Such strategies can however still be relatively stable or unstable, depending on whether or not the neutral mutants it allows for - which by definition do not have a selective advantage themselves - can open doors for other mutants, that do have a selective advantage. This paper defines robustness against indirect invasions in order to be able to discern between those two very different situations. Robustness against indirect invasions turns out to come with a very natural setwise generalisation of evolutionary stability; we prove that if a strategy is robust against indirect invasions, then this strategy and its (indirect) neutral mutants form a set that is asymptotically stable in the replicator dynamics.

1 Stable strategies and neutral mutants

Evolutionary stability is a concept in which strictness is important. A strategy is evolutionarily stable if it would (strictly) outperform any mutant in a post-invasion or post-entry situation, as long as this mutant constitutes a small enough proportion of the population. This ensures that such a strategy is asymptotically stable in the replicator dynamics; any mutant, if not arriving in too large proportions, would, in time, be pushed out again. A slightly weaker concept is neutral stability. This weaker version demands that a strategy should do at least as well as any mutant in a post-entry situation, again, as long as this mutant constitutes a small enough proportion of the population. Neutral stability thereby allows for ties between incumbent strategy and mutants. This naturally comes with weaker dynamic implications; the replicator dynamics do not necessarily push the population back towards the neutrally stable strategy, but we can at least be sure that locally the dynamics do not push a population away from it.

At first sight, and for many games, the concept of neutral stability functions perfectly well as a bordering case between evolutionary stability and instability. If we for instance look at games with payoff matrices as below,

$$\begin{bmatrix} 1 & 1 \\ 1 & a \end{bmatrix}, \tag{Ex. 1}$$

then strategy 1 is perfectly adequately described as evolutionarily stable if $a < 1$, as neutrally, but not evolutionarily stable if $a = 1$, and as instable if $a > 1$. In the set of games described for $a \in \mathbb{R}$ with this payoff matrix, the case where $a = 1$ could be considered to be degenerate, but there are many games in which neutral mutants occur naturally. Especially in extensive form games, where different strategy profiles can lead to the same actions being played, it is natural that there are neutral mutants. Mutants that only differ off the equilibrium path result in the same payoffs as an incumbent equilibrium strategy receives, both when played against the incumbent and when played against itself. These mutants therefore do equally well as the incumbent strategy against the post-entry population.

In extensive form games, neutral mutants not only occur naturally, but they also give rise to a new question. This question is: how much harm can a neutral mutant do? In the following two games, we will focus on the stability of strategy 1 in the matrices.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \tag{Ex. 2a}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 2 & 2 \end{bmatrix} \tag{Ex. 2b}$$

In both games strategy 1 is a neutrally stable strategy (an NSS) and strategy 2 is a neutral mutant of strategy 1. There is a difference, though. Because strategy 2 is not pushed out of the population in either one of the two games, one could imagine that if there is some random drift, then it is possible that strategy 2 attains a considerable share in the population. In the first game, this has no effect on the opportunities for strategy 3 to enter the population. This strategy has a selective disadvantage when strategy 1 is the only strategy in the population, and the selective disadvantage remains, whatever proportion strategy 2 attains in the population. In the second game, however, if the proportion of strategy 2 players in the population exceeds 50%, then strategy 3 will outcompete both strategies in the post-entry mix. In this game, a neutral mutant (strategy 2) can open the door for another strategy, which makes strategy 1 in Example 2b (much) less stable than in Example 2a.[1]

In order to discern how much harm neutral mutants can do, we introduce the concept of robustness against indirect invasions (RAII). This is done in Section 3. Robustness against indirect invasions turns out to be an attractive concept, because it has nice dynamic properties; we show, also in Section 3, that if a strategy is robust against indirect invasions, then the set consisting of this strategy and its neutral mutants is asymptotically stable in the replicator dynamics. Thereby it generalizes the equivalent result for evolutionarily stable strategies (ESS'es).

For single strategies, evolutionary stability is a characteristic that can relatively easily be checked. A classic result is that evolutionary stability as defined in Maynard Smith & Price (1973) is equivalent to a strategy being locally superior (Hofbauer, Schuster & Sigmund, 1979, Weibull, 1995, p45). The latter is much harder to check directly, but much easier shown to be a sufficient condition for asymptotic stability in the replicator dynamics. Local superiority therefore is a very useful in between step to link the much more applicable concept of evolutionary stability to asymptotic stability in the replicator dynamics.

For sets of strategies, the literature contains a few definitions of what it means to be evolutionarily stable. Thomas' (1985) original definition of an ES-set could be seen as a setwise hybrid of evolutionary stability and local superiority, while for instance Cressman's (1992) definition can be seen as a more natural setwise generalisation of evolutionary stability. Thomas definition, however, and the results in his paper turn out to be very useful to show that, if a strategy is robust against indirect invasions, then this strategy, together with its (indirect) neutral mutants, is an asymptotically stable set. Whether a strategy is robust against indirect invasions is relatively easily checked, and the set that includes its (indirect) neutral mutants is also easily constructed. This implies that such a set can play the same role in the setwise generalization as evolutionary stability plays in the singleton set case.

In Section 4 we compare RAII to the concept of robustness against equilibrium entrants

[1] The appendix contains phase plots for the examples that can be quite instructive.

(REE) from Swinkels (1992a). While more than reasonable in some settings, a disadvantage of this concept is that strategies that are REE need not even be Lyapounov stable, as we show with an example. Because these static concepts as well as the replicator dynamics are ways to get at properties of stochastic processes, we also describe how that should guide the application of these concepts. In particular the differences between applications in economics and biology are highlighted. We also discuss how they relate to interesting results in Balkenborg & Schlag (2001, 2007).

In the discussion we will also point out why the concept of robustness against indirect invasions is particularly useful when we look at evolution of strategies in repeated games.

2 Preliminaries

We will assume that the number of pure strategies is finite. Those pure strategies will then be denoted by $e^1, ..., e^n$, which are the vertices of the unit simplex Δ. We will denote strategies with lower case letters, and x_i will be the probability with which strategy x plays pure action e^i. Games are characterized by an $n \times n$ matrix A, and payoffs are defined by $\pi(x, y) = x^T Ay$.

3 Definitions and results

As a preparation for the definition, we will first describe what it is meant to exclude. Let x be a neutrally stable strategy (NSS). Suppose furthermore that there is an $\alpha \in (0, 1)$ and that there are strategies y and z for which the following holds:

$$\begin{aligned} \pi(x, x) &= \pi(y, x) \\ \pi(x, y) &= \pi(y, y) \\ \pi(z, (1-\alpha)x + \alpha y) &> \pi(x, (1-\alpha)x + \alpha y) \end{aligned}$$

The first two equations make y a neutral mutant of x, as they imply that $\pi(x, (1-\alpha)x + \alpha y) = \pi(y, (1-\alpha)x + \alpha y)$ for all $\alpha \in [0, 1]$. The third formalises what harm that could do. The neutral mutant y can by random drift attain a share α in the population for which a second mutant z outperforms both x and y (see Binmore & Samuelson, 1994). This possibility is to be excluded by the definition.

A simple way to exclude this is to demand that if x and y are neutral mutants, then for all z it should be that $\pi(z, (1-\alpha)x + \alpha y) \leq \pi(x, (1-\alpha)x + \alpha y)$. This would rule out a sequence of one neutral mutant followed by a mutant that has a selective advantage. Note that there is good reason not to demand *strict* inequality here; just as much as we want to allow for y to be a neutral mutant of x, we also want to allow for z's that are neutral mutants of both. In the example below, demanding strict inequality would make e^1 fail to qualify for our notion of stability, both in Example 3a and in Example 3b. Still, this is just a copy of Example 2, with one dimension added to the set of neutral mutants - e^2 and e^3 are indiscernable - and with clearly different dynamics; the replicator dynamics push all populations (back) towards the subsimplex spanned by e^1, e^2 and e^3 in the Example 3a, while in Example 3b, the dynamics will push away from the same subsimplex if e^2 and e^3 together make up more than 50% of the population.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{Ex. 3a}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 \end{bmatrix} \tag{Ex. 3b}$$

A drawback of the inequality not being strict is the following. While it excludes a sequence of *one* neutral mutant followed by a mutant that has a selective advantage, it does not exclude the possibility of a sequences of more than one neutral mutant, before a mutant with a strict advantage turns up. The following example illustrates that.

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} \tag{Ex. 4}$$

In this game there is a number of different equilibria. First e^4 is an ESS. Then e^3 is a Nash equilibrium, but not an NSS. Furthermore, e^2 is an NSS, but not an ESS. Finally e^1 is also an NSS, but not ESS

Looking at the replicator dynamics for this game (see Appendix A), we observe that there is a path along which drift and selection can take a population all the way from e^1 to e^4. If first the neutral mutant e^2 arises, and drift drives e^1 extinct, and second the mutant e^3 that is neutral for e^2 arises, and drift drives e^2 extinct, then e^4 can successfully invade and take over the population. So even though e^4 would be at a disadvantage facing any mix of e^1 and its neural mutant e^2, there is a path that starts in e^1 and ends in e^4 along which the population only encounters neutral drift and selection in the direction of the latter. (It should however be noted that e^3 is not neutral for e^1. Hence a reintroduction of e^1 anywhere between e^2 and e^3 throws the population back onto the boundary face between e^1 and e^2, because e^3 has a selective disadvantage anywhere but at the boundary face between e^2 and

e^3 itself). It is therefore reasonable to think that e^1 is more stable than e^2, and e^2 is more stable than e^3, while e^4 of course is the most stable.[2] Still there does exist a path from e^1 to e^4 on which the first two mutants do not have a selective disadvantage, and the third mutant has an actual advantage. Whether this possible escape route makes e^1 seriously unstable or not, can be a matter of debate.

One argument against taking this escape route too seriously is that, in this example, e^2 is an infinitely small part of the set of all convex combinations of e^1 and e^2, and the farthest from e^1. The idea of stability criteria is that they supposedly capture essentials of stochastic processes that have mutation and selection in them by separating the two. Mutations occur at the individual level and hence are small at the population level. This makes it reasonable to use local stability criteria; it may be that there are mutations that would not be neutralized by the (deterministic) replicator dynamics, but if they are enormous jumps in the population, it is unreasonable to require that they would be stabilized. Here, one could argue, it takes a sequence of two neutral, but still very large mutations in order to open the door for a third one.

There is however also a very strong argument in favour of taking this escape route seriously. Although these stability concepts are defined for infinite population models, of course they are only useful in as far as they describe large, but still finite populations. In the literature on finite populations, there is a pivotal role for the fixation probability of a neutral mutant.[3] One can imagine that after one individual mutates, all individuals in the population have equal chances of reproducing, if this mutation is a neutral one. This is not the same as to say that they all get exactly the same number of offspring; equal fitness only means that the probabilities are equal. Random drift can therefore in- or decrease the share of this mutant in the population. Because the state space of population compositions has two absorbing boundaries (one where all individuals are playing the incumbent strategy and one where all individuals are playing the mutant strategy), one can easily see that if mutations are rare, a stochastic process with rare mutations will spend most of its time at the two states that are absorbing for the stochastic process without mutations. The extremes of the set of convex combinations of two strategies therefore are not just any two points of this set; they are the points where the population finds itself most of the time, if

[2] Generalisations of this game are also possible, if we define a game with strategies e^1 to e^n as follows:

$$\pi\left(e^i, e^j\right) = \begin{cases} 1 & \text{if } i \leq j+1 \text{ and } i \neq n \\ 0 & \text{if } i > j+1 \\ 2 & i = j = n \end{cases}$$

If we take $n \geq 4$, then e^n is an ESS, e^{n-1} is a Nash equilibrium, but not an NSS, $e^1, ..., e^{n-2}$ are NSS, but not RAII. Still it is clear that the higher n, the more stable e^1 and its neutral mutant e^2 - or the boundary face between e^1 and e^2 - is.

[3] The fixation probability of a neutral mutant is $1/N$, where N is the size of the population. See for instance Chapter 6-8 in Nowak (2006) for intuitive and formal explanations.

mutations are rare. If mutations from any strategy in example 1 to any other strategy in example 1 are equally probable, but all rare, then going from e^1 to e^4 is just a matter of the right order of the mutants that succeed in taking over the population (first e^2, then e^3, then e^4). Here, one can assume that with time, this "code" will be broken.

Robustness against indirect invasions (RAII), as defined below, does exclude all higher order indirect invasions, including the ones described above. It turns out that robustness against indirect invasions does also have nice dynamic properties. This gives us another reason to take a special interest in this concept that does look at sequences of more than one neutral mutant.

As a preparation, we define three sets for any strategy x: the set of (evolutionarily) worse, equal and better performers against x. Formally, that is:

$$\begin{aligned}
&S_W(x) = \{y \mid \pi(y,x) < \pi(x,x) \text{ or } (\pi(y,x) = \pi(x,x) \text{ and } \pi(y,y) < \pi(x,y))\} \\
&S_E(x) = \{y \mid \pi(y,x) = \pi(x,x) \text{ and } \pi(y,y) = \pi(x,y)\} \\
&S_B(x) = \{y \mid \pi(y,x) > \pi(x,x) \text{ or } (\pi(y,x) = \pi(x,x) \text{ and } \pi(y,y) > \pi(x,y))\}
\end{aligned}$$

Note that being a neutral mutant is a symmetric thing; if $y \in S_E(x)$, then $x \in S_E(y)$. Also, these sets are disjoint, and together they cover the whole strategy space.

The last two sets help defining robustness against indirect invasions. Here we use superscripts to discern different mixed strategies, because subscripts are already in use to indicate the different entries of the vector that represents the strategy.

Definition 1 *A strategy x is robust against indirect invasions (RAII) if*

1) $S_B(x) = \varnothing$ and

$$\textit{2) } \nexists\ y^1, ..., y^n,\ n \geq 2, \textit{ such that } \begin{cases} y^1 \in S_E(x) \\ y^i \in S_E\left(y^{i-1}\right), \quad 2 \leq i \leq n-1 \\ y^n \in S_B\left(y^{n-1}\right) \end{cases}$$

It is not hard to check that strategy e^1 in Example 4 is not robust against indirect invasions.

If a strategy x is RAII, then it makes sense to group this strategy and its (indirect) neutral mutants together in a set.

$$\textbf{Definition 2 } S_{NM}(x) = \left\{ y \mid \exists\ y^1, ..., y^n, n \geq 1 \textit{ such that } \begin{array}{l} y^1 \in S_E(x) \\ y^i \in S_E\left(y^{i-1}\right), \quad 2 \leq i \leq n \\ y \in S_E(y^n) \end{array} \right\}$$

Note that $y \in S_E(y)$, so this set $S_{NM}(x)$ does include normal neutral mutants too; if y is a neutral mutant, take $n = 1$ and $y^1 = y$. It is also easy to see that if x is an ESS, then it is immediately also RAII, and $S_{NM}(x) = \{x\}$. This implies that this definition, for strategies that are RAII, subsumes singleton sets of strategies that are evolutionarily stable. It is also clear that if x is RAII and $y \in S_{NM}(x)$, then y is also RAII and $S_{NM}(y) = S_{NM}(x)$. The strategy x therefore does not play a special role in the set $S_{NM}(x)$.

We can illustrate these two definitions with another example.

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \tag{Ex. 5}$$

Here it is clear that all four pure strategies are RAII. Also, e^1 and e^2 are neutral mutants of each other, e^2 and e^3, e^3 and e^4, and e^4 and e^1. The set $S_{NM}\left(e^1\right)$ - or $S_{NM}\left(e^i\right)$ for $i = 1, .., 4$ - is the union of the convex hulls of e^1 and e^2, e^2 and e^3, e^3 and e^4, and e^4 and e^1 (see also Appendix A).

The set of a strategy and its (indirect) neutral mutants also has a very attractive general property. If x is RAII, then this set is asymptotically stable in the replicator dynamics. To see this, it suffices to show that $S_{NM}(x)$ is a (Thomas) ES-set (see Thomas, 1985, Definition 5, p 111, Weibull, 1996, Definition 2.6, p 51, or Schlag & Balkenborg, 2001, Definition 2, p 579).

Definition 3 *A set $X \subset \Delta$ is called a* (Thomas) evolutionarily stable set *if it is non-empty and closed, and if for each $x \in X$ the following holds:*

1) $\pi(y, x) \leq \pi(x, x)$ for each $y \in \Delta$ (that is: x is a symmetric Nash equilibrium strategy)

2) there is a neighbourhood U_x such that for all $y \in U_x$ for which $\pi(y, x) = \pi(x, x)$, the inequality $\pi(x, y) \geq \pi(y, y)$ holds, whereby $\pi(x, y) = \pi(y, y)$ implies $y \in X$.

This definition turns out to be very useful, even though it can be seen as a hybrid between a setwise version of evolutionary stability and a setwise version of local superiority. It starts out as the classical definition of evolutionary stability for single strategies by Maynard Smith & Price (1973), but then introduces a neighbourhood, as in the local superiority definition for single strategies by Hofbauer, Schuster & Sigmund (1979). Lemma 1 in Thomas (1985) implies that this hybrid definition however is equivalent to a setwise generalisation of local superiority (see also Proposition 2.10 in Weibul, 1996, p 51).

Proposition 4 *A set $X \subset \Delta$ is a (Thomas) ES-set if and only if it is non-empty and closed and if each $x \in X$ has some neighbourhood U_x such that $\pi(x,y) \geq \pi(y,y)$ for all $y \in U_x$ and $\pi(x,y) > \pi(y,y)$ for all $y \in U_x \backslash X$.*

It will also be useful to have a definition from Balkenborg & Schlag (2001, point *e)* in Theorem 3, p 585). We only have included non-emptyness and closedness, but one can show that closedness is in fact a redundant requirement.

Definition 5 *A set $X \subset \Delta$ is called a* (Balkenborg & Schlag) evolutionarily stable set *if it is non-empty and closed, and if for each $x \in X$ the following holds:*

1) $\pi(y,x) \leq \pi(x,x)$ for each $y \in \Delta$ (that is: x is a symmetric Nash equilibrium strategy)

2) if $\pi(y,x) = \pi(x,x)$, then the inequality $\pi(x,y) \geq \pi(y,y)$ holds, whereby $\pi(x,y) = \pi(y,y)$ implies $y \in X$.

From Theorem 3 in Balkenborg & Schlag - which shows equivalence of 5 definitions - it follows that a (Balkenborg & Schlag) ES-set is equivalent to a (Thomas) ES-set. Because the equivalence of those two is relatively easy, we can focus on those two only with two simple steps:

Lemma 6 *A set $X \subset \Delta$ is a (Balkenborg & Schlag) ES-set if and only if it is a (Thomas) ES-set.*

Proof. Balkenborg & Schlag $\Rightarrow$ Thomas: take $U_x = \Delta$

Thomas $\Rightarrow$ Balkenborg & Schlag: If *1)* holds (that is, it is a set consiting of Nash equilibrium strategies) for set X, but it it is not a (Balkenborg & Schlag) ES-set, then either

a) $\exists\ y$ for which $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) > \pi(x,y)$ or
b) $\exists\ y \notin X$ for which $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) = \pi(x,q)$.

Now consider $z = \alpha x + (1-\alpha) y$ with $\alpha \in (0,1)$.

If it is a), then for all $\alpha \in (0,1)$ we find that $\pi(z,x) = \alpha\pi(x,x) + (1-\alpha)\pi(y,x) = \pi(x,x)$ and $\pi(z,z) = \alpha^2\pi(x,x) + \alpha(1-\alpha)[\pi(x,y) + \pi(y,x)] + (1-\alpha)^2\pi(y,y) = \alpha\pi(x,x) + \alpha(1-\alpha)\pi(x,y) + (1-\alpha)^2\pi(y,y) > \alpha\pi(x,x) + (1-\alpha)\pi(x,y) = \pi(x,z)$. But then we can for any U_x choose α small enough for z to be an element of U_x, which implies that X is not a (Thomas) ES set.

If it is b), then for all $\alpha \in [0,1]$ we find that $\pi(z,x) = \pi(x,x)$ and $\pi(z,z) = \pi(x,z)$. Since $x \in X$ and $y \notin X$, and X is closed, there is at least one $\widehat{\alpha}$ such that $\widehat{z} = \widehat{\alpha}x +$

$(1-\widehat{\alpha})\,y \in X$, while there is no $U_{\widehat{z}}$ for which all $\alpha x + (1-\alpha)\,y$ that are elements of $U_{\widehat{z}}$ are also elements of X. But then X is, again, not a (Thomas) ES set. ■

The set $S_{NM}(x)$ can easily be shown to be a (Balkenborg & Schlag) ES-set.

Lemma 7 *If x is robust against indirect invasions, then $S_{NM}(x)$ is a (Balkenborg & Schlag) ES-set.*

Proof. This is relatively straightforward. Since $S_B(y) = \varnothing$ for all $y \in S_{NM}(x)$, this set consists of Nash equilibrium strategies only, which implies that 1) holds. If $S_{NM}(x)$ is not a (Balkenborg & Schlag) ES-set, then it must, again, be that either

a) $\exists\ y$ for which $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) > \pi(x,y)$ or
b) $\exists\ y \notin S_{NM}(x)$ for which $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) = \pi(x,q)$.

If it is a), then that contradicts that $S_B(y) = \varnothing$ for all $y \in S_{NM}(x)$.

If it is b), then we get a contradiction too, because $y \notin S_{NM}(x)$, while from $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) = \pi(x,y)$ it follows that $y \in S_{NM}(x)$. ■

The inverse implication also holds.

Lemma 8 *If X is an ES-set, with $x \in X$, then x is strongly robust against indirect invasions*

Proof. First observe that if $x \in X$ and if y is a neutral mutant of x - that is: $\pi(x,x) = \pi(y,x)$ and $\pi(x,y) = \pi(y,y)$ - then all strategies $y^{\alpha} = (1-\alpha)\,x + \alpha y$ for $\alpha \in [0,1]$ must also be elements of X; if not, then by closedness of ES-sets, there is a $y^{\alpha} \in X$ for which every neighbourhood U of y^{α} contains a strategy $y^{\beta} \notin X$, for which $\pi\left(y^{\alpha}, y^{\beta}\right) = \pi\left(y^{\beta}, y^{\beta}\right)$, which contradicts that X is an ES-set. This argument can be applied repeatedly in order to show that actually $S_{NM}(x) \subset X$.

Now suppose x is not RAII. This implies that either $S_B(x) \neq \varnothing$ or that $\exists y^1, ..., y^n$, $n \geq 2$, such that $y^1 \in S_E(x), y^i \in S_E\left(y^{i-1}\right), 2 \leq i \leq n-1$, and $y^n \in S_B\left(y^{n-1}\right)$. Because of the argument above, we can, without loss of generality, focus on the former (since we know that then $y^{n-1} \in X$, in which case we would focus on $S_B\left(y^{n-1}\right) \neq \varnothing$ instead).

But if $S_B(x) \neq \varnothing$, then there is a y such that either $\pi(y,x) > \pi(x,x)$ or $\pi(y,x) = \pi(x,x)$ and $\pi(y,y) > \pi(x,y)$. This implies that X is not a (Balkenborg & Schlag) ES-set. ■

Together, Lemma 7 and 6 pave the way from robustness against indirect invasions to asymptotic stability; if x is RAII, then $S_{NM}(x)$ is a (Balkenborg & Schlag) ES-set by Lemma 7, hence a (Thomas) ES-set by Lemma 6 and asymptotically stable by Corollary 3 in Thomas (1985).

For completeness, we also give Definition 3f from Cressman (1992, p191). It is straightforward that this coincides with what Balkenborg & Schlag (2001, p576) call a simple ES-set.

Definition 9 *A set $X \subset \Delta$ is called a* (Cressman) evolutionarily stable set *if it is non-empty and closed, and if for each $p \in X$ the following holds:*

1) $\pi(y,x) \leq \pi(x,x)$ for each $y \in \Delta$ (that is: x is a symmetric Nash equilibrium strategy)

2) for each $y \notin X$ for which $\pi(y,x) = \pi(x,x)$ holds, $\pi(x,y) > \pi(y,y)$.

Balkenborg & Schlag's (2001) Theorem 3 encompasses equivalence of a (Cressman) and a (Balkenborg & Schlag) ES-set.[4]

I would like to conclude this section with an example of an asymptotically stable set, which it is not an ES set. In the example below, there is no strategy that is RAII.

$$\begin{bmatrix} 1 & 1 & 0 & 2 \\ 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad \text{(Ex. 6)}$$

Still, the union of the plane spanned by e^1, e^2 and e^3 and the plane spanned by e^3, e^4 and e^1 is a (Weibull) ES* set (see the definition below). From Weibull, Proposition 3.13, we know that this set is asymptotically stable.

Definition 10 *A set $X \subset \Delta$ is a (Weibull) ES*-set if it is nonempty and closed and if each $p \in X$ has some neighbourhood U_p such that $\pi(p,q) > \pi(q,q)$ for all $q \in U_p \backslash X$*

The RAII concept does not only provide a shortcut in determining that this game has no ES-set, but, more importantly, here it is also informative about the dynamics we can expect.

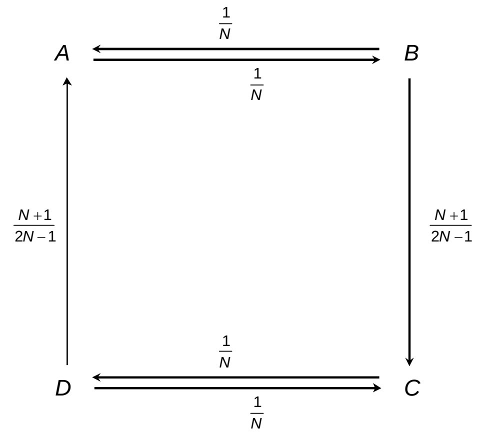
[4] Lemma 4, page 193, in Cressman (1992) suggest that a (Cressmann) ES-set is (setwise) locally superior - in the sense of Proposition 4. This is true, but the proof of the Lemma is not correct. It is important to realise that the proof there (bottom line page 193) invokes an argument from Theorem 15.5 in Hofbauer & Sigmund (1988), which is the same as Theorem 6.4.1 in Hofbauer & Sigmund (1998) or Propositions 2.5 and 2.6 in Weibull (1995). Those results concern ESS'es, that by definition do not have neutral mutants, and those proofs use the fact that the set Z_x is compact, and that $b_x(t) > 0$ on Z_x. This implies that b_x attains a minimum larger than 0 on Z_x. Here, and in Lemma 4 in Cressman (1992), neutral mutants are allowed, which implies that $b_x(t) > 0$ only on $Z_x \backslash S_E(x)$, which is no longer a compact set. This by itself therefore does not exclude the possibility that there is a sequence $t^i \in Z_x \backslash S_E(x)$ such that $\lim_{i \to \infty} b_x(t) = 0$. Appendix B contains a proof with the same approach that does overcome the lack op compactness, but it is considerably longer.

Although the replicator dynamics predict no movement at all when the population finds itself in the two pure Nash equilibria, the fact that there are indirect invasions suggests that we might expect that both of them will be left in time. Since the way out of A is through its neutral mutant B and then to C, and the way out of C is through its neutral mutant D and then back to A, we thus expect this population to move "clockwise", with stickyness at the two pure equilibria.

Because infinite population analysis is only useful inasfar as it is informative about dynamics in large, but finite populations, it makes sense to see if this is in line with what we find when we look at stochastic dynamics with low mutation probabilities in finite populations. With this game we know from From Fudenberg & Imhof (2006) that a Moran process with rare stochastic mutations will spend almost all of its time in the four pure states, and that it will be sufficient to look at the fixation probabilities of single mutants in those pure states. The fixation probabilities have simple explicit expressions (see for instance Nowak, 2006, page 110 or Fudenberg & Imhof, 2006, page 256) some of which are even extremely simple or obvious for this example. With strong selection and N denoting the population size, $\rho_{A\to B}$,$\rho_{B\to A}$,$\rho_{C\to D}$ and $\rho_{D\to C}$ are $\frac{1}{N}$, since they are pairs of neutral mutants. Fixation probabilities $\rho_{A\to C}$,$\rho_{A\to D}$,$\rho_{B\to D}$,$\rho_{C\to A}$,$\rho_{C\to B}$,$\rho_{D\to B}$ are all zero, since the single mutants have payoff 0 and are chosen for reproduction with probability 0. The remaining two are a slightly less obvious (the computation is in Appendix C), but they do give simple results; $\rho_{B\to C} = \rho_{D\to A} = \frac{N+1}{2N-1}$. This gives us the following transition matrix.

$$\begin{matrix} \frac{N-1}{N} & \frac{1}{N} & 0 & 0 \\ \frac{1}{N} & \frac{N^2-4N+1}{2N^2-N} & \frac{N+1}{2N-1} & 0 \\ 0 & 0 & \frac{N-1}{N} & \frac{1}{N} \\ \frac{N+1}{2N-1} & 0 & \frac{1}{N} & \frac{N^2-4N+1}{2N^2-N} \end{matrix}$$

Leaving out the zero's and the diagonal elements, this can be pictured as follows:



It is obvious that rare mutations and strong selection exclude making full counter-clockwise circles for any population size. Also, in the limit for $N \to \infty$, the invariant distribution becomes $\left[\frac{1}{2}, 0, \frac{1}{2}, 0\right]$ and we get clockwise movement only, when we restrict our attention to the four pure states (that is: we will not observe transitions from B back to A nor from D back to C).

Of course this is just an example that allows for simple computations. It nonetheless does indicate that indirect invasions - and robustness against it - is a useful concept. It shows that the replicator dynamics can be insufficient to describe the dynamics in the limit of infinite populations. This implies that the existence of an NSS has limited value for describing dynamic behaviour of the system; the implication it has for Lyapounov stability of the replicator dynamics is only of limited use here, since the replicator dynamics do not capture random drift. The indirect invasions furthermore do indicate the direction in which the population will move. Finding circles here is therefore of additional value; with the Rock-Scissors-Paper game, we are glad that we do not just know under which conditions the Nash equilibrium is stable, but also that the replicator dynamics circles around it (see Weibull, 1995). The same applies here; we are not just happy that the concept of indirect invasions help us determine whether or not ES-set exists, but also that they indicate the circles the dynamics will make.

4 Relations with other (setwise) criteria

4.1 Robustness against equilibrium entrants and equilibrium evolutionary stability

From the definitions, it is clear that the following inclusions hold:

$$\Delta^{ESS} \subset \Delta^{RAII} \subset \Delta^{NSS} \subset \Delta^{NE}$$

where Δ^{ESS} is the set of ESS'es, Δ^{RAII} is the set of equilibria that are RAII, Δ^{NSS} is the set of NSS'es, and Δ^{NE} is the set of Nash equilibria. Robustness against indirect invasion is therefore a stability criterion that is weaker than evolutionary stability.

Having such a weaker stability criterion can be very useful if the set of evolutionarily stable strategies is empty. There are however also other weaker stability criteria. A well known one is robustness against equilibrium entrants (REE, Swinkels 1992a, see also Swinkels, 1992b) and its set valued version of equilibrium evolutionary stability (EES, also in Swinkels 1992a). The attraction of these concepts is that they take an intermediate position with regard to how much anticipation, or how much rationality, players are assumed to have. While traditional applications of the Nash equilibria tend to assume that there are no limits to how much or how well a player can anticipate, the evolutionary refinement, quite to the contrary, assumes no anticipation whatsoever. Robustness against equilibrium entrants takes an intermediate position here, by disregarding possible mutants that are themselves not best responses to the post-entry mix. If there are no mutants that, when entering in not too large proportions, are best responses to the post-entry mix, then this strategy is robust against equilibrium entrants.

Definition 11 *A strategy $x \in \Delta$ is robust against equilibrium entrants (REE) if there exists some $\overline{\epsilon} \in (0,1)$ such that for all $y \neq x$ and $\epsilon \in (0, \overline{\epsilon})$, y is not a best response to $\epsilon y + (1 - \epsilon) x$.*

This is Definition 7 from Swinkels (1992a), only slightly reformulated as in definition 2.5 from Weibull, 1995. An equivalent, and sometimes easier to check definition from Swinkels (1992a) is

Definition 12 *A strategy $x \in \Delta$ is robust against equilibrium entrants (REE) if*

x is a Nash equilibrium of the game

x is the only Nash equilibrium of the game, if we restrict the game to only include those strategies that are best replies to x.

This concept allows for some quite remarkable links with other equilibrium refinements, such as properness and 'never a weak best response'.

The motivating example in Swinkels (1992a) is a 3 by 3 game, equivalent to an extensive form game, where both individuals can veto a subgame in the first stage. The subgame then is a prisoners dilemma, for which the cooperative outcome has a higher and the Nash equilibrium has a lower payoff than the payoff to the players if the subgame is vetoed. In normal form, that translates to the following game, where playing e^1 means vetoing the subgame, e^2 means not vetoing the subgame and playing cooperate, and e^2 means not vetoing the subgame and playing defect.

$$\begin{bmatrix} 3 & 3 & 3 \\ 3 & 4 & 0 \\ 3 & 5 & 1 \end{bmatrix} \tag{Ex. 7}$$

For this game one can show that all populations in the interior of the unit simplex converge to e^1 (see also Appendix A). For general 3 by 3 games, one can show that the same actually holds for any p that is REE. This does however not hold in general for games with a finite number of strategies. The following example shows that for larger games, there are limits to how well this concept can be linked with stability in the replicator dynamics.

$$\begin{bmatrix} 6 & 6 & 6 & 6 \\ 6 & 8 & 9 & 0 \\ 6 & 0 & 8 & 9 \\ 6 & 9 & 0 & 8 \end{bmatrix} \tag{Ex. 8}$$

Strategy e^1 in this game is REE; it is in fact the unique Nash equilibrium of this game, which makes it fit Definition 12. It is, however, not an NSS, and therefore immediately also not RAII. If we want to look at the dynamic behaviour after mutant invasions, then it is worth seeing that if we restrict the game to only include e^2, e^3 and e^4, then the equilibrium puts probability $\frac{1}{3}$ on each of those strategies. This strategy, played against itself, would earn a payoff of $5\frac{2}{3}$, which is less than 6. Therefore, if these three strategies would enter together in exactly equal proportions, they would be pushed out by the replicator dynamics. The equilibrium of the game restricted to e^2, e^3 and e^4, however, is unstable; it is a Rock-Scissors-Paper game for which the replicator dynamics spiral away from the equilibrium (see Weibull, 1995, p77). The smallest of deviations from equal proportions of the mutant therefore would send the proportions of e^2, e^3 and e^4 spiralling away. Even if in the beginning average payoffs would still be lower than 6, while spiralling away they would at some point become larger than 6, pushing e^1 out of the population in the long run. This implies that any invasion y for which $y_1, y_2, y_3 > 0$ and for which *not* $y_1 = y_2 = y_3$ - thereby only excluding a very special set of mutants - will lead to the replicator dynamics pushing the population away from the equilibrium, however little q differs from the equilibrium. As almost all populations in the interior of the unit simplex do not converge to e^1, this example is in relatively sharp contrast to example 4. (Figure 6 in Appendix A nicely depicts this; the black line in it represents the mutations from which the replicator dynamics take the population back to e^1, while the other three trajectories start only fractions away from there, and yet all end up on the subsimplex opposite to the REE equilibrium).

Because the set valued version of this concept (EES) coincides with the point-valued one (REE) for singleton sets, stability in the replicator dynamics can also not be guaranteed for equilibrium evolutionarily stable sets.

Examples 4 and 5 are also useful for discussing different approaches. In a situation where individuals (can) anticipate what their opponent does in the next stage, it can be preferable to use the REE concept; it formalises a good reason why people would play strategy e^1 in example 4, which implies that both veto the prisoners dilemma being played, as they expect the Nash equilibrium to be played in the subgame. On the other hand, we could also assume no anticipation whatsoever, but just look at how well possible mutants would do. In that case we would expect that if mutants are rare, and we start at a monomorphous population where all individuals play e^1, then a mutant e^2 could invade, and take over the population. This new monomorphous state would now be susceptible to invasion by e^3, which in turn could be invaded by e^1. Therefore, depending on whether or not we allow players to anticipate what others would do, we either use the REE concept and predict e^1, or we do not and predict a cycle from e^1 to e^2 to e^3 and back to e^1. Please note that with anticipation, e^3 does not have to actually be present in the population to cast a shadow over e^2. In a context with no anticipation, e^2 is only hurt by e^3 after it mutates into the population, but with anticipation, the danger of e^3 as a possibility excludes e^2 being played, even without anyone actually playing e^3.

Example 5 is a little more complicated. For this game, it is actually a bit hard to decide how much anticipation we should bestow on players if we want to choose an intermediate position, since the post-entry dynamics are not trivial to compute. None of the mutants threaten e^1 according to the REE criterion; the ones with unequal shares of the other three pure strategies are discarded because best responding to any on those would require a strategy in which two of the last three pure strategies are not used, and the ones with equal shares of the last three pure strategies are discarded because e^1 would outperform it. Still any mutant strategy, apart from the ones with equal shares of the last three pure strategies, would take the population away from e^1, with no natural way back.

4.2 More general fitness functions and different possibilities for uniform invasion barriers

Balkenborg and Schlag (2001) look at ES-sets under more general fitness functions. This calls for a richer set of definitions, and apart from (Thomas) ES-sets and (Balkenborg & Schlag) ES-sets, they also introduce simple, pointwise uniform and uniform evolutionarily stable sets. Above, we have summarized the theory for single strategies - see Weibull (1995) - as follows: evolutionary stability and local superiority are equivalent, and imply asymptotic stability in the replicator dynamics. A more detailed summary would actually be that there are *three* static concepts that are equivalent (evolutionary stability, the existence of a uniform invasion barrier, and local superiority) and that imply asymptotic stability in the replicator dynamics.

$$\begin{tikzcd}
\textbf{ESS} \arrow[r, shift left] & \textbf{UIB} \arrow[l, shift left] \arrow[r, shift left] & \textbf{LS} \arrow[l, shift left] \arrow[r, shift left] & \textbf{AS} \arrow[l, shift left, "\times" marking]
\end{tikzcd}$$

A (Thomas) ES set is a hybrid between an setwise generalisation of an ESS and one of Local Superiority. We have furthermore focussed mostly on alternative setwise definitions of an ESS and a setwise generalisation of LS. The in between step of the existence of a uniform invasion barrier is however important in the setwise generalisations in Balkenborg & Schlag (2001), where they explore different possibilities of what a uniform invasion barrier could be in a setwise generalisation (every element of a set could be required to have a uniform invasion barrier of its own, or one could require that there be an overall uniform invasion barrier).

Another interesting paper in this domain is Balkenborg & Schlag (2007), although the setting there is somewhat different, as it concerns a multipopulation model for (possibly) asymmetric games, while we focus on a single population model for symmetric games (see also Binmore & Samuelson, 1994).

5 Conclusions and discussion

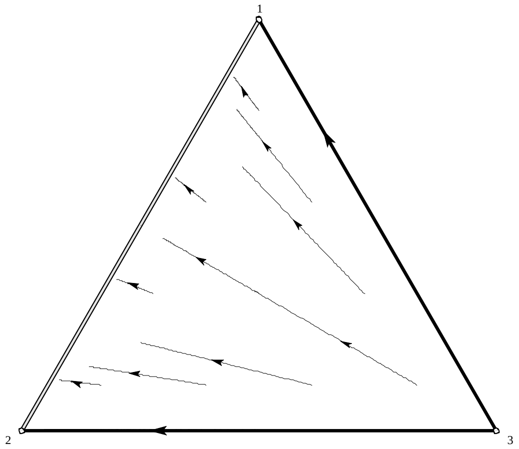
In games where neutral mutants occur naturally, such as extensive form games, and for strategies that are neutrally stable, it is worth checking whether or not these neutral mutants, that do not have a selective advantage themselves, can be harmful or not by opening doors for other mutants that do have a selective advantage if the proportion of those neutral mutants in the population is high enough. Robustness against indirect invasions distinguishes between these two situations. If a strategy is robust against indirect invasions, then the set that consists of this strategy and its (indirect) neutral mutants is asymptotically stable in the replicator dynamics. This set thereby is a natural setwise generalization of the ESS concept which is relatively easy to check, while it may not always be so easy to find (the different types of) ES-sets without the concept of robustness against indirect invasions.

A special example is repeated games. In a companion paper, Julián García and I apply this new concept to repeated games with discounting, in which neutral mutants also occur naturally. There we show that no strategy is robust against indirect invasions; for every strategy we can find stepping stone paths out of equilibrium. With the results presented here, we know that this implies that there is no (Thomas) or (Balkenborg & Schlag) ES-set, which would certainly be harder to figure out without the concept of robustness against indirect invasions.

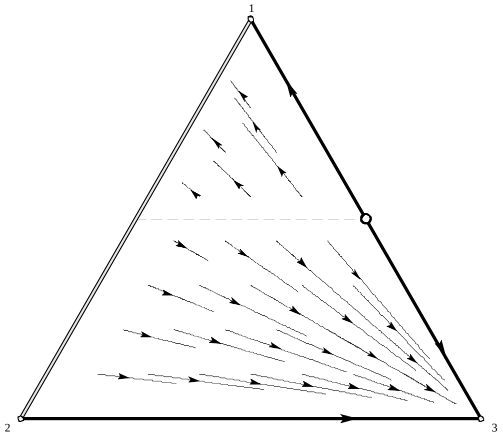
6 Acknowledgements

Phase plots were made with Dynamo Version 1.4.1 by William H. Sandholm and Emin Dokumaci (see Sandholm & Dokumaci, 2007). This is an absolutely great service to users. I would also like to thank Larry Samuelson for suggesting to present the concept of robustness against indirect invasions separate from the results concerning repeated games, so that they can be applied to extensive form games in general, and for suggesting to link them to ES-sets or dynamics. I would also like to thank Ido Polak for the derivation in Appendix C and Gerard van der Laan and Peter Spreij.

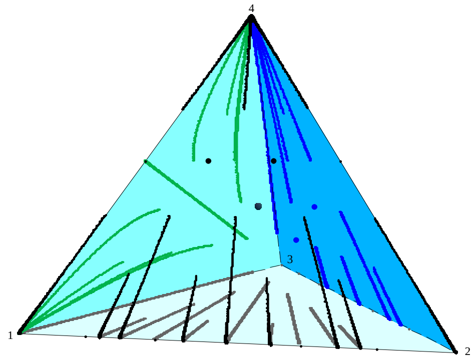
A Phase plots for examples



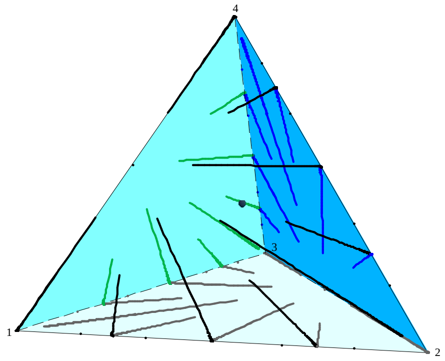
Example 2a



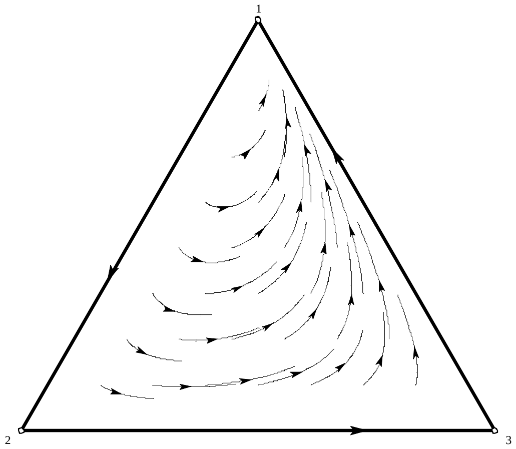
Example 2b



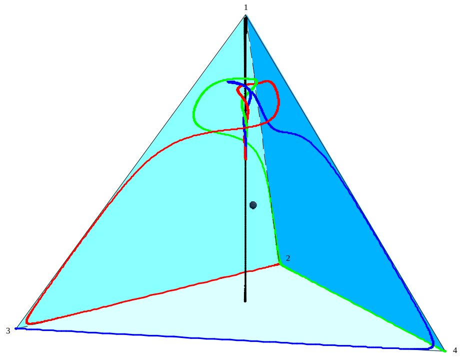
Example 4



Example 5



Example 7



Example 8

B Alternative proof

The following theorem combines Proposition 4 and Lemmas 6 and 7 in one result. The proof uses the approach from Theorem 15.5 in Hofbauer & Sigmund (1988), which is the same as Theorem 6.4.1 in Hofbauer & Sigmund (1998) or Propositions 2.5 and 2.6 in Weibull (1995) .

Theorem 13 *If there are only finitely many pure strategies, and x is strongly robust against indirect invasions, then $S_{NM}(x)$ is an ES-set.*

Proof. The crucial step is to show that any element $y \in S_{NM}(x)$ has a neighbourhood U_y such that $\pi(y,t) > \pi(t,t)$ for all strategies $t \in U \backslash S_{NM}(x)$. Because $y \in S_{NM}(x)$ implies that y is also strongly robust against indirect invasions, and that $S_{NM}(x) = S_{NM}(y)$, we can, without loss of generality, restrict ourselves to proving that x has a neighbourhood U such that $\pi(x,t) > \pi(t,t)$ for all strategies $t \in U \backslash S_{NM}(x)$.

In order to show that this indeed holds, if x is SRII, we use ingredients from the proofs of propositions 2.5 and 2.6 in Weibull (1996, p43-46), that reproduce results from Hofbauer, Schuster & Sigmund (1979). We do have to build up the crucial set Z_x of boundary faces of the unit simplex that do *not* contain x in two steps though. The reason is that in our case this set can also contain neutral mutants, which implies we cannot directly apply the arguments used in this proof for the case where x is an ESS (see also the discussion below this proof).

Let $Z_{x,W}$ be the union of all boundary faces of the unit simplex that do not contain x, nor any of its (direct!) neutral mutants; that is,

$$Z_{x,W} = \{z \in \Delta \mid \forall y \in S_E(x), \exists\, i \text{ such that } y_i > 0 \text{ and } z_i = 0\}$$

Please note that $x \in S_E(x)$. This implies that if there are no neutral mutants, and hence x is an ESS, the above definition immediately gives Z_x, as defined in Weibull (1995, p44) and as reproduced below. Also it is clear that $Z_{x,W} \subset S_W(x)$, since x is SRII and the equal performers, that is, strategies in $S_E(x)$, are excluded by the definition.

Define the function $f_x(\epsilon, z)$ that indicates, as a function of the mutant z and the share ϵ of this mutant in the population, how much better, or how much worse, x does against the post-entry mix than z.

$$f_x(\epsilon, z) = \pi(x,x) - \pi(z,x) - \epsilon\left(\pi(x,x) - \pi(z,x) + \pi(z,z) - \pi(x,z)\right)$$

If there is a $\delta \in [0,1]$ such that x outperforms z as long as $\epsilon \in (0,\delta)$, then this is an invasion barrier. The highest possible invasion barrier for x against z is denoted by $b_x(z)$. Formally:

$$b_x(z) = \sup\{\delta \in [0,1] : f_x(\epsilon, z) > 0\ \forall \epsilon \in (0,\delta)\}$$

Let Z_x be the union of all boundary faces of the unit simplex that do not contain x; that is,

$$Z_x = \left\{z \in \Delta \mid \exists\ i \text{ such that } x_i > 0 \text{ and } z\left(e^i\right) = 0\right\}$$

There are two things worth noticing. The first is that the function $b_x(z)$ is continuous on $Z_x \backslash S_E(x)$; because x is strongly robust against indirect invasions, $S_B(x) = \varnothing$, and therefore $Z_x \backslash S_E(x) \subset S_W(x)$. Therefore, for x and z fixed, $f_x(\epsilon, z) = 0$ for at most one ϵ, which we denote by ϵ_0. If $\epsilon_0 \in (0,1)$, then $\pi(x,x) > \pi(z,x)$ and $\pi(z,z) > \pi(x,z)$, and hence $b_x(z) = \epsilon_0 = \frac{\pi(x,x)-\pi(z,x)}{\pi(x,x)-\pi(z,x)+\pi(z,z)-\pi(x,z)}$. Otherwise, $b_x(z) = 1$. This makes b_x a continuous function.

The second thing worth noticing is that any $t \in Z_x$ can be written as a convex combination of an element of $Z_{x,W}$ and a neutral mutant of x already in Z_x; for any $t \in Z_x$, there is an $z \in Z_{x,W}$, a $y \in S_E(x) \cap Z_x$ and an $a \in [0,1]$ such that $t = az + (1-a)y$. So if indeed $z \in Z_{x,W}$ and $y \in S_E(x) \cap Z_x$, then straightforward algebra, using $\pi(x,x) = \pi(y,x)$ and $\pi(x,y) = \pi(y,y)$, leads to

$$\begin{aligned}
f_x(\epsilon,t) &= \pi(x,x) - \pi(t,x) - \epsilon\left(\pi(x,x) - \pi(t,x) + \pi(t,t) - \pi(x,t)\right) \\
&= a\left(\pi(x,x) - \pi(z,x)\right) - \epsilon\left(\begin{array}{c} a\left(\pi(x,x) - \pi(z,x)\right) + \\ a^2\pi(z,z) + a(1-a)\left\{\pi(z,y) + \pi(y,z)\right\} + (1-a)^2\pi(y,y) - \\ \left(a\pi(x,z) + (1-a)\pi(x,y)\right) \end{array}\right) \\
&= a\left\{\pi(x,x) - \pi(z,x) - \epsilon\left(\begin{array}{c} \pi(x,x) - \pi(z,x) + \pi(z,z) - \pi(x,z) + \\ (1-a)\left(\pi(z,y) - \pi(z,z) + \pi(y,z) - \pi(y,y)\right) \end{array}\right)\right\}
\end{aligned}$$

Now first we show that b_x attains a minimum larger than 0 on $Z_x \backslash S_E(x)$. In order to do so, suppose that b_x does not, which implies that there is a sequence $t^1, t^2, \ldots$, with $t^i \in Z_x \backslash S_E(x)\ \ \forall\ i$, such that $\lim_{i\to\infty} b_x\left(t^i\right) = 0$. These t^i can be written as $t^i = a_i z^i + (1-a_i)y^i$ with $z^i \in Z_{x,W}$, $y^i \in S_E(x) \cap Z_x$ and $a_i \in (0,1]$. This implies that there is an n such that $0 < b_x\left(t^i\right) < 1$ for all $i \geq n$ and hence that

$$b_x\left(t^i\right) = \frac{\pi(x,x) - \pi\left(z^i,x\right)}{\left\{\begin{array}{c} \pi(x,x) - \pi\left(z^i,x\right) + \pi\left(z^i,z^i\right) - \pi\left(x,z^i\right) + \\ (1-a_i)\left(\pi\left(z^i,y^i\right) - \pi\left(z^i,z^i\right) + \pi\left(y^i,z^i\right) - \pi\left(y^i,y^i\right)\right) \end{array}\right\}}$$

Also both nominator and denominator must be larger than 0 for all $i \geq n$. Because π is bounded on Z_x, we know that the assumption that $\lim_{i\to\infty} b_x\left(t^i\right) = 0$ implies that $\lim_{i\to\infty} \pi(x,x) - \pi\left(z^i,x\right) = 0$. The sequence z^i, however, must have an accumulation point z^* in $Z_{x,W}$. We can therefore make a subsequence z^j such that $z^j \to z^*$, and since $\pi(x,y) = x^T A y$ is a bilinear function, this implies that $\pi(x,x) - \pi(z^*,x) = 0$.

The sequence y^j in turn must have an accumulation point in $S_E(x) \cap Z_x$. We can therefore make a further subsequence y^k such that $y^k \to y^*$.

First we show that $\lim_{k\to\infty} a_k = 0$. Suppose this is not the case, and $0 < \lim_{k\to\infty} a_k = \alpha^* \leq 1$. Then we can take $t^* = a^* z^* + (1-a^*)y^*$. Around this t^* we can construct a compact subset of $Z_x \cap S_W(x)$ on which b_x is a continuous function (this is just the familiar argument for the normal ESS case). This implies that $b_x(t^*) = \lim_{k\to\infty} b_x(t^k) = 0$. But this contradicts that $t^* \in S_W(x)$. Hence $\lim_{k\to\infty} a_k = 0$.

Now we focus on convex combinations of any point z on $Z_{x,W}$ for which $\pi(x,x) - \pi(z,x) = 0$, and any neutral mutant $y \in S_E(x) \cap Z_x$. If we take $t = az + (1-a)y$ in the function $f_x(\epsilon, t)$ as worked out above, we find, since $\pi(x,x) - \pi(z,x) = 0$, that

$$f_x(\epsilon,t) = a\left\{-\epsilon\left(\begin{array}{c}\pi(z,z) - \pi(x,z) + \\ (1-a)(\pi(z,y) - \pi(z,z) + \pi(y,z) - \pi(y,y))\end{array}\right)\right\}$$

For all $a \in (0,1]$ we know that $t = az + (1-a)y \in S_W(x)$, because $z \in Z_{x,W}$ and $y \in S_E(x) \cap Z_x$. Hence we know that

$$\pi(z,z) - \pi(x,z) + (1-a)(\pi(z,y) - \pi(z,z) + \pi(y,z) - \pi(y,y)) < 0 \text{ for all } a \in (0,1],$$

and in particular that

$$\pi(z,z) - \pi(x,z) + \pi(z,y) - \pi(z,z) + \pi(y,z) - \pi(y,y) \leq 0$$

or

$$(1) \qquad \pi(z,y) + \pi(y,z) \leq \pi(x,z) + \pi(y,y)$$

Now we return to the subsequence $t^k = a_k z^k + (1-a_k) y^k$ and split the z^k's and y^k's themselves in two parts as well.

Because $y^k \to y^*$, we can rewrite the sequence of y^k's as follows:

$$y^k = \beta_k \widehat{y}^k + (1-\beta_k) y^*, \text{ with } \widehat{y}^k \in S_E(x) \cap Z_x \forall k \text{ and } \lim_{k\to\infty} \beta_k = 0.$$

As both y^k and y^* are neutral mutants, we know that $\pi(y^k, x) = \pi(y^*, x) = \pi(x,x)$. This implies that also $\pi(\widehat{y}^k, x) = \pi(x,x)$. Also from being neutral mutants we know that $\pi(y^k, y^k) = \pi(x, y^k)$ and $\pi(y^*, y^*) = \pi(x, y^*)$. If we would let $y^k(\beta)$ denote $\beta\widehat{y}^k + (1-\beta) y^*$, then one could also write that as:

$$\pi(y^k(\beta_k), y^k(\beta_k)) = \pi(x, y^k(\beta_k)) \text{ and } \pi(y^k(0), y^k(0)) = \pi(x, y^k(0))$$

Now x is SRII, and, since we found that $\pi(\widehat{y}^k, x) = \pi(y^*, x) = \pi(x,x)$, we also know that $\pi(y^k(\beta), x) = \pi(x,x)$ for all $\beta \in [0,1]$. Together this implies that for any $\beta \in [0,1]$ it must be true that $\pi(y^k(\beta), y^k(\beta)) \leq \pi(x, y^k(\beta))$. With equality for $\beta = \beta_k > 0$ and for $\beta = 0$, it must be that $\pi(y^k(\beta), y^k(\beta)) = \pi(x, y^k(\beta))$ for all $\beta \in [0,1]$. This implies that:

$$(2) \qquad \pi(\widehat{y}^k, y^*) = \pi(y^*, y^*) = \pi(x, y^*) \text{ and } \pi(\widehat{y}^k, \widehat{y}^k) = \pi(y^*, \widehat{y}^k) = \pi(x, \widehat{y}^k) \forall k$$

The z^k's will be separated in two parts slightly differently:

$$z^k = \gamma_k \widehat{z}^k + (1-\gamma_k)\, \widetilde{z}^k, \text{ with } \widetilde{z}^k \in Z_{x,W} \text{ and } \pi(x,x) - \pi\left(\widetilde{z}^k, x\right) = 0,$$
$$\text{and with } \widehat{z}^k \in Z_{x,W} \text{ and } \pi(x,x) - \pi\left(\widehat{z}^k, x\right) > 0 \text{ and } \lim_{k\to\infty} \gamma_k = 0.$$

This is possible since we know that for the limit z^* equality holds: $\pi(x,x) - \pi(z^*, x) = 0$.

For any $t^k = a_k z^k + (1-a_k)\, y^k$ in the sequence we can therefore compute the maximum invasion barrier

$$b_x\left(t^k\right) = \frac{\pi(x,x) - \pi\left(z^k, x\right)}{\left\{\begin{array}{c} \pi(x,x) - \pi\left(z^k,x\right) + a_k \pi\left(z^k,z^k\right) - \pi\left(x,z^k\right) + \\ (1-a_k)\left(\pi\left(z^k,y^k\right) + \pi\left(y^k,z^k\right) - \pi\left(y^k,y^k\right)\right) \end{array}\right\}} =$$

using $\pi(x,x) - \pi\left(\widetilde{z}^k, x\right) = 0$ we find that this equals

$$= \frac{\gamma_k\left(\pi(x,x) - \pi\left(\widehat{z}^k, x\right)\right)}{\left\{\begin{array}{c} \gamma_k\left(\pi(x,x) - \pi\left(\widehat{z}^k,x\right)\right) + a_k \pi\left(z^k,z^k\right) - \pi\left(x,z^k\right) + \\ (1-a_k)\left(\pi\left(z^k,y^k\right) + \pi\left(y^k,z^k\right) - \pi\left(y^k,y^k\right)\right) \end{array}\right\}} >$$

and because $b_x\left(t^k\right) \in (0,1)\, \forall k \geq n$, this is larger than

$$> \frac{\gamma_k \min_{xx-\widehat{z}x}}{\left\{\begin{array}{c} \gamma_k \min_{xx-\widehat{z}x} + a_k \pi\left(z^k,z^k\right) - \pi\left(x,z^k\right) + \\ (1-a_k)\left(\pi\left(z^k,y^k\right) + \pi\left(y^k,z^k\right) - \pi\left(y^k,y^k\right)\right) \end{array}\right\}} >$$

where $\min_{xx-\widehat{z}x} = \min_i \pi(x,x) - \pi\left(e^i, x\right) > 0$ (if this were 0, then $b_x\left(t^k\right)$ would not be larger than 0).

We will now work out the denominator term by term, using (2). This is not a very attractive bit of algebra, bit I see no way to avoid it.

$$\begin{array}{c} \gamma_k \min_{xx-\widehat{z}x} + a_k \left\{(\gamma_k)^2 \pi\left(\widehat{z}^k,\widehat{z}^k\right) + \gamma_k(1-\gamma_k)\left(\pi\left(\widehat{z}^k,\widetilde{z}^k\right) + \pi\left(\widetilde{z}^k,\widehat{z}^k\right)\right) + (1-\gamma_k)^2 \pi\left(\widetilde{z}^k,\widetilde{z}^k\right)\right\} \\ -\gamma_k \pi\left(x,\widehat{z}^k\right) - (1-\gamma_k)\,\pi\left(x,\widetilde{z}^k\right) + \\ (1-a_i)\left\{\begin{array}{c} \beta_k\gamma_k\pi\left(\widehat{z}^k,\widehat{y}^k\right) + (1-\beta_k)\gamma_k\pi\left(\widehat{z}^k,y^*\right) + \beta_k(1-\gamma_k)\pi\left(\widetilde{z}^k,\widehat{y}^k\right) + (1-\beta_k)(1-\gamma_k)\pi\left(\widetilde{z}^k,y^*\right) \\ +\beta_k\gamma_k\pi\left(\widehat{y}^k,\widehat{z}^k\right) + (1-\beta_k)\gamma_k\pi\left(y^*,\widehat{z}^k\right) + \beta_k(1-\gamma_k)\pi\left(\widehat{y}^k,\widetilde{z}^k\right) + (1-\beta_k)(1-\gamma_k)\pi\left(y^*,\widetilde{z}^k\right) \\ -\beta_k\pi\left(\widehat{y}^k,\widehat{y}^k\right) - (1-\beta_k)\,\pi\left(y^*,y^*\right) \end{array}\right\} = \end{array}$$

which we rearrange as

$$= \begin{array}{c} \gamma_k \min_{xx-\widehat{z}x} + a_k \left\{(\gamma_k)^2 \pi\left(\widehat{z}^k,\widehat{z}^k\right) + \gamma_k(1-\gamma_k)\left(\pi\left(\widehat{z}^k,\widetilde{z}^k\right) + \pi\left(\widetilde{z}^k,\widehat{z}^k\right)\right) + (1-\gamma_k)^2 \pi\left(\widetilde{z}^k,\widetilde{z}^k\right)\right\} \\ -\gamma_k \pi\left(x,\widehat{z}^k\right) - a_k(1-\gamma_k)\,\pi\left(x,\widetilde{z}^k\right) + \\ (1-a_i)\left\{\begin{array}{c} \beta_k\gamma_k\pi\left(\widehat{z}^k,\widehat{y}^k\right) + (1-\beta_k)\gamma_k\pi\left(\widehat{z}^k,y^*\right) + \beta_k(1-\gamma_k)\pi\left(\widetilde{z}^k,\widehat{y}^k\right) + (1-\beta_k)(1-\gamma_k)\pi\left(\widetilde{z}^k,y^*\right) \\ +\beta_k\gamma_k\pi\left(\widehat{y}^k,\widehat{z}^k\right) + (1-\beta_k)\gamma_k\pi\left(y^*,\widehat{z}^k\right) + \beta_k(1-\gamma_k)\pi\left(\widehat{y}^k,\widetilde{z}^k\right) + (1-\beta_k)(1-\gamma_k)\pi\left(y^*,\widetilde{z}^k\right) \\ -\beta_k\pi\left(\widehat{y}^k,\widehat{y}^k\right) - (1-\beta_k)\,\pi\left(y^*,y^*\right) - (1-\gamma_k)\,\pi\left(x,\widetilde{z}^k\right) \end{array}\right\} < \end{array}$$

Because $\widetilde{z}^k \in Z_{x,W}$ and $\pi\left(x,x\right) - \pi\left(\widetilde{z}^k, x\right) = 0 \forall k$, we know that $\pi\left(\widetilde{z}^k, \widetilde{z}^k\right) - \pi\left(x, \widetilde{z}^k\right) < 0$ and that $\pi\left(\widetilde{z}^k, y\right) + \pi\left(y, \widetilde{z}^k\right) \leq \pi\left(x, \widetilde{z}^k\right) + \pi\left(y,y\right)$ if y is a neutral mutant (see (1)). This implies that the denominator is smaller than

$$
\begin{array}{ll}
 & \gamma_k \min_{xx-\widehat{z}x} + a_k \left\{ \left(\gamma_k\right)^2 \pi\left(\widehat{z}^k, \widehat{z}^k\right) + \gamma_k \left(1-\gamma_k\right) \left(\pi\left(\widehat{z}^k, \widetilde{z}^k\right) + \pi\left(\widetilde{z}^k, \widehat{z}^k\right)\right) - \gamma_k \left(1-\gamma_k\right) \pi\left(\widetilde{z}^k, \widetilde{z}^k\right) \right\} \\
 & \qquad - \gamma_k \pi\left(x, \widehat{z}^k\right) + \\
< & \left(1-a_i\right) \left\{ \begin{array}{c} \beta_k \gamma_k \pi\left(\widehat{z}^k, \widehat{y}^k\right) + \left(1-\beta_k\right) \gamma_k \pi\left(\widehat{z}^k, y^*\right) \\ + \beta_k \gamma_k \pi\left(\widehat{y}^k, \widehat{z}^k\right) + \left(1-\beta_k\right) \gamma_k \pi\left(y^*, \widehat{z}^k\right) \\ - \beta_k \gamma_k \pi\left(\widehat{y}^k, \widehat{y}^k\right) - \left(1-\beta_k\right) \gamma_k \pi\left(y^*, y^*\right) \end{array} \right\}
\end{array}
$$

Returning to the invasion barrier, this implies that

$$b_x\left(t^k\right) >$$

$$
> \frac{\gamma_k \min_{xx-\widehat{z}x}}{\left\{ \begin{array}{c} \gamma_k \min_{xx-\widehat{z}x} + \gamma_k a_k \left\{ \gamma_k \pi\left(\widehat{z}^k, \widehat{z}^k\right) + \left(1-\gamma_k\right) \left(\pi\left(\widehat{z}^k, \widetilde{z}^k\right) + \pi\left(\widetilde{z}^k, \widehat{z}^k\right)\right) - \left(1-\gamma_k\right) \pi\left(\widetilde{z}^k, \widetilde{z}^k\right) \right\} \\ - \gamma_k \pi\left(x, \widehat{z}^k\right) + \\ \gamma_k \left(1-a_i\right) \left\{ \begin{array}{c} \beta_k \pi\left(\widehat{z}^k, \widehat{y}^k\right) + \left(1-\beta_k\right) \pi\left(\widehat{z}^k, y^*\right) \\ + \beta_k \pi\left(\widehat{y}^k, \widehat{z}^k\right) + \left(1-\beta_k\right) \pi\left(y^*, \widehat{z}^k\right) \\ - \beta_k \pi\left(\widehat{y}^k, \widehat{y}^k\right) - \left(1-\beta_k\right) \pi\left(y^*, y^*\right) \end{array} \right\} \end{array} \right\}} >
$$

All these payoffs can be replaced by their maximum or minimum values, which exist, if we take these maxima and minima over the obvious different compact parts of the simplex.

$$
> \frac{\gamma_k \min_{xx-\widehat{z}x}}{\left\{ \begin{array}{c} \gamma_k \min_{xx-\widehat{z}x} + \gamma_k a_k \left[\gamma_k \max_{\widehat{z},\widehat{z}} + \left(1-\gamma_k\right) \left(\max_{\widehat{z},\widetilde{z}} + \max_{\widetilde{z},\widehat{z}}\right) + \left(1-\gamma_k\right) \max_{\widetilde{z},\widetilde{z}} \right] - \gamma_k \min_{x,\widehat{z}} + \\ \gamma_k \left(1-a_i\right) \left\{ \begin{array}{c} \beta_k \max_{\widehat{z},\widehat{y}} + \left(1-\beta_k\right) \max_{\widehat{z},y^*} \\ \beta_k \max_{\widehat{y},\widehat{z}} + \left(1-\beta_k\right) \max_{y^*,\widehat{z}} \\ - \beta_k \min_{\widehat{y},\widehat{y}} - \left(1-\beta_k\right) \min_{y^*,y^*} \end{array} \right\} \end{array} \right\}}
$$

This gives us a lower bound of $b_x\left(t^k\right)$ for each k. But for $\alpha_k \to 0$, $\beta_k \to 0$ and $\gamma_k \to 0$, this sequence of lower bounds converges to

$$\frac{\min_{xx-\widehat{z}x}}{\min_{xx-\widehat{z}x} - \min_{x,\widehat{z}} + \max_{\widehat{z},y^*} + \max_{y^*,\widehat{z}} - \min_{y^*,y^*}} > 0$$

This implies that also $\lim_{k\to\infty} b_x\left(t^k\right) > 0$ which contradicts the assumption that $\lim_{i\to\infty} b_x\left(t^i\right) = 0$. Therefore we conclude that b_x does attain a minimum larger than 0 on $Z_x \backslash S_E\left(x\right)$.

From here, we can follow the "only if" part of proposition 2.6 in Weibull (1995) quite closely again. Let $\overline{\epsilon}$ be the minimum of b_x on $Z_x \backslash S_E\left(x\right)$ (x's uniform invasion barrier). Let

$$V = \left\{t \in \Delta : t = \epsilon y + \left(1-\epsilon\right) x \text{ for some } y \in Z_x \text{ and } \epsilon \in \left[0, \overline{\epsilon}\right)\right\}$$

Since Z_x is a closed set not containing x, there exists a neighbourhood U of x such that $U \cap \Delta \subset V$. Suppose that $y \neq x, y \in U \cap \Delta \backslash S_E(x)$. Then $y \in V \backslash S_E(x)$, and $\pi(y, \epsilon y + (1-\epsilon)x) < \pi(x, \epsilon y + (1-\epsilon)x)$, since $\overline{\epsilon}$ is the minimum of b_x on $Z_x \backslash S_E(x)$. By bilinearity of π this inequality is equivalent with $\pi(\epsilon y + (1-\epsilon)x, \epsilon y + (1-\epsilon)x) < \pi(x, \epsilon y + (1-\epsilon)x)$, which we rewrite as $\pi(t,t) < \pi(x,t)$.

The fact that x has a neighbourhood U such that $\pi(x,t) > \pi(t,t)$ for all strategies $t \in U \cap (S_{NM}(x))^C$ is enough to show that $S_{NM}(x)$ is an ES*-set. But since $S_{NM}(x)$ equals $S_E(x)$ locally, and since for t within $S_E(x)$ we have $\pi(x,t) = \pi(t,t)$, we find that $S_{NM}(x)$ is also an ES-set. ■

C Fixation probabilities

The matrix between strategies B and C is

$$\begin{matrix} 1 & 0 \\ 2 & 1 \end{matrix}$$

This gives a fixation probability (see for instance Nowak, 2006, page 110) of C in B of

$$\rho_{B\to C} = \frac{1}{1 + \sum\limits_{k=1}^{N-1} \prod\limits_{i=1}^{k} \frac{N-i-1}{2N-i-1}}$$

Focussing on the sum, we find - with some suggestive notation - that

$$\begin{aligned}
\sum_{k=1}^{N-1} \prod_{i=1}^{k} \frac{N-i-1}{2N-i-1} &= \frac{N-2}{2N-2} + \frac{N-2}{2N-2}\frac{N-3}{2N-3} + \frac{N-2}{2N-2}\frac{N-3}{2N-3}\frac{N-4}{2N-4} + ... + \frac{N-2}{2N-2}\frac{N-3}{2N-3}...\frac{1}{N+1} \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{2N-3}\left(1+\frac{N-4}{2N-4}\left(...\left(1+\frac{3}{N+3}\left(1+\frac{2}{N+2}\left(1+\frac{1}{N+1}\right)\right)\right)...\right)\right)\right) \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{2N-3}\left(1+\frac{N-4}{2N-4}\left(...\left(1+\frac{3}{N+3}\left(1+\frac{2}{N+2}\left(\frac{N+2}{N+1}\right)\right)\right)...\right)\right)\right) \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{2N-3}\left(1+\frac{N-4}{2N-4}\left(...\left(1+\frac{3}{N+3}\left(1+\frac{2}{N+1}\right)\right)...\right)\right)\right) \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{2N-3}\left(1+\frac{N-4}{2N-4}\left(...\left(1+\frac{3}{N+3}\left(\frac{N+3}{N+1}\right)\right)...\right)\right)\right) \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{2N-3}\left(1+\frac{N-4}{2N-4}\left(...\left(1+\frac{3}{N+1}\right)...\right)\right)\right) \\
&\vdots \\
&= \frac{N-2}{2N-2}\left(1+\frac{N-3}{N+1}\right) \\
&= \frac{N-2}{2N-2}\left(\frac{2N-2}{N+1}\right) = \frac{N-2}{N+1}
\end{aligned}$$

Thereby we find that

$$\rho_{B\to C} = \frac{1}{1+\sum_{k=1}^{N-1}\prod_{i=1}^{k}\frac{N-i-1}{2N-i-1}} = \frac{1}{1+\frac{N-2}{N+1}} = \frac{1}{\frac{2N-1}{N+1}} = \frac{N+1}{2N-1}$$

(This derivation is from Ido Polak)

References

[1] Balkenborg, D. and K.H. Schlag. (2001). "Evolutionarily stable sets, " *International Journal of Game Theory* **29**, 571–595.

[2] Balkenborg, D. and K.H. Schlag. (2007). "On the evolutionary selection of sets of Nash equilibria," *Journal of Economic Theory* **133**, 295-315.

[3] Binmore, K. and L. Samuelson (1994). "Drift," *European Economic Review* **38**, 859-867.

[4] Cressman, R. (1992). "Evolutionarily stable sets in symmetric extensive two-person games, " *Mathematical Biosciences* **108**, 179-201.

[5] Fudenberg, D. and L. Imhof (2006). "Imitation processes with small mutations," *Journal of Economic Theory* **131**, 251-262

[6] Hofbauer, J., P. Schuster and K. Sigmund. (1979). "A note on evolutionary stable strategies and game dynamics," *Journal of Theoretical Biology* **81**, 609-612.

[7] Maynard Smith, J. and G.R. Price (1973). "The logic of animal conflict," *Nature* **246**, 15-18.

[8] Nowak, M.A. (2006) *Evolutionary dynamics: exploring the equations of life.* Harvard University Press, Cambridge, MA.

[9] Sandholm, W.H. and E. Dokumaci. (2007). Dynamo: Phase Diagrams for Evolutionary Dynamics. Software suite. http://www.ssc.wisc.edu/~whs/dynamo.

[10] Hofbauer, J. and K. Sigmund. (1988). *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, Cambridge.

[11] Hofbauer, J. and K. Sigmund. (1998). *Evolutionary Games and Population Dynamics.* Cambridge University Press, Cambridge.

[12] Swinkels, J.M. (1992a). "Evolutionary stability with equilibrium entrants," *Journal of Economic Theory* **57**, 306-332.

[13] Swinkels, J.M. (1992b). “Evolution and strategic stability: From Maynard Smith to Kohlberg and Mertens,” *Journal of Economic Theory* **57**, 333-342.

[14] Thomas, B. (1985). “On evolutionarily stable sets,” *Journal of Mathemathical Biology* **22**, 105-115.

[15] Van Veelen, M. (2006). “Evolution of strategies in repeated games with discounting,” Tinbergen Institute discussion paper.

[16] Weibull, J.W. (1995). *Evolutionary Game Theory*. Cambridge MA: MIT Press.