Equilibrium selection in repeated games[1]

Sam Jindani[2]

26 January 2020

The folk theorem tells us that a wide range of payoffs can be sustained as equilibria in an infinitely repeated game. But this creates a problem of indeterminacy: how can we predict which of the many possible payoffs players will receive? Existing results about learning in repeated games suggest that players may converge to an equilibrium, but fail to say anything about selection between equilibria. I propose a stochastic learning rule that selects a subgame-perfect equilibrium of the repeated game in which the payoffs correspond to the Kalai–Smorodinsky bargaining solution. This result highlights the parallels between the canonical bargaining problem and equilibrium selection in repeated games.

JEL classification: C70, C72, C73

Keywords: repeated games, learning, bargaining problem, equilibrium selection

1 Introduction

The folk theorem for infinitely repeated games[3] is a powerful and fundamental result, but it creates a problem of indeterminacy: if any individually rational payoff profile can be supported when players are sufficiently patient, how can we predict what the outcome of the game will be? Yet models of learning for repeated-game strategies focus on convergence to equilibrium, rather than selection between equilibria.

1 I thank my supervisors, Peyton Young and Tom Norman, as well as Alan Beggs, Yuval Heller, Ed Hopkins and Meg Meyer for their insightful comments. I also thank seminar participants at the Transatlantic Theory Workshop at Northwestern University and at the Gorman Workshop at the University of Oxford. I gratefully acknowledge financial support from the Economic and Social Research Council (award number ES/J500112/1) and the University of Oxford Department of Economics.

2 Department of Economics, University of Oxford.
Email: sam.jindani@gmail.com.

3 Aumann 1976; Rubinstein 1979; Fudenberg and Maskin 1986.

Consider two players playing an infinitely repeated game with discounting and perfect monitoring. In Kalai and Lehrer's seminal analysis (1993), the players start with a belief about their opponent's strategy; they best respond given their belief and update it as the game is played. Each player is trying to guess how their opponent will behave, but their opponent's behaviour depends in turn on the opponent's guess about how the first player will behave. This is known as the *interactive learning problem*. Nevertheless, Kalai and Lehrer show that Bayesian players will converge to a Nash equilibrium of the repeated game, provided the priors satisfy a 'grain-of-truth' assumption: the initial beliefs must be compatible with the eventual play. The key insight is that rational learning, by itself, can lead to players correctly predicting their opponent's play, from which it follows that play will correspond to a Nash equilibrium.[4]

Near-rational learning can also yield convergence to equilibrium. In particular, Foster and Young (2003) consider agents who form bounded-memory beliefs (or hypotheses) about their opponent's strategy and play ε-best replies to those beliefs. They periodically test their hypotheses and form new ones if the observed behaviour differs too much from the expected behaviour. Foster and Young show that under this learning rule, players will tend to converge to a subgame-perfect equilibrium of the repeated game. The fact that players best respond with noise means that off-equilibrium paths can be visited and allows convergence to subgame-perfect equilibrium rather than simple Nash equilibrium.

However the literature on learning for repeated-game strategies does not address the question of selection between equilibria.[5] I present a model of near-rational learning in two-player, infinitely repeated games that yields sharp selection results. For any stage game, the learning rule selects a subgame-perfect equilibrium of the corresponding repeated game in which the payoffs received by each player correspond to the Kalai–Smorodinsky bargaining solution of the payoff space (Kalai and Smorodinsky 1975). The results are achieved under plausible assumptions about the players: they form beliefs based on evidence;

4 The grain-of-truth assumption is somewhat restrictive (Nachbar 1997, 2005), although the literature features a number of positive results for different settings or when one considers approximate Nash equilibria (Jordan 1995; Nyarko 1998; Sandroni 1998; Norman 2019).

5 This is in contrast to the literature on learning for stage-game actions, where there are well-established results about the selection of risk-dominant equilibria (Kandori, Mailath and Rob 1993; Young 1993) or efficient equilibria (Pradelski and Young 2012).

they are rational with high probability, in the sense of optimising given their beliefs; and the learning rules are uncoupled.[6]

The Kalai–Smorodinsky solution is the unique bargaining solution satisfying Pareto optimality, symmetry, invariance to affine transformations, and a condition called monotonicity. The monotonicity axiom replaces Nash's independence of irrelevant alternatives axiom (Nash 1950). For example, in the asymmetric prisoner's dilemma of figure 1, we can view the set of feasible and individually rational payoffs profiles of the game as the set of feasible outcomes of a bargaining problem, and we can view the minmax payoff profile $(0, 0)$ as the disagreement point. The Kalai–Smorodinsky payoff profile of the game is then $(2.4, 1.6)$. Note that this is on the Pareto frontier and gives more to the player with the higher maximum payoff.

The result highlights the similarities between the canonical bargaining problem and the problem of equilibrium selection in repeated games, both of which concern settings where there are a number of possible outcomes with different payoff implications for each player and the question is how the players will split the surplus. It is therefore not surprising that a solution concept in one framework should carry over to the other. The result also leads to a new interpretation of the Kalai–Smorodinsky solution as the organic outcome of a learning process.[7]

The model works as follows: Players form bounded-memory beliefs about their opponent's strategy based on recent histories of play. They reject beliefs when their opponent plays something that conflicts with their belief too often. When they hold a belief they usually play according to a bounded-memory strategy that is a best response to the belief, but sometimes make mistakes.[8] After rejecting a belief, they become uncertain and spend a few periods forming a new belief based on their opponent's recent actions. During this time

6 A learning rule is *uncoupled* if it does not require the player to have knowledge of her opponent's payoff function to implement it; it is *completely uncoupled* if it is uncoupled and does not require knowledge of the opponent's past actions (Young 2009). The rule in the present paper is uncoupled but not completely uncoupled.

7 Young (1998) shows that a model of learning for stage-game actions can converge to the Kalai–Smorodinsky solution in the context of a bargaining game.

8 Note that there is a distinction between the bounded-memory strategies that players follow when they hold a belief and the broader strategies that are determined by the learning rule. The latter are what is typically understood by 'strategy' in a repeated game. For the sake of simplicity, and since the latter will not explicitly play a part in the paper, I will refer to the bounded-memory strategies simply as strategies.

	C	D
C	$2, 2$	$-1, 3$
D	$5, -1$	$0, 0$

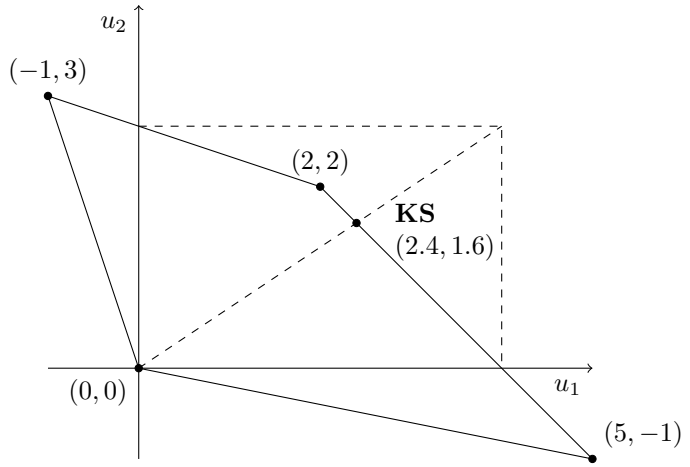

Figure 1: Game matrix and payoff space for the asymmetric prisoner's dilemma.

they experiment and play more randomly. Under this learning rule, states that do not correspond to a subgame-perfect equilibrium will tend to be unstable because they involve one or both players having a wrong belief. The assumption that players tend to play more randomly when they are uncertain is key in making the model tractable, since it ensures that any equilibrium can be easily reached from a state in which either player is uncertain.

The payoffs in the stable states are determined by the way in which players make mistakes. I assume that the probability of an agent making a mistake in a given state is decreasing in the expected discounted utility received in that state. That is, a more dissatisfied player is more likely to make mistakes than a satisfied one.[9] I show that under this assumption, the model selects the Kalai–Smorodinsky payoffs, in the sense that all stochastically stable states involve players choosing strategies that constitute a subgame-perfect equilibrium and that approximately yield the relevant payoffs.[10] This is because in a state in which one of the players expects a low utility, that player will make mistakes relatively frequently, rendering the state unstable. The use of stochastic stability as a solution concept is key to making the model tractable. The learning rule is uncoupled and does not impose any restrictions on players'

9 Experimental results establish a relationship between players' payoffs and their probability of deviating from best response. In a setup where players learn stage-game actions, Mäs and Nax (2016) find that a player who experiences a payoff decrease in one period is more likely to deviate in the following period.

10 A *stochastically stable* state of a stochastic process is, roughly speaking, one that recurs with non-negligible frequency in the long run as noise vanishes.

initial beliefs.

The learning rule in this paper builds on two learning rules in particular. First, the approach of Foster and Young (2003) discussed above also features players who form beliefs about their opponent's strategy given past play, are near rational given their beliefs, and reject them when they conflict with observed behaviour. In order to make the model more tractable and obtain equilibrium selection results, I rule out mixed strategies. This simplifies the analysis considerably, since it allows players to count mistakes rather than infer them probabilistically.[11] Second, the learning rule is also related to procedures by Young (2009) and by Pradelski and Young (2012) for stage-game actions. These rules share the feature that players can be in two different modes: one stable mode in which they play according to a fixed strategy with some deviations, and one mode of random search, in which they experiment with different strategies. The random search mode in those models is triggered by low payoffs, whereas in the present model it is brought about by the opponent's play not conforming to the player's beliefs. This is known as *fast and slow search* and has parallels in other fields, such as computer science, and in nature.[12]

The rest of this paper is organised as follows: in section 2, I set out the learning rule; in section 3, I state and prove the selection result.

2 The learning rule

Let $\mathscr{G}$ be a normal-form, two-player game with finite action sets A_1 and A_2 and payoff functions u_1 and u_2. Let a_1 and a_2 denote typical elements of A_1 and A_2, respectively. Define $A \coloneqq A_1 \times A_2$, with typical element a. Suppose players do not use mixed actions. This assumption buys us considerable tractability, as will be made clear below. The game $\mathscr{G}$ is repeated infinitely many times and players rank infinite payoff sequences according to the discounting criterion; that is, player i assigns utility $U_i(\mathbf{v}_i)$ to any sequence of payoffs $\mathbf{v}_i = (v_i^t)_{t=1}^{\infty}$,

11 A complication is that because the model both rules out mixed stage-game strategies and requires repeated-game strategies to be of bounded memory, the standard folk theorem does not apply. To guarantee the existence of equilibria supporting desired payoff pairs, I use a result due to Barlo, Carmona and Sabourian (2016).

12 For instance bees move from flower to flower in a small neighbourhood as long as they find sufficient nectar; but if the nectar yield becomes too low, they search for another neighbourhood further away. This is known as 'near-far search' (Motro and Shmida 1995).

where

$$U_i(\mathbf{v}_i) \coloneqq (1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}v_i^t,$$

where $\delta \in (0,1)$ is a constant shared by both players. Call the repeated game $\mathscr{H}$. Define the maximum and minimum payoffs in $\mathscr{G}$:

$$\bar{v} \coloneqq \max_{i\in\{1,2\},\, a\in A} u_i(a), \text{ and}$$
$$\underline{v} \coloneqq \min_{i\in\{1,2\},\, a\in A} u_i(a).$$

Let V be the convex hull of the set of feasible payoff pairs in $\mathscr{G}$. Define the minmax payoffs for each player as usual, except that mixed actions are disallowed. That is, player i's minmax payoff is

$$w_i \coloneqq \min_{a_j\in A_j}\max_{a_i\in A_i} u_i(a_i,a_j).$$

For the sake of convenience, let us normalise the payoffs such that $w_i = 0$ for all i. Let V^* be the set of feasible and strictly individually rational payoff pairs; namely

$$V^* \coloneqq \{v \in V : v \gg (0,0)\}.$$

Assume that $\mathscr{G}$ is such that V^* is non-empty. Define the maximum feasible payoff for each player:

$$\bar{v}_i^* \coloneqq \sup_{v\in V^*} v_i.$$

(Note that we have to take the supremum since V^* is not compact.) Then let x be the *Kalai–Smorodinsky payoff profile*; that is

$$x_i \coloneqq \arg\max_{v\in V^*}\min_{i=1,2}\frac{v_i}{\bar{v}_i^*}.$$

The Kalai–Smorodinsky payoff profile equalises the ratios of gains across players, *i.e.* $x_1/\bar{v}_1^* = x_2/\bar{v}_2^*$. Let

$$\rho \coloneqq \frac{x_i}{\bar{v}_i^*}.$$

Players follow a particular learning rule, which I will call R^ε, where ε is a small positive number that captures players' propensity for mistakes. The rule works as follows: In a given period, player i is in one of two *moods*. In the *certain* mood, she has a belief about her opponent's strategy and plays

according to a strategy that is a best response, but sometimes makes mistakes; in the *uncertain* mood, she is unsure of her opponent's strategy and attempts to guess it, and, because she is unsure, she is more prone to playing randomly.[13]

Players' beliefs and strategies are restricted to being of memory k; one can think of k as the *complexity* of players' strategies. Fix k to be some arbitrary strictly positive integer.

Definition 1. For any subgame of $\mathscr{H}$, the *ℓ-history* at that subgame is the ℓ-tuple of the action profiles played in the ℓ most recent periods.

Definition 2. A strategy is of *memory k* if for any two subgames with the same k-history, the strategy prescribes the same action.

As in Foster and Young 2003, restricting strategies to being of memory k is necessary in order to define the process generated by the learning rule as a finite Markov chain: transition probabilities must depend on a finite history, otherwise the number of states would be infinite. Let S_i be the set of memory-k strategies for player i, with typical element s_i.

I first describe how players act in each of the moods, before explaining how transitions between them occur. The two moods are:

> *Certain* $c(s_i, s_j)$: Player i has a belief about j's behaviour, which is a strategy $s_j \in S_j$. She also has a strategy $s_i \in S_i$ that is a best response to s_j in every subgame of $\mathscr{H}$.[14] In each period, i makes a mistake with probability $\varepsilon^{f(U_i/\bar{v}_i^*)}$, where $\varepsilon \in [0, 1)$, f is a strictly positive and strictly increasing function, and U_i is i's predicted continuation payoff i given s_i, s_j, and the current k-history. If a mistake is made, i plays an action from A_i at random with uniform probability.[15] If no mistake is made, i plays according to s_i.
>
> *Uncertain* $u(s_i)$: Each period, i plays an action from A_i according to a fully mixed distribution $p(s_i, h)$, where h is the current k-history.

13 Moods are simply states of the automaton that determines a player's behaviour. I use the term to distinguish from the states of the Markov chain that will be defined below, following Young 2009.

14 Note that there exists a best response to s_j that belongs to S_i.

15 I use the uniform distribution for ease of exposition. It suffices that the distribution is fully mixed, including in the limit as $\varepsilon \to 0$.

In the certain mood, the lower the player's continuation payoff, the more likely she is to make mistakes. This will imply that equilibria in which at least one player receives low utility will tend to be unstable. The distribution p in the uncertain mood is unimportant – for instance we could imagine that $p(s_i, \cdot)$ is always close to s_i. What matters is that it is fully mixed and that it does not depend on ε.[16] Then for ε small enough, mistakes in the uncertain state are arbitrarily more likely than mistakes in the certain state.

I now describe how transitions between moods occur. Fix strictly positive integers m and τ. In broad terms, a player in the certain mood looks back m periods to check her beliefs. If there were τ or fewer deviations relative to her beliefs in those m periods, she stays certain; otherwise, she becomes uncertain with positive probability. A player in the uncertain mood looks back m periods and with positive probability forms a new belief based on her opponent's actions in those periods. Thus m can be thought of as the players' *memory*, while τ is their *tolerance* for discrepancies between actions and their beliefs.

In order to determine whether an action in a given period was a deviation, players need to know what happened in the k periods prior to that. Therefore we need m to be larger than k. The larger m, the easier it is for players to identify wrong beliefs. If m is larger than k but not by much, then players may forget deviations too soon to notice that their belief is wrong. It turns out that

$$m > 2k + \tau$$

is sufficient for the results below. We also need

$$\tau > 2k.$$

This means that players should not be too sensitive to mistakes.

I assume that players include their own deviations in the count for transitioning to the uncertain state. This is not critical to the model but simplifies analysis. It is reasonable if one considers that players realise that their mistakes will tend to cause their opponent to change strategy whenever the opponent has a belief that is roughly correct.

16 More precisely, the distribution p can depend on ε as long as it is fully mixed in the limit as $\varepsilon \to 0$.

I will say that an m-history is *inconsistent* with a strategy profile if it contains more than τ deviations by either player:

Definition 3. Let h be an m-history. Let h' be the $(m-k)$-history consisting of the $m-k$ most recent action profiles in h. Let $s_1 \in S_1$ and $s_2 \in S_2$. Given h, count the deviations from s_1 and s_2 in h'. If the number of deviations is strictly greater than τ, h is *inconsistent* with the strategies s_1 and s_2; otherwise, h is *consistent* with s_1 and s_2. If the number is zero, h is *fully consistent* with s_1 and s_2.

Players will start to reject beliefs when the history is inconsistent with their beliefs, in the sense defined above. In particular, if a player's belief is correct and her opponent makes $\tau+1$ mistakes in a row, she will reject her belief with positive probability. Note that if mixed actions were allowed, players could not count mistakes. Instead, they would have to infer them probabilistically, which would significantly complicate the analysis (as in Foster and Young 2003).

Transitions occur as follows:

> *Certain* $c(s_i, s_j)$: As long as the most recent m-history is consistent with s_i and s_j, player i stays certain. Otherwise i goes to $u(s_i)$ with probability $b \in (0,1)$ each period.
>
> *Uncertain* $u(s_i)$: Each period, a transition occurs with probability $b' \in (0,1)$. If a transition occurs, pick a belief $s_j \in S_j$ according to a distribution $q(s_i, h)$, where h is the most recent m-history. Pick one of the best responses to s_j in S_i at random with uniform probability. Call the chosen strategy s_i. Player i goes to $c(s_i, s_j)$. Otherwise, i stays uncertain.

The probabilities b and b' are arbitrary. The distribution $q(s_i, h)$ can be any distribution provided that the probability of choosing the correct belief when one exists is strictly positive. In particular, $q(\cdot, h)$ could be such that the player is more likely to choose strategies that could have generated the opponent's actions in h, or could have generated actions similar to those ones.

Looking ahead, note that there are two ways in which a given pair of certain moods might be unstable: First, if the players' strategies do not constitute a subgame-perfect equilibrium – that is, if at least one of the players has an incorrect belief – then at some point one of the players is likely to become uncertain as her opponent plays actions that conflict with her belief. Second,

if at least one player has a low expected utility (relative to her maximum feasible payoff) then she is likely to make mistakes, which will eventually make her opponent become uncertain. Therefore a stable state will correspond to a subgame-perfect equilibrium in which payoffs are the Kalai–Smorodinsky payoffs.

Finally, suppose the function f is $\mathbb{R} \to \mathbb{R}^{++}$ and is continuous. Assume the range of f is small given τ, in the sense that

$$(\tau + 1)f(\underline{v}/\bar{v}) > \tau f(1). \tag{1}$$

(Recall that f is strictly increasing.) Thus a high utility reduces the probability of making a mistake, but not too much. To see what this implies, consider a given sequence of mistakes. The total probability of this sequence depends on both the number of mistakes in the sequence and the continuation payoffs of the players at the moment of making each mistake, and this probability is decreasing in both the number and the payoffs. Inequality 1 implies that the former effect dominates the latter. This will ensure that states that don't correspond to subgame-perfect equilibria are unstable relative to ones that do, even if they are states in which players expect high utility.

This completes the definition of the learning rule R^ε.

Suppose both players use R^ε. This determines the transition matrix of a Markov chain for which a state is a pair of moods, one for each player, and an m-history. Let M_i be the set of i's possible moods. Let $Z \coloneqq M_1 \times M_2 \times A^m$ be the set of states, with typical element z. Denote the transition matrix of the Markov chain by P^ε.

3 The result

I now state the result formally. In broad terms, I show that provided k and δ are large enough, the stochastically stable states of the process P^ε must involve (i) players' beliefs being correct, (ii) players' strategies constituting a subgame-perfect equilibrium, and (iii) players receiving discounted payoffs arbitrarily close to the Kalai–Smorodinsky payoff profile in every period.

When ε is strictly positive, the process P^ε is irreducible and aperiodic and therefore has a unique stationary distribution, μ^ε. The approach used in the stochastic stability literature is to consider the limit of this distribution as ε

becomes small, namely

$$\mu^* \coloneqq \lim_{\varepsilon \to 0} \mu^\varepsilon,$$

as μ^* is generally more tractable than μ^ε.[17] The *stochastically stable* states of P^ε are defined to be the states that have positive probability in μ^*. Since μ^ε corresponds to the long-run frequency distribution of P^ε, the stochastically stable states are the states that are observed in the long-run with non-negligible frequency when ε is small.

Let Ω be the set of recurrent classes of P^0, with typical element ω. Note that Ω is a set of subsets of Z. I will say that $\omega \in \Omega$ is a *limit set* of P^ε. Note that if at least one state of a limit set is stochastically stable, so are the others; I will say that a limit set is stochastically stable if it contains stochastically stable states. Note, moreover, that within a given limit set, players must always have the same strategies and beliefs, since any transition between moods happens via the uncertain mood, and, as we will see below, any state in which at least one player is uncertain is not in any limit set. I will say that players use *constant strategies* within a limit set.

The main result is as follows:

Theorem 1. In any stochastically stable limit set of P^ε, players' beliefs are correct and players use constant strategies that constitute a subgame-perfect equilibrium of $\mathscr{H}$. Moreover, for any $\lambda > 0$, the continuation payoff of each player i in every state of any stochastically stable limit set is at least $x_i - \lambda$, provided δ and k are large enough.

Thus the learning rule selects a subgame-perfect equilibrium that yields the Kalai–Smorodinsky payoff profile.

The condition on δ and k is to ensure that there exists an equilibrium that yields payoffs close enough to the Kalai–Smorodinsky payoff pair. Because mixed strategies are ruled out and memory is bounded, the standard Fudenberg and Maskin (1986) folk theorem does not apply. Instead, I make use of the following theorem, which is a reformulation of theorem 1 in Barlo, Carmona and Sabourian 2016:

Theorem A (Barlo, Carmona and Sabourian 2016)**.** For all $v \in V^*$ and $\lambda > 0$, there exists a k-memory pure subgame-perfect equilibrium of $\mathscr{H}$ such

17 The seminal paper in game theory is Foster and Young 1990. Ellison 2000 provides a good overview of the type of analysis used here.

that each player i players receive continuation payoffs within λ of v_i at every subgame.

The proof of theorem 1 proceeds by a series of lemmas. Throughout, suppose that δ is large enough to guarantee that at least one subgame-perfect equilibrium exists. (Mixed actions are disallowed, so there may not be a stage-game Nash equilibrium.)

First, I show that no state in which a player is uncertain can belong to a limit set of the Markov chain.

Lemma 1. Any state in which at least one of the players is uncertain is not in a limit set of P^ε.

Proof. First suppose that both players are uncertain. With positive probability under P^0, both players play in a way that is compatible with some subgame-perfect equilibrium s for m periods and then switch to the certain mood with strategies and beliefs corresponding to s. From this new state, the players stay certain, so the initial state is not in a recurrent class of P^0.

Next, suppose one player is uncertain and the other is certain. With positive probability under P^0, the uncertain player remains uncertain and plays actions that conflict with the certain player's belief for sufficiently long to make the history inconsistent with the belief. Then with positive probability the certain player becomes uncertain. Note that we require $m > \tau + k$ to guarantee that the uncertain player's actions count as mistakes up to τ periods in the future. Then we are in the case of the previous paragraph, so the initial state is not in a recurrent class of P^0. □

Next, I show that every subgame-perfect equilibrium corresponds to a limit set.

Lemma 2. Let s be a subgame-perfect equilibrium. Then there exists $\omega \in \Omega$ such that in every state of ω, players are certain with strategies s and correct beliefs, and the history is fully consistent with s.

Proof. Consider a state $z \in Z$ such that both players are certain with strategies s and correct beliefs, and the history is fully consistent with s. (Recall that an m-history is fully consistent with a strategy profile if there were no deviations from the strategy profile in the past $m - k$ periods.) Following z, no player leaves the certain mood under R^0 and the history stays fully consistent with

s. Either z is in a recurrent class of P^0 or the process eventually reaches a recurrent class of P^0, so the claim follows. □

Lemma 2 states that every subgame-perfect equilibrium corresponds to at least one limit set, but it is worth noting that a given subgame-perfect equilibrium may correspond to multiple limit sets. For instance, in the prisoner's dilemma the grim trigger equilibrium corresponds to two limit sets: one in which players cooperate and one in which players defect.

There may also be limit sets in which beliefs are incorrect, for instance when the discrepancies occur on histories that are not reached in the set, or if the interval between each discrepancy is large compared to the memory of the process. In lemma 4, we will see that such limit sets are less stable than those in which beliefs are correct.

Before turning to lemma 3, I introduce some further notation. (This follows Ellison 2000.) Let $z, z' \in Z$. The transition $z \to z'$ is *possible* if for all $\varepsilon > 0$, $P^\varepsilon(z, z') > 0$. If a transition $z \to z'$ is possible, define its *cost* to be the unique real number $c(z, z')$ such that

$$\lim_{\varepsilon \to 0} \frac{P^\varepsilon(z, z')}{\varepsilon^{c(z,z')}}$$

exists and is strictly positive. A finite sequence of states $(z_1, z_2, \ldots, z_n)$ is a *path* from z_1 to z_n if, for all $t < n$, $z_t \to z_{t+1}$ is possible. Its cost is the sum of the pairwise costs, namely

$$c(z_1, z_2, \ldots, z_n) \coloneqq \sum_{t=1}^{n-1} c(z_t, z_{t+1}).$$

Let $\omega, \omega' \in \Omega$. A sequence $(z_1, z_2, \ldots, z_n)$ is a path from ω to ω' if it is a path from z_1 to z_n and $z_1 \in \omega$ and $z_n \in \omega'$. Let $Q(\omega, \omega')$ be the set of all paths from ω to ω'. Define the *resistance* of the transition from ω to ω' to be

$$r(\omega, \omega') \coloneqq \min_{q \in Q(\omega, \omega')} c(q).$$

The resistance $r(\omega, \omega')$ is a measure of how likely the transition from ω to ω' is relative to the size of ε; the higher $r(\omega, \omega')$, the less likely the transition. As $\varepsilon \to 0$, transitions that cost 1 become arbitrarily more likely than transitions

that cost 2, say. Define the minimum outgoing resistance or *radius* of ω to be

$$r(\omega) \coloneqq \min_{\omega' \neq \omega} r(\omega, \omega').$$

The radius $r(\omega)$ is a measure of how difficult it is to exit ω.

In general, one can determine the stochastically stable limit sets by analysing the resistances of the various transitions. In this model, the analysis is simplified by the following fact:

Lemma 3. Let $\omega, \omega' \in \Omega$. Then

$$r(\omega, \omega') = r(\omega).$$

Proof. By definition, $r(\omega, \omega') \geq r(\omega)$. I show that ω' can be reached by a path that costs $r(\omega)$. Let z' be some state in ω'; denote its strategy profile by s' and its m-history by h'. Consider any path from ω to some other recurrent class that costs $r(\omega)$. This path must pass through a state in which at least one player is uncertain. Start by copying the path until such a state is reached. Call the uncertain player i. Then the following events occur without further cost: i stays in uncertain mood and plays in such a way as to make j switch from certain to uncertain; for m periods, both players play in such a way that the history is h'; in the following period, they become certain with the strategies and beliefs of z'. Then we have reached ω'. Thus $r(\omega, \omega') \leq r(\omega)$, and therefore $r(\omega, \omega') = r(\omega)$. □

Lemma 3 means that it is as easy to go from ω to ω' as it is to go from ω to ω'', in the sense that both transitions have the same resistance. This is because any transition out of ω involves passing through a state in which at least one player is uncertain, and from any such state there is a path to any limit set that has zero resistance. Lemma 3 will allow us to determine the stochastically stable states from the radiuses alone.

Let $\Omega^* \subseteq \Omega$ be the set of limit sets such that in every state players are certain with correct beliefs and the history is fully consistent with the strategies. Lemma 2 implies that Ω^* is non-empty (provided at least one subgame-perfect equilibrium exists).

I now show that limit sets in which beliefs are incorrect are less stable, in the sense of having a lower radius, than limit sets in which beliefs are correct.

Lemma 4. Let $\omega \in \Omega^*$. Let $\omega' \in \Omega$ be a recurrent class in which at least one belief is incorrect. Then $r(\omega) > r(\omega')$.

Proof. First, I argue that $r(\omega) \geq (\tau+1)f(\underline{v}/\bar{v})$. (Recall that $\bar{v}$ is the maximum payoff across both players and f is strictly increasing, so that $f(\underline{v}/\bar{v})$ is a lower bound on the cost of a mistake.) Denote the strategies of ω by s_1 and s_2. The beliefs are correct, and so are equal to the strategies. For a change of mood to occur, it must be the case that the m-history is inconsistent with s_1 and s_2. Since the history is fully consistent in every state of ω by assumption, the players must make at least $\tau + 1$ mistakes in total, each of which costs at least $f(\underline{v}/\bar{v})$. Thus $r(\omega) \geq (\tau + 1)f(\underline{v}/\bar{v})$.

Next, I argue that $r(\omega') \leq \tau f(1)$. (Note that $f(1)$ is an upper bound on the cost a mistake.) Again, denote the strategies by s_1 and s_2. Let s_j' be i's belief about j's strategy. At least one of the players has a wrong belief by assumption; suppose that this is true of i, *i.e.* $s_j' \neq s_j$. I claim that it is sufficient for the players to make τ mistakes in total for i to switch to the uncertain mood: pick one of the k-histories for which s_j' and s_j differ, and suppose that in the next k periods i and j make mistakes in such a way as to arrive at that history – this takes $\ell \leq 2k$ mistakes (remember that $\tau > 2k$ so $\ell < \tau$); in the next period, suppose both i and j do not make mistakes, so that j plays an action that conflicts with s_j'; at this stage, the m-history contains at least $\ell + 1$ mistakes from i's point of view; thereafter suppose both players make mistakes each period until the m-history is inconsistent for i, at which point i switches to uncertain. Then a total of $\ell' \leq \tau$ mistakes have been made, each costing less than $f(1)$, so $r(\omega') \leq \tau f(1)$. Moreover, this takes at most $k + 1 + \lceil \tau/2 \rceil$ periods, so $m > 2k + \tau$ is sufficient to ensure that the history is in fact inconsistent for i in the last period.

It then follows from inequality 1 that $r(\omega) > r(\omega')$. □

Thus limit sets in which beliefs are incorrect are less stable than those in which beliefs are correct. Intuitively, this is because a player with an incorrect belief will tend to observe deviations when her opponent isn't making mistakes, which reduces the total number of mistakes required to leave a limit set. Note the role of inequality 1 and the assumptions that $\tau > 2k$ and $m > 2k + \tau$ in the proof. Inequality 1 implies that the number of mistakes is more important in determining the cost of a transition than the payoffs of the players making the mistakes, so that a limit set with high payoffs but incorrect beliefs has a

lower radius than a limit set with low payoffs but correct beliefs. The condition $\tau > 2k$ ensures that it is possible to reach a history in which the incorrect belief diverges from the strategy without adding extra mistakes. And the condition $m > 2k + \tau$ ensures that players recognise mistakes sufficiently far in the past.

Next, I show that any limit set in which players' payoffs do not attain the Kalai-Smorodinsky payoff profile x is unstable, in the sense that for appropriate δ and k there exists another limit set that has a higher radius. The existence of such a limit set relies on theorem A.

Lemma 5. Fix $\lambda \in (0, 1)$ and let $z \in Z$ be a state in a limit set $\omega \in \Omega^*$ in which at least one player i's continuation payoff is weakly less than λx_i. Then there exists $\omega' \in \Omega^*$ such that $r(\omega') > r(\omega)$, provided δ and k are large enough.

Proof. Let i be a player whose continuation payoff is weakly less than λx_i. Consider the following path out of ω: Start in z, and suppose that player i makes a mistake – this costs weakly less than $f(\lambda\rho)$. Then for the next τ periods the player with the minimum cost in that period makes a mistake, at which point the players become uncertain. Note that since x is on the Pareto frontier of V, in each period at least one player's continuation payoff must be x_i or less. Therefore the mistakes cost weakly less than $\tau f(\rho)$ in total. Thus

$$r(\omega) \leq \tau f(\rho) + f(\lambda\rho).$$

Note that this requires $m > k + \tau$.

Next, by theorem A we can pick δ and k large enough such that there exists a subgame-perfect equilibrium s of $\mathscr{H}$ in which both players' continuation payoffs in every subgame are weakly greater than ηx_i, where $\eta \in (0, 1)$ is defined to satisfy

$$(\tau + 1)f(\eta\rho) > \tau f(\rho) + f(\lambda\rho). \tag{2}$$

Let ω' be a limit set in which players' strategies and beliefs are s. Then $r(\omega') \geq (\tau + 1)f(\eta\rho)$, so by inequality 2 we have

$$r(\omega') > r(\omega),$$

as required. □

Roughly speaking, lemma 5 implies that by choosing appropriate k and δ, we can ensure that in every state of the limit set in Ω^* with maximal radius,

players receive payoffs arbitrarily close to the Kalai–Smorodinsky payoff pair. (Note that in lemma 5 differences from x are expressed multiplicatively for ease of exposition, whereas in theorem 1 they are expressed additively.)

The final step of the proof is to show that limit sets with non-maximal radiuses are not stochastically stable. I use the following theorem:

Theorem B (Ellison 2000)**.** Let $\omega, \omega' \in \Omega$. If $r(\omega) > r(\omega', \omega)$, then ω' is not stochastically stable.[18]

Intuitively, Ellison's theorem rules out a given limit set ω' as stochastically stable if there is a second limit set ω that can be easily reached from ω' and that is hard to exit. One might have supposed that $r(\omega, \omega') > r(\omega', \omega)$ would be sufficient to rule out ω', however this is not the case because it might be easy to go from ω to other limit sets and easy to reach ω' from other sets, say. It turns out that $r(\omega) > r(\omega', \omega)$ is sufficient to rule out ω'.

In our case, $r(\omega) > r(\omega', \omega)$ is equivalent to $r(\omega) > r(\omega')$ by lemma 3. So limit sets with non-maximal radiuses are not stochastically stable. Then theorem 1 follows.

4 Conclusion

I have shown how to construct a stochastic learning rule that results in players playing a subgame-perfect equilibrium of the repeated game most of the time in the long run. The learning rule selects the Kalai–Smorodinsky bargaining solution, which highlights the parallels between the canonical bargaining problem and the problem of equilibrium selection in repeated games. The results are achieved under plausible assumptions about the players: they form beliefs based on evidence; they are rational with high probability, in the sense of optimising given their beliefs; and the learning rules are uncoupled, meaning that players do not require knowledge of their opponent's payoffs.

It is worth noting that the selection result depends on exactly how players make mistakes. Under the learning rule, players are more likely to make mistakes the lower their continuation payoff relative to their maximum feasible payoff ($U_i/\bar{v}_i^*$ in the notation above). If instead we had specified the

18 This is theorem 3 in Ellison 2000. One can also use the main theorem, although doing so requires slightly more notation.

mistake probability to depend only on the absolute value of the continuation payoff (that is, just U_i), then the learning rule would have selected the maxmin, or Rawlsian, solution (Rawls 1971; Kalai 1977). Note that the Kalai–Smorodinsky and maxmin solutions both involve maximising some symmetric social welfare function that is increasing in the players' utilities. As a result, they are efficient. That the learning rule selects efficient equilibria is intuitive: if players make mistakes when they expect a low repeated-game utility, equilibria in which players have low repeated-game utility will be unstable, and the process will favour efficient outcomes.

References

Aumann, R. 1976. 'Agreeing to disagree'. *Annals of Statistics* 4: 1236–1239.

Barlo, M., G. Carmona and H. Sabourian. 2016. 'Bounded memory Folk Theorem'. *Journal of Economic Theory* 163: 728–774.

Ellison, G. 2000. 'Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution'. *Review of Economic Studies* 67: 17–45.

Foster, D. P., and H. P. Young. 1990. 'Stochastic evolutionary game dynamics'. *Theoretical Population Biology* 38: 219–232.

———. 2003. 'Learning, hypothesis testing, and Nash equilibrium'. *Games and Economic Behavior* 45: 73–96.

Fudenberg, D., and E. Maskin. 1986. 'The Folk Theorem in repeated games with discounting or with incomplete information'. *Econometrica* 54: 533–554.

Jordan, J. S. 1995. 'Bayesian learning in repeated games'. *Games and Economic Behavior* 9: 8–20.

Kalai, E. 1977. 'Proportional solutions to bargaining situations: interpersonal utility comparisons'. *Econometrica* 45: 1623–1630.

Kalai, E., and E. Lehrer. 1993. 'Rational learning leads to Nash equilibrium'. *Econometrica* 61: 1019–1045.

Kalai, E., and M. Smorodinsky. 1975. 'Other solutions to Nash's bargaining problem'. *Econometrica* 43: 513–518.

Kandori, M., G. J. Mailath and R. Rob. 1993. 'Learning, mutation, and long run equilibria in games'. *Econometrica* 61: 29–56.

Mäs, M., and H. H. Nax. 2016. 'A behavioral study of 'noise' in coordination games'. *Journal of Economic Theory* 162: 195–208.

Motro, U., and A. Shmida. 1995. 'Near-far search: An evolutionarily stable foraging strategy'. *Journal of Theoretical Biology* 173: 15–22.

Nachbar, J. H. 1997. 'Prediction, optimization, and learning in repeated games'. *Econometrica:* 275–309.

———. 2005. 'Beliefs in repeated games'. *Econometrica* 73: 459–480.

Nash, J. 1950. 'The bargaining problem'. *Econometrica* 18: 155–162.

Norman, T. W. 2019. 'Bayesian learning in coordination games'. Working paper.

Nyarko, Y. 1998. 'Bayesian learning and convergence to Nash equilibria without common priors'. *Economic Theory* 11: 643–655.

Pradelski, B. S. R., and H. P. Young. 2012. 'Learning efficient Nash equilibria in distributed systems'. *Games and Economic Behavior* 75: 882–897.

Rawls, J. 1971. *A theory of justice.* Cambridge, MA: Harvard University Press.

Rubinstein, A. 1979. 'Equilibrium in supergames with the overtaking criterion'. *Journal of Economic Theory* 21: 1–9.

Sandroni, A. 1998. 'Necessary and sufficient conditions for convergence to Nash equilibrium: The almost absolute continuity hypothesis'. *Games and Economic Behavior* 22: 121–147.

Young, H. P. 1993. 'The evolution of conventions'. *Econometrica* 61: 57–84.

———. 1998. 'Conventional contracts'. *Review of Economic Studies* 65: 773–792.

———. 2009. 'Learning by trial and error'. *Games and Economic Behavior* 65: 626–643.