Erkenn
DOI 10.1007/s10670-013-9467-y

ORIGINAL ARTICLE

Probe and Adjust in Information Transfer Games

Simon M. Huttegger · Brian Skyrms · Kevin J. S. Zollman

Received: 1 March 2013 / Accepted: 1 March 2013

Abstract We study a low-rationality learning dynamics called *probe and adjust*. Our emphasis is on its properties in games of information transfer such as the Lewis signaling game or the Bala-Goyal network game. These games fall into the class of *weakly better reply games*, in which, starting from any action profile, there is a weakly better reply path to a strict Nash equilibrium. We prove that probe and adjust will be close to strict Nash equilibria in this class of games with arbitrarily high probability. In addition, we compare these asymptotic properties to short-run behavior.

1 Introduction

Recently there has been an increasing interest in explaining high-rationality outcomes in games and decisions by low-rationality means. The central question of this foundational program is whether equilibrium behavior can be attained by simple learning methods. The literature on learning in games considers a large variety of learning methods that make different assumptions about the rationality and the computational capacity of players (for overviews see Fudenberg and Levine 1998; Young 2004). It should be noted that the degree of rationality or cognitive

S. M. Huttegger (✉) · B. Skyrms
Department of Logic and Philosophy of Science, University of California, Social Science Plaza A, Irvine, CA 92697, USA
e-mail: shuttegg@uci.edu

B. Skyrms
e-mail: bskyrms@uci.edu

K. J. S. Zollman
Department of Philosophy, Carnegie Mellon University, Baker Hall 135, Pittsburgh, PA 15213-3890, USA
e-mail: kzollman@andrew.cmu.edu

Published online: 22 March 2013

sophistication of agents can vary greatly even on a low-rationality approach. Two paradigmatic models of learning in games, *reinforcement learning* and *fictitious play*, can be used to illustrate this point. Fictitious play is a very simple form of Bayesian learning; it is assumed that opponent choices form an exchangeable sequence and that the player starts with a Dirichlet prior over the space of probability distributions of per-period opponent play. This leads to a two-parameter family of predictive probabilities for the next move of the opponent. As a decision rule a fictitious play learner uses best response to her beliefs; that is, she chooses the strategy that maximizes her expected payoff, where expectation is taken relative to her probabilities (Brown 1951).[1]

A fictitious play learner has a model of the world—she observes the frequencies of opponent choices—and she is optimizing against that model—she chooses the strategy which is best given her beliefs. Contrariwise, a reinforcement learner simply chooses a strategy with a probability that's proportional to its cumulative past payoff. Thus, fictitious play can be regarded as being higher up the learning hierarchy than reinforcement learning. Reinforcement learners, despite their modest resources, perform well in many decision problems and games (Beggs 2005; Hopkins and Posch 2005). Often the qualitative dynamic behavior resulting from reinforcement learning is very similar to that of fictitious play (Hopkins 2002).[2]

In this paper we study a learning method which is even more minimal than reinforcement learning. This learning method, which is called *probe and adjust* in Skyrms (2010),[3] is payoff-based like reinforcement learning; this means that both learning rules only use information about past payoffs in order to make decisions. But, unlike reinforcement learning, probe and adjust only uses payoff information from the last period of play. In this respect probe and adjust represents one of the simplest possible learning rules in games.[4] It is therefore useful in determining the degree to which outcomes previously justified in terms of very sophisticated reasoning can nonetheless be achieved by very simple agents.

We will focus on games that model strategic aspects of information transfer. After describing probe and adjust learning in the next section, our analysis proceeds in three steps. We apply probe and adjust to a well-known game of information transfer called the *Bala-Goyal network game* (Bala and Goyal 2000). After that, we report some results on probe and adjust in Lewis signaling games. These two games (and several others) are instances of a more general class of games that is studied in Huttegger (2012). In the relevant class of games there are weak improvement paths to strict Nash equilibria. We show that probe and adjust players will play strict Nash

[1] The learning part of fictitious play is equivalent to Carnap's basic system of inductive logic.

[2] The fictitious play process referred to here uses so-called smoothed best responses that result from perturbations of the payoffs.

[3] A somewhat more complicated form of probe and adjust was introduced by Kimbrough and Murphy (2009) in an analysis of tacit collusion in oligopoly pricing. See also Marden et al. (2009) and Young (2009).

[4] This is meant to be informal but intuitively plausible. As one of the reviewers pointed out, this claim could be made more precise by studying the informational and computational requirements of probe and adjust and compare it to other learning rules such as reinforcement learning. This could be done, for instance, by considering finite state automata as in Binmore and Samuelson (1992).

equilibria with arbitrarily high probability in the long run in these games. Moreover, we will point to some further information transfer games that fall under this class of games. We conclude by discussing related literature and pointing out some limitations of our results.

2 Probe and Adjust

Our agents are very simple indeed. They are repeatedly engaged in playing a stage game. Most of the time they use the previous period's action. But sometimes a clock rings for one of them, or a whistle blows for another one, and they experiment by choosing one of their actions at random. Clocks and whistles are independent, so different players can try new actions in the same period. Experimenting with a new action is a probe. If an agent probes, she compares the payoff from choosing the new action to the previous period's payoff. If the new payoff is higher, she continues with playing her new action; if it is lower, she switches back to her old one. This is the "adjust" part of probe and adjust.

More formally, we consider an N-person game Γ in strategic form. Let A_i be the set of player i's actions.[5] Let $A = A_1 \times \cdots \times A_n$ be the set of action profiles. Let $u_i : A \to \mathbb{R}$ be player i's utility function, which yields her payoffs. Let a_{in} be player i's action at time n and a_{-in} the action profile chosen by i's opponents at time n. Then $u_i(a_{in}, a_{-in}) = u_i(a_n)$ is i's payoff at time n, where $a_n \in A$ is the action profile at time n. We can now describe probe and adjust as follows:

Initialization: Random initial action profile; each player i chooses an action a_{i0} from A_i uniformly at random.
Probe: Each player exhibits inertia with high probability and probes with low probability; more specifically, for each n, with probability $1 - \varepsilon, a_{in} = a_{i(n-1)}$; and with probability ε, i chooses a_{in} from A_i uniformly at random.
Adjust: If player i probed, then she adjusts her action $a_{i(n+1)}$ as follows:

$$a_{i(n+1)} = \begin{cases} a_{in} & \text{if } u_i(a_{n-1}) < u_i(a_n) \\ a_{i(n-1)} & \text{if } u_i(a_{n-1}) > u_i(a_n); \end{cases}$$

in case of a tie, we assume that i chooses $a_{i(n+1)} = a_{i(n\text{-}1)}$ with some fixed probability q and $a_{i(n+1)} = a_{in}$ with probability $1 - q$, where $0 < q < 1$.

Notice that we use a uniform probe probability ε for all players. Our results would also hold under the slightly more general assumption of possibly different probe probabilities that are of the same order as they go to zero.[6] Since using different probe probabilities would unnecessarily complicate our arguments, we restrict ourselves to the uniform case. The same is true of the probability for

[5] We use actions instead of strategies since, strictly speaking, we consider infinitely repeated games where a strategy specifies a choice at each play of the stage game. Because we only study payoff-based methods that players use as they play a game repeatedly, what is adjusted are the strategies of the stage game, a.k.a., the actions.

[6] We want to exclude cases where one player probes with probability ε and another with probability ε^2. The proof of Theorem 4 would not apply in this case.

breaking ties. Our results would continue to hold if players were allowed to have different values for q as long as each q is strictly between zero and one.

There are two main differences between probe and adjust and reinforcement learning. Firstly, reinforcement learning uses cumulative payoffs to evaluate an action, while probe and adjust only uses the payoff resulting from the choice of an action when it is probed as compared to the previous round's payoff. And, secondly, a reinforcement learner keeps track of all her actions' cumulative payoffs, which are potentially infinite. Probe and adjust, on the other hand, only has to remember last period's action and payoff and compare it to this period's payoff and strategy.

Probe and adjust can be described as a *better reply dynamics* in cases where the choices of the opponents stay fixed during a probe event. In this case, if a probe event leads to accepting a new strategy it will be at least as good a reply to the profile of opponent choices as the old strategy, and sometimes a better reply. Notice that this only holds if opponent choices are fixed. A probe and adjust event takes two rounds until it is completed, for the player must probe in one round and adjust given her payoff experiences in the next. Nothing in our setup prevents one player from probing while another is still adjusting. For example, suppose that player i probes at time n and player j at time $n + 1$. If i finds that her payoff from probing at n was less than her payoff at $n - 1$, she will switch back to playing her old action at time $n + 1$. j compares her payoff at time $n + 1$ with her payoff at time n. But i might be playing a different strategy at $n + 1$ than she was playing at n. As a consequence, j might not evaluate her probe strategy against a static background of opponent choices and might end up not choosing a better reply.

A learning rule similar to probe and adjust has recently been proposed by Marden et al. (2009).[7] As we will see, there are some differences between this rule and ours, especially with regard to the games we study in this paper. To facilitate a comparison, let us state Marden et al.'s algorithm explicitly. Marden et al.'s learning rule is couched in terms of two additional concepts, that of a baseline action a^b_{in} and that of a baseline utility u^b_{in}. A player's choices are determined by her baseline strategy except for cases where she chooses to experiment. If she experiments, she chooses a new action and adopts it as her new baseline action if and only if it yields a higher payoff than the baseline payoff; in this case she also updates her baseline payoff. Otherwise, she abides by her old baseline action and payoff. More formally, the algorithm is given by the following scheme:

Initialization: Each player i chooses an action a_{i0} from A_i uniformly at random, and sets $a^b_{i0} = a_{i0}$ and $u^b_{i0} = u_i(a_0)$.

Action selection: For each n, with probability $1 - \varepsilon, a_{in} = a^b_{in}$; and with probability ε, i chooses a_{in} from A_i uniformly at random.

Baseline action and baseline utility update: Each player compares $u_i(a_n)$ and u^b_{in} and updates the baseline action as follows:

1. If player i experimented ($a_{in} \neq a^b_{in}$), then

[7] See Young (2009) for an interesting modification of this learning rule.

$$a^b_{i(n+1)} = \begin{cases} a_{in} & \text{if} \quad u^b_{in} < u_i(a_n) \\ a^b_{in} & \text{if} \quad u^b_{in} \geq u_i(a_n); \end{cases}$$

likewise,

$$u^b_{i(n+1)} = \begin{cases} u_i(a_n) & \text{if} \quad u^b_{in} < u_i(a_n) \\ u^b_{in} & \text{if} \quad u^b_{in} \geq u_i(a_n). \end{cases}$$

2. If player i did not experiment, then

$$a^b_{i(n+1)} = a^b_{in}$$
$$u^b_{i(n+1)} = u_i(a_n)$$

There are two main differences between this learning algorithm and probe and adjust. In the first place, Marden et al.'s method uses a specific rule for breaking ties, namely, always stay with your old choice, whereas probe and adjust randomizes; thus, their tie-breaking rule is a limiting case of ours. We will see that this has important consequences. A second difference is that the baseline utility switches back to its previous value if experimentation was not successful. In probe and adjust, the last period's payoff serves as baseline payoff regardless of whether the player just finished probing or not.

Despite their differences, we would like to stress that both probe and adjust and Marden et al.'s learning method represent reasonable candidates for minimal-rationality learning. For both methods, the learner needs no information about the environment and nearly no memory. The only requirement is that she needs to be able to make pairwise payoff comparisons. So, one might think that rules like probe and adjust won't get us very far. In the rest of the paper we will argue to the contrary; probe and adjust seems to be quite successful in games that involve the transfer of information.

3 The Bala-Goyal Game

We now turn to studying some of the properties of probe and adjust in a game where simple trial and error learning rules seem particularly apt. The game was originally conceived by Venkatesh Bala and Sanjeev Goyal (2000), who also serve as its namesakes. This game is a simple network game where players can connect to each other in order to receive information. Each player has a unique piece of information that can be transmitted through any number of players—information can come from far away in the network. The objective of each player is to collect as much information as possible. Since connections to other players are costly, the players face a coordination task to design a network where there are as few connections as possible while maintaining information paths between every player. Since the number of players in an information network can be very large, it is desirable to study whether simple payoff-based dynamics such as probe and adjust lead to

overall efficient network structures or whether the players need to have more information about the structure of the game to do that.

We shall focus on the simplest version of the Bala-Goyal game where there is no information decay as it is transferred through the network, and where information flows only in one direction along connections. Toward the end of the paper we discuss what our results for this game imply for other, more sophisticated versions of the game.

In this simplest version of the Bala-Goyal game there are N players. A player can choose to connect to any of the other players. Thus, an action of player i consists in choosing whether or not to connect to j, for each $j \neq i$. This implies that each player has $\sum_{k=1}^{N-1} \binom{N-1}{k} = 2^{N-1}$ actions. If player i connects to player j, i pays a cost $c \in (0,1)$[8] and receives a piece of information that has a value of 1. In addition, i gets all the pieces of information j gets via her direct and indirect links; that is to say, if there is a chain of connections from i_1 to i_2, and so on, all the way to i_k, then i_1 gets all the information from i_2 to i_k.

An action profile is most easily pictured as a directed graph, where each edge corresponds to a player, and where a directed edge between i and j means that i connects to j. There is a one-to-one correspondence between directed graphs and action profiles. Let us denote the directed graph associated with $a \in A$ as $G(a)$. Then the payoff to player i when a is the action profile chosen by the players is

$$u_i(a) = m_i - k_i c$$

where m_i is the number of players j such that there is a directed path from i to j in $G(a)$; k_i is the number of players j such that there is a directed edge from i to j in $G(a)$.

Efficient information networks have a very simple structure in this version of the Bala-Goyal game. To make this more precise, two important facts have to be noted:

1. An action profile a in the Bala-Goyal game is a Nash equilibrium if, and only if, $G(a)$ is minimally strongly connected.
2. An action profile is a strict Nash equilibrium if, and only if, $G(a)$ is a circle.[9]

The first fact asserts that in a Nash equilibrium there are no unnecessary connections. A graph is minimally connected if (i) it is strongly connected[10] and (ii) removing any one of the edges would result in a disconnected graph. Minimal

[8] In their original presentation of the game, Bala and Goyal allow for arbitrarily high costs which creates essentially three different games with different properties. One where $c \leq 1$, another where $n - 1 \geq c > 1$, and finally one where $c > n - 1$. The case where $c > n - 1$ is not terribly interesting as the cost of obtaining information is so high that it is not worth obtaining it (each player has a dominant strategy). While the case where $n - 1 \geq c > 1$ is very interesting, it has rather different properties from the game where $c \leq 1$ and space prevents its discussion here. We also ignore the case of $c = 1$ since it implies that a player is indifferent between linking and not linking to some other player.

[9] See Bala and Goyal (2000) for proofs.

[10] A directed graph is strongly connected if there is a directed path between each two vertices. Henceforth we will use the term "connected" to mean strongly connected.

connectedness therefore means that in a Nash equilibrium all players get all the information and that costs are not higher than they need to be.

The second fact states that strict Nash equilibria—action profiles that result in a strictly lower payoff for a unilaterally deviating agent—are associated to a special kind of minimally connected graph: the circle. In a circle, there is exactly one directed path that connects all players (up to isomorphism). Thus, each player pays the cost of only one connection while receiving all the information from the other players. Note that there are many action profiles that correspond to a circle configuration.

It is quite easy to see why a graph that is not connected cannot be a Nash equilibrium. If a graph is not connected, then at least one player would benefit from establishing a connection to some other player (since $c \in (0, 1)$ and the value of each piece of information is equal to 1). It is likewise easy to see that a graph that is connected but not minimally connected cannot be associated with a Nash equilibrium. At least one of the players sustains a connection she might as well drop without losing any information. It is also easy to see that the circle has to be a strict Nash equilibrium. If a player deviates from a circle, then she either has more connections than necessary or disrupts the amount of information she receives; in both cases her payoff will be strictly lower. It can also be shown that the circle is the unique Pareto efficient configuration of the Bala-Goyal game (Bala and Goyal 2000).[11]

This analysis amounts to a static characterization of stable action profiles (Nash equilibria) and Pareto efficiency. It does not answer the question whether players will actually choose one of these action profiles. This question can be answered by studying players who repeatedly play the Bala-Goyal game and adjust their choices by learning from experience. Bala and Goyal (2000) themselves study a learning rule called *best response with inertia*. According to this rule, a player usually chooses her action from the previous period of play; but sometimes a clock rings and she best responds to the current action profile of her opponents. In case of a payoff tie, each payoff-equivalent action is chosen with positive probability. It is assumed that only one agent can best respond at a time.

Note the similarity of this learning rule to probe and adjust. Both learning rules exhibit inertia, and both adjust their choice of action at random times. In order to best respond, a player needs to know her own payoff matrix and observe *all* of her opponents' choices. Such significant knowledge of the opponents' actions are not needed by probe and adjust. In many games, this may be too strong an assumption. Even in a moderately large network game, we think that it is too much to ask for.

Best response with inertia can be viewed as a very naive kind of belief learning where a player believes with probability one that her opponents will choose the same action profile as in the previous period. Despite this naivite, Bala and Goyal prove that if all players use best response with inertia, then they will eventually come to choose a circle with probability one.

[11] A strategy profile is Pareto efficient if no player can be made better off by manipulating the choices of any number of players without making some other agent worse off.

Huttegger and Skyrms (2008) investigate two further types of learning rules for the Bala-Goyal game: reinforcement learning and fictitious play. Fictitious play is a more sophisticated kind of belief learning than best response with inertia, and it can be demonstrated in special cases that it will also successfully learn the circle in the Bala-Goyal game. But the same sort of criticisms apply to fictitious play, for it also presupposes knowledge of the global structure of the game. This is not true for reinforcement learning. But Huttegger and Skyrms present analytical and numerical evidence which demonstrates that reinforcement learners will not always converge to playing the circle.

This leaves one with the question whether there is any kind of simple learning rule that leads to the ring in the Bala-Goyal game without using global information about the structure of the game. We provide a solution to this problem by showing that probe and adjust will learn to play the circle in the following sense:

Proposition 1 *In the N-player Bala-Goyal game, for any probability* $p < 1$, *if the probe rate* $\varepsilon > 0$ *is sufficiently small, then the profile of player actions* a_n *is a circle configuration for all sufficiently large n with probability at least p.*

Thus, by making probes infrequent (small ε) players will eventually play a circle configuration with very high probability. Proposition 1 is a quite straightforward consequence of the more general Theorem 4.

Consider a simpler process than probe and adjust where only one player can probe and adjust at a time and no other player can start probing before a probe and adjust event is over (see Huttegger and Skyrms 2012). In this case one can look at an embedded Markov chain that ignores periods of play where nothing changes and focus on the probe and adjust events. This defines a Markov chain on pre-probe and post-probe events. The only absorbing states of this Markov chain are circles. Moreover, there is a positive probability path from every action profile to one of the absorbing states. It then follows from standard results in Markov chain theory that the process will converge to one of the absorbing states with probability one. The analysis of the full probe and adjust dynamics requires more sophisticated techniques since a player can probe while another one is adjusting.

Returning to our more complex probe and adjust process, one can see that once this system is in the optimal state it is likely to remain there, because departing requires a probe (of the right sort) and then an immediately subsequent probe (of the right sort). But finding the optimal state in the first place can be quite difficult. There are $(N-1)!$ cycle configurations in a group of N individuals, but there are a total of $2^{N(N-1)}$ possible network configurations. Thus, as the number of individuals grows the problem of finding the optimal configuration becomes increasingly more difficult.

In the long-run such a state will be reached, but one might also be interested in the probability that such a state is found in the short- to medium-run. Figure 1 shows the proportion of 10,000 iterations that a simulated community of probe-and-adjusters spent in the optimal state.[12] Here we see that finding the optimal state is essentially impossible in a relatively large number of iterations when the groups are

[12] These graphs are for simulations with $c = 0.5$, but the results are characteristic for other costs as well.

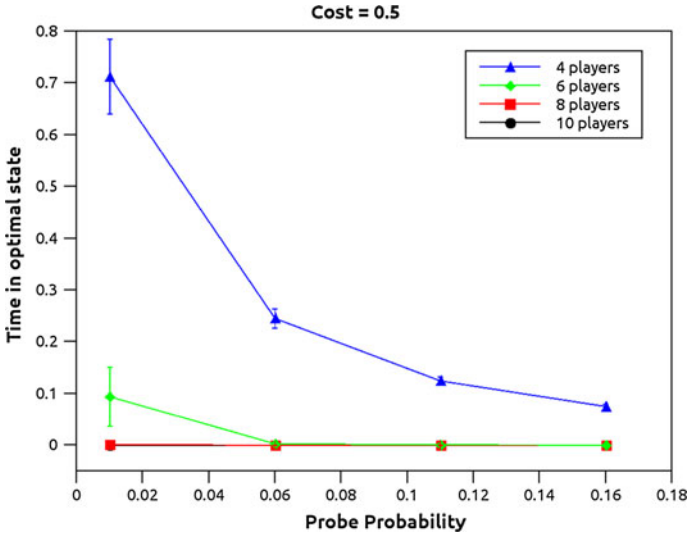

Fig. 1 Simulations results showing the proportion of time spent in the optimal state out of 10,000 iterations of the Bala Goyal network formation game with $c = 0.5$ and $p = 0.5$

even moderately large. We do not think this will be a unique problem with probe and adjust—it is simply a feature of the very large action space.

Perhaps surprisingly very small probe probabilities outperform larger probe probabilities. Higher probe probabilities are a blessing and a curse—they help the community to find the optimal state more quickly (or perhaps at all), but they also increase the probability that there will be subsequent probes that will take the system away from the state. In the Bala-Goyal game it appears that the negatives outweigh the positives.

Time in the optimal state is not the only potential feature of interest, however. Non-optimal states are nonetheless ordered in terms of their superiority, and one might want to know how probe and adjust fares even when it fails to find the optimal state.[13] Figure 2 shows the average payoff of communities of probe-and-adjusters.

When measured by this standard it remains the case that lower probe probabilities outperform higher ones. However, the degree of difference is smaller for games with more players than it is for games with fewer players. This is a result of the normalization—change in payoff from increasing the probe probability is the same absolute amount for games of all sizes. When this is normalized to the maximum payoff of the game, this amount of harm is larger for smaller games where the maximum obtainable payoff is smaller.

[13] The difference between these two measures is obscured by in-the-limit analysis since the limit average payoff will approach the optimal payoff as the probability of optimal play approaches one. In finite time, however, these two measures of performance need not necessarily coincide.

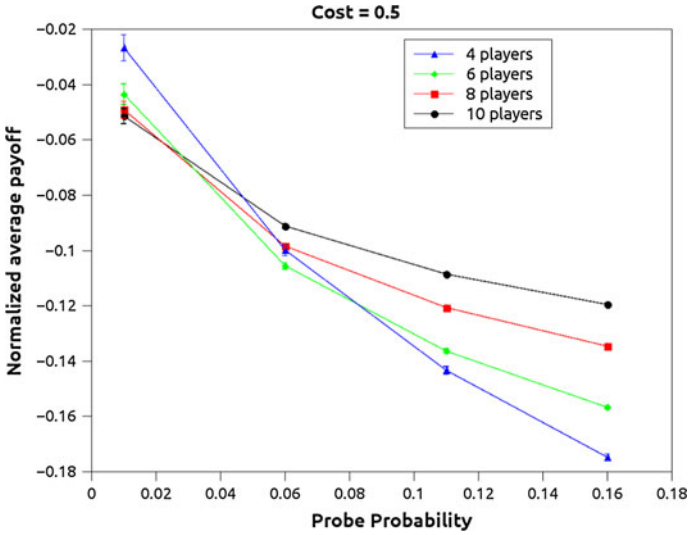

Fig. 2 Simulation results showing the adjusted average payoff for 10,000 iterations of the Bala Goyal network formation game with $c = 0.5$ and $p = 0.5$. The y-axis numbers represent how much worse than optimal the system performs measured in fractions of the optimal payoff

4 Signaling Games

Signaling games capture other aspects of information transfer. They were invented by Lewis (1969) for the purpose of explicating conventional meaning. Since the 1990s, there is a growing body of literature that investigates learning and evolution in this class of games (e.g. Huttegger 2007; Pawlowitsch 2008; Skyrms 1996; Wärneryd 1993). In this section we show that in the limit probe and adjust learning is very successful in signaling games.

A Lewis signaling game is a two-player game between a sender and a receiver. The sender knows which of a finite number of states of the world obtained, but the receiver does not. The sender can send one of a finite number of messages to the receiver. Upon receiving a message, the receiver chooses an act. It is assumed that each act is the right one for exactly one state of the world. More precisely, the interaction between the sender and the receiver is a success if the receiver chooses the right act for the state; otherwise it's a failure. Note that this presupposes complete common interest between the sender and the receiver. In game theoretic terms, the sender receiver game is a *partnership game* since both players get the same payoff in each outcome.

More formally, we are going to consider Lewis signaling games with the same number of states, signals and acts. Let m be the number of states, signals and acts. The sender's strategy specifies what message to send for each of the m states of the world. The receiver's strategy identifies what act to choose for each signal. By putting a probability distribution over the states of the world, one can calculate the expected payoffs of sender and receiver, which will necessarily be equal. Lewis identified specific outcomes of this game, which he named "signaling systems," as

being of particular importance. A signaling system is a strategy profile where by virtue of the signals the sender and the receiver always coordinate states and acts successfully. For this to be the case, the sender strategy is a one-to-one mapping between states and signals and the receiver strategy is a corresponding one-to-one mapping between signals and acts. Signaling systems are the only strict Nash equilibria of these games. There are many other Nash equilibria, though. In particular, there are so-called partial pooling equilibria where some signals are used for more than one state. Partial pooling equilibria can be evolutionarily significant.[14]

Similar to the Bala-Goyal game, we are interested whether there exist simple learning procedures that learn to signal in a Lewis signaling game, i.e., converge to a signaling system. One candidate that has been studied is reinforcement learning. In Argiento et al. (2009) it is shown that if $m = 2$, reinforcement learners converge to playing one of the two signaling systems with probability 1. However, if $m \geq 3$, this is not true anymore; see Hu et al. (2012), who also treat the general case of signaling games with m states and acts and k signals, where k may not be equal to m.

Thus, similar to the situation in the Bala-Goyal game, there is the question of whether there are simple payoff based dynamics that will learn to play signaling systems for all finite m. One such learning method, called win-stay lose-randomize, was shown to learn to signal in Barrett and Zollman (2007). But the definition of this learning rule requires that one can easily define what constitutes a success or a failure in a signaling game. This easy categorization is not possible in most games. Probe and adjust, on the other hand, does not suffer from this drawback. As we assert in the next proposition, probe and adjust will also learn to signal with high probability in the long run. We should be very clear about what the stage game is, however. It is not the extensive form of the Lewis signaling game, but the corresponding strategic form. The payoffs of the strategic form are obtained by fixing a probability distribution over the states of Nature and calculating expected payoffs for the players relative to this probability distribution.[15] The action a probe-and-adjuster chooses is a strategy of the strategic form, and the payoff she receives is an expected payoff. This is a serious idealization that we hope to remove in the future. As the framework stands so far, we need to assume that probe and adjust works on stage games in strategic form.

Proposition 2 *In the Lewis signaling game, for any probability $p < 1$, if the probe rate $\varepsilon > 0$ is sufficiently small, then the profile of player actions a_n is a signaling system for all sufficiently large n with probability at least p.*

As with the corresponding result for the Bala-Goyal game, Proposition 2 is a corollary of the more general Theorem 4. The heuristic can again be understood in terms of a process where only one agent is allowed to probe and adjust at a time (see Skyrms 2012). Signaling systems are the only absorbing states in this process, and there is a positive probability path from each action profile to a signaling system. This implies that the simplified probe and adjust process will converge to a signaling system with probability one. We shall see in the next section that the heuristic is

[14] See Huttegger (2007) and Pawlowitsch (2008) for the technical details.

[15] Proposition 2 does not depend on the probability distribution over the states of nature.

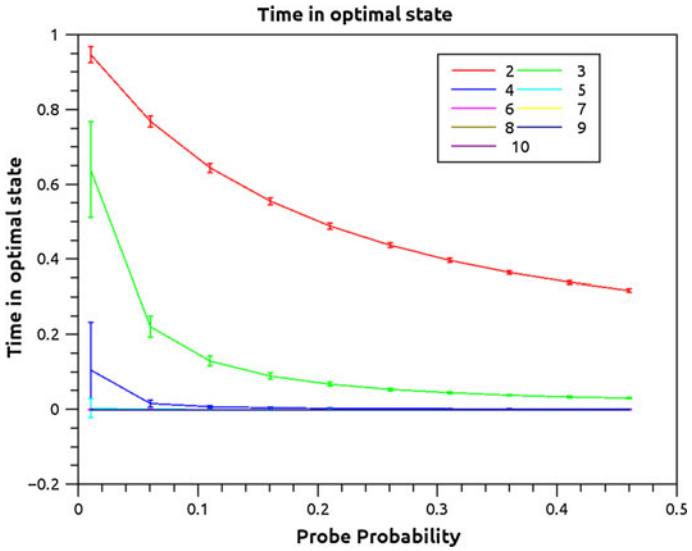

Fig. 3 Simulation results showing the proportion of iterations in optimal state (out of 100,000 iterations) for 9 different Lewis signaling games with equiprobable states and $p = 0.5$

actually useful in understanding the common structure of the Bala-Goyal game and the Lewis signaling game.

Like the Bala-Goyal game, finding a signaling system equilibrium in the short- and medium-run becomes increasingly more difficult as the number of states, signals, and acts multiplies. For a game with m states, signals, and acts, there are $m!$ signaling systems. However, there are m^m possible sender (or receiver) strategies. Again, we find that the proportion of optimal actions goes to zero as the game becomes more complex.

Figure 3 illustrates the proportion of time a community of probe and adjust learners spends in a signaling system. Like the Bala-Goyal game, one finds that the problem becomes intractably hard relatively fast. When $m \geq 6$ none of the tested probe probabilities ever found an optimal state. Also like the Bala-Goyal game, lower probe probabilities outperform higher probe probabilities.

Like the Bala-Goyal game, one might be interested in the performance of probe and adjust beyond merely obtaining optimality. Figure 4 shows the average payoff for several Lewis signaling games. We find similar results as before, as the problem becomes more difficult the average performance drops. In addition, we continue to find that lower probe probabilities are far superior.

5 Weakly Better Reply Games

In this section we briefly describe a certain class of games for which probe and adjust has particularly interesting convergence properties (see Theorem 4 below). As it turns out, the Bala-Goyal network game and the Lewis signaling game are both

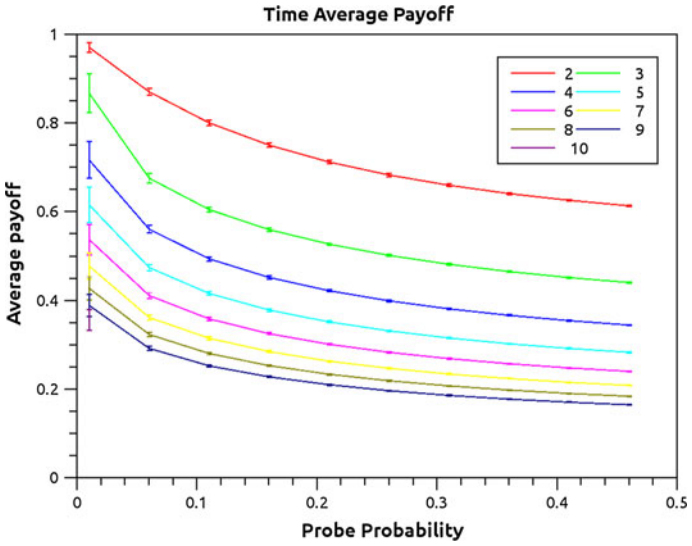

Fig. 4 Simulation results showing average payoff (over 100,000 iterations) for 9 different Lewis signaling games with equiprobable states and $p = 0.5$

instances of this class of games. Let us explain why. Suppose that in a Lewis signaling game players start at a profile that is not a signaling system. Then there is a path of profiles $a^1, \ldots, a^m$ where a^1 is the initial profile and a^m is a signaling system and where the players get from a^k to a^{k+1} by one of them choosing a weakly better reply to the opponents' choices. A weakly better reply is an action that gets you at least as high a payoff against some opponent profile as your current choice does. The same is true for the Bala-Goyal game.

Let us generalize this idea. A *weakly better reply path* is a sequence of action profiles $a^1, \ldots, a^m$ such that for each successive pair a^k, a^{k+1} there is exactly one player i with $a_i^k \neq a_i^{k+1}$ and $u_i(a^{k+1}) \geq u_i(a^k)$; i.e., exactly one player deviates from a profile and does not get a lower payoff by doing so. We will say that a game is a *weakly better reply game* if for any profile a there exists a weakly better reply path starting at a and ending at some strict Nash equilibrium.

We can now state the following result.

Proposition 3 *The Bala-Goyal game and the Lewis signaling game are weakly better reply games.*

Sketch of Proof. For the Bala-Goyal game, it is shown in Bala and Goyal (2000) that there is a positive probability path from any profile a to the circle c for best response with inertia. Bala and Goyal specify a sequence of networks such that at each stage only one player selects a weak best reply to the current action profile and the sequence leads to a circle. For Lewis signaling games, it follows from results in Wärneryd (1993) that at a non-signaling system profile a sender can always choose an alternative map from states to signals that makes use of the unused signals without lowering her payoff (i.e., expected payoff). From there, the receiver can

improve both players' payoffs by responding to the unused signals. This implies that there is a weakly better response path to a signaling system. Cf. also Skyrms (2012) on this property of Lewis signaling games. □

It follows from the proof that, although there are weakly better reply paths to strict Nash equilibria in the Bala-Goyal game and in signaling games, they need not be strict best replies. Thus, if one wants to show that probe and adjust will play strict Nash equilibria of these games in the long run, payoff ties have to be broken randomly. This will be true for many games that involve payoff ties generically (i.e., games with a non-trivial extensive form).

Weakly better reply games are similar to *weakly acyclic games*. In this class of games one requires better reply paths, instead of weakly better reply paths, to lead to a pure Nash equilibrium. On a better reply path, $u_i(a^{k+1}) > u_i(a^k)$ for exactly one player. It is quite straightforward to see that the payoff based learning rule by Marden et al. is designed for good performance in weakly acyclic games, since it stays with a choice exactly when probing does not lead to a higher payoff. Indeed, it is proved in Marden et al. (2009) that by using this algorithm players choose a pure Nash equilibrium with arbitrarily high probability in the long run. Note that this does not mean that their algorithm will choose a *strict* Nash equilibrium. However, it can be shown that this is true for probe and adjust in weakly better reply games.

Theorem 4 *Let* Γ *be an N-player weakly better reply game. Then for any probability* $p < 1$, *if the probe rate* $\varepsilon > 0$ *is sufficiently small, then the profile of player actions* a_n *is a strict Nash equilibrium of* Γ *for all sufficiently large n with probability at least p.*

The proof of this result is sketched in the appendix. It rests on the resistance tree method for determining the stochastically stable states of a game Young (1993, 1998). The intuition behind Theorem 4 is the following. If we suppose that the probe probability ε is sufficiently small, then it is exceedingly unlikely that more than one player will probe at a time. Thus, except on very rare occasions, the process will look very much like the embedded Markov chain that was briefly described in the previous sections. Additionally, the probability for following paths toward a strict Nash equilibrium is much higher than following one away from it.

It is now also easy to see that Theorem 4 implies Propositions 1 and 2 since. According to Proposition 3, both games are weakly better reply games. Moreover, the circle configuration and the signaling systems are the only strict Nash equilibria.

Let us now briefly compare probe and adjust to the learning rule of Marden et al. The latter does not have the same kind of behavior in the Bala-Goyal game or in the Lewis signaling game. In particular, it may get stuck at inefficient Nash equilibria. Suppose that in a Bala-Goyal game there are three players, and that player 1 visits players 2 and 3, while 2 visits 1 and 3 also visits 1. This is a Nash equilibrium profile since the corresponding directed graph is minimally connected. Player 1 does not have an alternative best response to the the other players' action profile, but player 2 might as well visit player 3 and vice versa. Thus probe and adjust can get out of this inefficient Nash equilibrium with one player probing at a time (see Fig. 5). But the Marden et al. rule cannot, since players stay with their baseline strategy in case of a tie. Therefore, the Marden et al. learning method needs at least

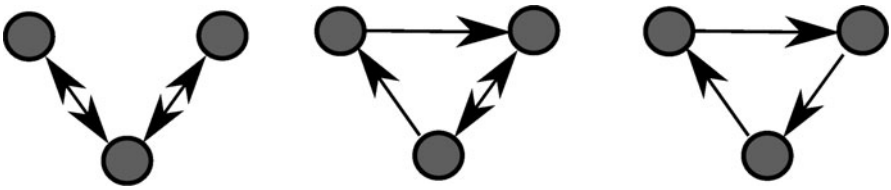

Fig. 5 An illustration of the weakly better response path from a Nash equilibrium to the cycle in the Bala-Goyal game

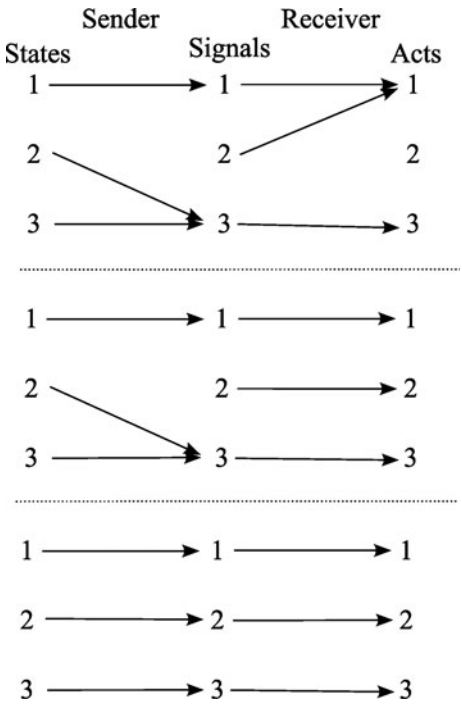

Fig. 6 An illustration of a weakly better response path from a Nash equilibrium to a signaling system in the Lewis signaling game

two simultaneous probes in order to escape the inefficient Nash equilibrium. This can lead to this profile being played with positive probability as the probe probability ε becomes small.

The same is true for the Lewis signaling game, where the Marden et al. rule can get stuck at partial pooling equilibria (see Fig. 6). As in the case of the inefficient equilibrium above, it takes two simultaneous probe events for this learning rule to escape a partial pooling equilibrium. Moreover, it is not clear whether Marden et al.'s rule will converge playing Nash equilibria since the paths leading to Nash equilibria are not better reply paths. In general, one can expect probe and adjust to be better than Marden et al.'s rule in games with payoff ties; for example, in many games that have a non-trivial extensive form.

We would like to emphasize once more that probe and adjust as described in this paper is being applied to games in strategic or normal form. This is different from the approach taken by Skyrms (2012), Argiento et al. (2009) and Hu et al. (2012). The issue whether what is learned are strategies or acts is particularly relevant for extensive form games like the signaling game. If learning is applied to acts, no strategic reasoning on the part of the players is assumed. It remains to be seen to what extent our results of this paper carry over to this more basic setting.

6 Other Information Transfer Games

The results so far suggest that probe and adjust is a fairly successful rule in information transfer games, at least asymptotically. The deeper reason for this is expressed by Theorem 4. We think that many information transfer games have the structure of weakly better reply games because information transfer games are often characterized by a strong degree of common interest between the players, which results in weakly better reply paths leading to strict Nash equilibria. Let us give just two examples. We think that the actual list of relevant examples is much longer.

One could expand the Lewis signaling game in terms of signal chains, where one sender signals to another sender about states of the world; the second signaler signals to a third one about the signals she received. This chain can be extended until a receiver uses the signals in order to determine what act to choose. If there is common interest between all these players, then this game will be a weakly better reply game. Therefore, probe and adjust will converge to playing the signaling systems of a signaling chain in the long run. The same is true for other, more complex signaling games that are based on the Lewis signaling game, where there is more than one sender and more than one receiver; see Skyrms (2010) for more on these variations.

We also conjecture that a combination of Lewis signaling games and the Bala-Goyal network game will be a weakly better reply game. Showing this in detail would take more space than we allow ourselves in this paper. The same is true for the Bala-Goyal game with $n - 1 > c > 1$, a case that was not covered in this paper.

7 Conclusion

In this paper we investigated probe and adjust in the context of information transfer games. We have shown that this minimal-rationality learning rule often converges on playing the efficient outcomes in these games, although there are several caveats as to the speed of convergence. These caveats seem to be related to the size of the strategy space, and thus seem to lead to problems for learning rules in general. Another important open problem concerns probe and adjust for extensive form games.

Our results reinforce the general message that relatively unsophisticated individuals can nonetheless construct complex and efficient social structures or complex and efficient languages. This helps to promote a broader program in game theory that

shows that often the cognitive sophistication traditionally assumed by game theorists is unnecessary in even complex social interactions.

Acknowledgments We would like to thank two anonymous referees for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. EF 1038456 and SES 1026586. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Huttegger was supported by National Science Foundation grant EF 1038456. Zollman was supported by NSF grants EF 1038456 and SES 1026586.

Appendix: Sketch of Proof for Theorem 4

The argument follows the same steps as the proof of Theorem 3.2 in Marden et al. (2009); more details for the case of weakly better reply games are given in Huttegger (2012).

We start with a specification of a state space of the process. Let $a = (a_1, \ldots, a_N)$ be an action profile. Let $a^p = (a_1^p, \ldots, a_N^p)$ be such that $a_i^p = 0$ if player i does not probe and $a_i^p = a_i'$ if player i probes and a_i' was player i's choice of action in the previous period. Let A denote the set of all action profiles a and A^p the set of all action profiles a^p. Then probe and adjust is a Markov chain on $A \times A^p$. $(a, \mathbf{0})$ denotes the state where action profile a is chosen and no one probes. In the *unperturbed* Markov process where $\varepsilon = 0, (a, \mathbf{0})$ is a recurrence class for each a; each state where at least one player probes is transient.

The proof now proceeds by applying the resistance tree method for analyzing stochastic stability as developed in Young (1993). Our goal is to show that, in weakly better reply games, only states $(a, \mathbf{0})$ where a is a strict Nash equilibrium are *stochastically stable*. This means that only these states have positive probability of being observed in the long run as the parameter ε tends to 0. Notice that this yields the assertion of Theorem 4. By a famous result in Young (1993), stochastically stable states of a process are the ones that have *minimum stochastic potential*. This quantity can be specified by determining the *resistances* of going from one state to another. Suppose that $(a, \mathbf{0})$ and $(b, \mathbf{0})$ are two states with $a \neq b$. (By a result in Young (1993) it is enough to consider the resistances between recurrent states of the unperturbed process.) Look at all the ways of getting from $(a, \mathbf{0})$ to $(b, \mathbf{0})$ by having agents probe and adjust. Note, for any intermediate step, how many agents probe. Add the number of all these agents. This is the resistance of the specific path from $(a, \mathbf{0})$ to $(b, \mathbf{0})$. Let r_{ab} be the minimum resistance over all possible paths from a to b.

For any state $(a, \mathbf{0})$, we then construct a *tree rooted at* $(a, \mathbf{0})$. This is a graph with vertices consisting of all states $(b, \mathbf{0})$ such that for any such state there is a unique directed path to $(a, \mathbf{0})$. If there is a directed edge from $(b, \mathbf{0})$ to $(c, \mathbf{0})$ in this graph, we associate the weight r_{bc} with it. The *resistance of the tree* rooted at $(a, \mathbf{0})$ is the sum of all the weights of the edges. The *stochastic potential of* $(a, \mathbf{0})$ is the minimum resistance of all trees rooted at $(a, \mathbf{0})$. Since these states coincide with the stochastically stable states of the probe and adjust process, we have to show that $(a, \mathbf{0})$ has minimum stochastic potential only if a is a strict Nash equilibrium.

There are three basic lemmata. The first is obvious since leaving a state where no player probes requires at least one player to probe.

Lemma 5 *For any state* $(a, \mathbf{0})$, *the resistance of leaving* a *is at least 1.*

The proof of the second lemma is based on the defining property of weakly better reply games.

Lemma 6 *For any state* $(a, \mathbf{0})$ *there is a finite sequence of transitions to a state* $(a^*, \mathbf{0})$ *such that* a^* *is a strict Nash equilibrium and the resistance of each step in the transition is 1.*

The third lemma states that leaving a strict Nash equilibrium state has resistance at least 2. This should be clear from the definition of strict Nash equilibrium; either at least two players have to probe simultaneously or one immediately following the other so that each of them gets at least the same payoff as at the strict Nash equilibrium state.

Lemma 7 *For any state* $(a, \mathbf{0})$ *with* a *a strict Nash equilibrium, any path from* $(a, \mathbf{0})$ *to some other state* $(b, 0)$ *has resistance at least 2.*

The proof can now be completed by showing that whenever a tree T is rooted at a state $(b, \mathbf{0})$ where b is not a strict Nash equilibrium, it can be transformed into a tree T'; rooted at a strict Nash equilibrium state such that the stochastic potential of T'; is strictly less than the stochastic potential of T. By Lemma 6 there exists a path from $(b, \mathbf{0})$ to $(a, \mathbf{0})$ such that a is a strict Nash equilibrium and the resistance at each step on the path is equal to 1. Add the edges corresponding to the path to T and delete the old edges from T on the path. This results in a tree rooted at $(a, \mathbf{0})$. On the path, each old outgoing edge had resistance at least 1 by Lemma 5, except $(a, \mathbf{0})$ which had resistance at least 2 by Lemma 7. Thus the new tree has strictly less stochastic potential.

References

Argiento, R., Pemantle, R., Skyrms, B., & Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic Processes and their Applications, 119,* 373–390.

Bala, V., & Goyal, S. (2000). A noncooperative model of network formation. *Econometrica, 68*, 1129–1181.

Barrett, J., & Zollman, K. J. S. (2007). The role of forgetting in the evolution and learning of language. *Journal of Experimental and Theoretical Artificial Intelligence, 21*, 293–309.

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory, 122*, 1–36.

Binmore, K., & Samuelson, L. (1992). Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory, 57*, 278–305.

Brown, G. W. (1951). Iterative solutions of games by fictitious play. In *Activity analysis of production and allocation* (pp. 374–376). New York: Wiley.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. Cambridge, MA: MIT Press.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica, 70*, 2141–2166.

Hopkins, E., & Posch, M. (2005). Attainability of boundary points under reinforcement learning. *Games and Economic Behavior, 53*, 110–125.

Hu, Y., Skyrms, B., & Tarrés, P. (2012). *Reinforcement learning in signaling games*. Working paper, arXiv:1103.5818.

Huttegger, S.M. (2007). Evolution and the explanation of meaning. *Philosophy of Science, 74*, 1–27.

Huttegger, S. M. (2012). Probe and adjust. *Biological Theory* (forthcoming).
Huttegger, S.M., Skyrms, B. (2008). Emergence of information transfer by inductive learning. *Studia Logica, 89*, 237–256.
Huttegger, S. M., & Skyrms, B. (2012). Emergence of a signaling network with "probe and adjust. In B. Calcott, R. Joyce, K. Sterelny (Eds.), *Signaling, commitment, and emotion*. Cambridge, MA: MIT Press.
Kimbrough, S.O., & Murphy, F. H. (2009). Learning to collude tacitly on production levels by oligopolistic agents. *Computational Economics, 33*, 47–78.
Lewis, D. (1969). *Convention: A philosophical study*. Harvard, MA: Harvard University Press.
Marden, J. R., Young, H. P., Arslan, G., & Shamma, J. S. (2009). Payoff-based dynamics for multiplayer weakly acyclic games. Siam *Journal of Control and Optimization, 48*, 373–396.
Pawlowitsch, C. (2008). Why evolution does not always lead to an optimal signaling system. *Games and Economic Behavior, 63*, 203–226.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.
Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford: Oxford University Press.
Skyrms, B. (2012). Learning to signal with 'probe and adjust'. *Episteme, 9*, 139–150.
Wärneryd, K. (1993). Cheap talk, coordination and evolutionary stability. *Games and Economic Behavior, 5*, 532–546.
Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*, 57–83.
Young, H. P. (1998). *Individual strategy and social structure. An evolutionary theory of institutions*. Princeton, NJ: Princeton University Press.
Young, H. P. (2004). *Strategic learning and its limits*. Oxford: Oxford University Press.
Young, H. P. (2009). Learning by trial and error. *Games and Economic Behavior, 65*, 626–643.