Absolute Expediency and Imitative Behaviour

Antonio J. Morales[1]
centrA
and
University of Malaga

January 2002

[1]This paper is based on some parts of my PhD thesis supervisored by Tilman Börgers at UCL. I really thanks him for his advice and support. Financial support from the Bank of Spain is gratefully acknowledged.

Abstract

This paper analyses a model of learning by imitation, where besides the decision maker, there is a population of individuals facing the same decision problem. We analyse a property called Absolute Expediency, which requires that the decision maker's expected payoff increases from one round to the next for every decision problem and every state of the population. We give a simple characterisation of the expediency property and show that its basic feature is proportional imitation: the change in the probability attached to the played action is proportional to the difference between the received and the sampled payoff (the sampled payoff plays the role of an aspiration level).

1 Introduction

Learning theory has become a central part of economic theory over the last decade. This is due to the fact that it is precisely through a learning process that we arrive to the majority of our decisions and decision making is one of the basic tasks in economics. An extensive overview of this subject is provided by Fudenberg and Levine (1998).

One of the main weaknesses in this literature is the variety of learning rules and that the results obtained are quite often particular to the learning rules. In a recent paper, Börgers et al (2001) have tried to put into order the huge variety of rules considered in the literature by focusing on a particular property: *absolute expediency*.[1] Their approach is so general that encompasses almost all learning rules whose input is the *received* payoff. This property requires that the performance of the agent improves from one period to the next provided the environment stays the same. This property has two particular advantages: first, it refers to agent's behaviour and therefore it can be tested in experiments. Second, the experimental verification is more easily achieved because the property is concerned with the short run performance of the agent, and it is precisely the short run behaviour what it is observed in experiments.

Although they do not provide a complete characterisation of absolute expedient rules, they find necessary as well as some sufficient conditions. Expediency requires that the decision maker behave *as if* he used a modified version of Cross' (1973) learning rule, in which the adjusted probabilities are proportional to the payoff received. However, it encompasses richer rules than the Cross' one as for example, rules which incorporate "similarity" relationships between actions.

In this paper, we extend their theoretical analysis to cover situations in which the agent can observe, in addition to his payoff, the performance of other individuals who are also engaged in the same decision situation. This opens the door to the analysis of learning by imitation which is also believed to be a major source for human learning.

In our imitation scenario, a crucial issue is where the payoffs distributions come from.[2] We will consider the simplest framework in which our analysis

[1] This expression was first considered in the literature on machine learning (Lakshmivarahan and Thathachar (1973)).

[2] In Börgers et al (2001) this question is irrelevant because they deal with individual learning. Therefore, their approach can be applied to single decision problems as well as

can be conducted by assuming that all members of the population are facing the *same* decision problem. It will be assumed that payoff realisations are independent across rounds. In addition, we need to consider how payoff realisations across individuals are related in a *given* round. Among others, there are two extreme cases: (i) payoffs realisations are independent across individuals (this case is referred to as Independent Events Condition), and (ii) every individual choosing the same action receives the same payoff (this case will be named Common Events Condition). These are the cases that will be considered in this paper.

Our analysis will be conducted in what Börgers et al. (2001) call "local" model of learning: Only two periods of time are considered ("today" and "tomorrow") and the decision maker's behaviour "today" is taken as given and fixed. This sort of models can be considered as the "reduced" form of more general learning models. To see how such a model of learning can be derived from a fully specified learning model with a general state space the reader is referred to Section 8 of Börgers et al. (2001).

The decision maker's behavior today is described by a probability distribution, which indicates how likely is that the decision maker plays each of his strategies. After playing his strategy and receiving his payoff, he has the opportunity to observe the payoff and the action taken by a member of a population. With this information, the decision maker updates his probability distribution. To simplify the complexity of the characterisation analysis, we will restrict the class of learning rules that the decision maker can use. The restriction is intended to capture the essence of imitative behaviour: given that imitation is the act of copying others' actions, we will assume that the decision maker can only update the probabilities attached to the played and the sampled actions.

A learning rule is called absolutely expedient if it increases the expected change in expected payoffs from one round to the next for every decision problem and every state of the population. We find that the characterisation of absolutely expedient rules is the same under the Common Events and the Independent Events condition. The basic feature is the proportional imitation, meaning that the change in the probability attached to the played action is proportional to the difference between the received and the sampled payoff. Specifically, the probability attached to the played action is decreased (increased) if its payoff is smaller (greater) than the sampled payoff. Hence,

games.

the sampled payoff plays the role of an aspiration level.

A related proportional imitation component is also found in Schlag (1998) although in a quite different framework. Schlag (1998) considers *pure* strategy imitation rules in a population of agents all of whom are updating their behaviour. He axiomatizes *strictly improving* rules[3] and finds that improving rules imitate higher payoff strategies with a probability which is proportional to the payoff difference.

The rest of the paper is organised as follows. In Section 2, we introduce the framework in which the analysis is conducted. Section 3 contains the definition of absolute expediency. The characterisation results are contained in Section 4. Finally, Section 5 concludes. All the proofs are contained in the Appendix.

2 Framework

A decision maker faces a decision problem. He has to choose one strategy from a finite set S of strategies which has at least two elements. Throughout this paper we keep S fixed. We assume that the decision maker knows S. Let $\mathbb{E}$ be the set of states of Nature. An event e occurs in accordance with a probability distribution p. The payoff received by a decision maker is a function $\pi : S \times \mathbb{E} \to [0,1]$. Note that we have normalized payoffs to be between zero and one.[4] This motivates the following definition:

Definition 1 *An environment E is a specification of a probability distribution p and a payoff function π.*

As we said in the introduction, only two periods of time will be considered: "today" and "tomorrow". The decision maker's behaviour "today" will be exogenous and described by a probability distribution σ. The distribution σ specifies for each pure strategy s how likely it is that the decision maker chooses s today. We make the following assumption.

Assumption 1. *For every $s \in S$ the probability $\sigma(s)$ is strictly positive.*

[3] A rule is strictly improving if when used by all members of the population, the population average expected payoff increases from one round to the next.

[4] The substantial assumption here is that the decision maker knows *some* upper and *some* lower bound for payoffs. Note that this assumption is made everywhere in the literature.

This assumption implies that there is a positive probability that the decision maker plays today the best strategies. This assumption rules out situations in which the decision maker has to find out "good" strategies and focuses on the task of detecting "good" strategies, which is by far easier than the former one.

The decision maker first chooses a strategy s from S and then he observes the payoff received for that choice. After that, he observes the action and the payoff experienced by a member of a population whose members are facing the same decision problem.

The population is described by a probability distribution $\theta \in \Delta(S)$ which specifies the proportion of individuals playing each pure strategy. Assuming uniform sampling, the probability that the decision maker samples an individual playing strategy s is precisely $\theta(s)$. We will refer to θ as the population state.

We will assume that payoff realisations are stochastically independent across today and tomorrow. Regarding how payoff realisations across individuals are related today, two polar cases will be considered.

(a) *Common Events Condition*: The state of Nature is realized. This state of Nature is common to every decision maker.

(b) *Independent Events Condition*: The state of Nature is independently realized across decision makers.

The decision maker's behaviour "tomorrow" is governed by a learning rule.

Definition 2 *A* learning rule L *is a function* $L : S \times [0,1] \times S \times [0,1] \to \Delta(S)$

The distribution $L(s, x, s', y)$ is the state of the decision maker "tomorrow" if his state "today" is σ, the pure strategy which he chose was s, the payoff received was x, the sampled strategy was s' and the sampled payoff was y. For every $s'' \in S$, we denote by $L(s, x, s', y)(s'')$ the probability which $L(s, x, s', y)$ assigns to s''.

In this paper we shall focus on learning rules which satisfy the following two assumptions.

Assumption 2: *Continuity assumption. For any $s \in S$ the learning rule L is continuous in payoffs.*

This is a technical assumption which will allow us to focus on smooth functions in searching for absolutely expedient rules. The key assumption is the following:

Assumption 3: *Imitative assumption.* $L(s, x, s', y)(s'') = \sigma(s'')$ *for all* $s'' \neq s, s'$ *and all* $x, y \in [0, 1]$.

This assumption is intended to capture the essence of imitative behaviour. It states that the probabilities attached to "non-observed" strategies are not updated. The learning rules satisfying this assumption will be called *imitative rules*. Note that more general learning rules will clearly imply a more general and richer structure, at the cost of complicating the analysis.

In the next section, we will introduce the property we are interested in.

3 Expediency

In this section we define the expediency property over the complete set of learning rules, i.e. without assuming the imitative assumption. We will impose the imitative assumption in the next section, where the characterisation analysis will be undertaken.

For given environment E, we define for every strategy $s \in S$: $\pi_s \equiv \int_{\mathbb{E}} \pi(s, e) dp(e)$, i.e. π_s is the expected payoff associated with strategy s. We denote by S^* the set of expected payoff maximising strategies, i.e.: $S^* \equiv \{s \in S \mid \pi_s \geq \pi_{s'} \text{ for all } s' \in S\}$.

For given learning rule L, environment E and population state θ, we can define a function f which assigns to every pure strategy s, the expected change in the probability attached to s.

For the Common Events Condition we have:

$$f(s) = \sum_{s' \in S} \sigma(s') \sum_{s'' \in S} \theta(s'') \int_e \left[L(s', \pi(s', e), s'', \pi(s'', e))(s) - \sigma(s) \right] dp(e)$$

for $s \in S$.

While for the Independent Events Condition $f(s)$ is defined as

$$\sum_{s' \in S} \sigma(s') \sum_{s'' \in S} \theta(s'') \iint_{e', e''} \left[L(s', \pi(s', e'), s'', \pi(s'', e''))(s) - \sigma(s) \right] dp(e') dp(e'')$$

for $s \in S$.

And for any subset $\widetilde{S} \subseteq S$,

$$f(\widetilde{S}) = \sum_{s \in \widetilde{S}} f(s)$$

We also define a function g which gives the expected change in expected payoffs. Formally,

$$g = \sum_{s \in S} f(s)\pi_s$$

Definition 3 *A learning rule B is* expedient *if for all environments E with $S^* \neq S$ and all population states θ: $g > 0$.*

A learning rule is therefore *expedient* if for all non-trivial environments,[5] the decision maker's expected payoff increases for every environment and every population state, provided the environment stays the same. The next section characterises the class of expedient imitative rules.

4 Expedient Imitative Rules

The aim of this paper is the characterisation of the class of absolutely expedient learning rules within the class of imitative rules. Before tackling it, we will state two useful implications of the imitative assumption over the *actual* and *expected* movement of learning rules.

First, consider strategies $s, s' \in S$ with $s \neq s'$. Note that it is enough to specify $L(s, x, s', y)(s)$ to completely specify any *imitative* learning rule because $L(s, x, s', y)(s'') = 0$ for all $s'' \neq s, s'$ by the imitative assumption and $L(s, x, s', y)(s') = \sigma(s) + \sigma(s') - L(s, x, s', y)(s)$ because $L(s, x, s', y)$ is a probability distribution.

Second, there are two circumstances under which the probability attached to a particular action s happens to be updated: (i) either such action s is actually played by the decision maker (this happens with probability $\sigma(s)$) or (ii) such action is sampled by the decision maker (and this happens with probability $\theta(s)$). Note that if the population is not playing such action s, i.e. $\theta(s) = 0$ (and this might occur because there are no restrictions placed on the population state), the unique force is (i). Furthermore, if the unique

[5] An environment is called *trivial* if every strategy gets the same expected payoff.

strategy present in the population is strategy $s' \neq s$, i.e. $\theta(s') = 1$, then the formula for the expected movement is further simplified.

We will follow the approach taken in Börgers et al. (2001) by defining and characterising in first place a seemingly unrelated property.

Definition 4 *A behaviour rule L is* unbiased *if for all environments E with $S^* = S$ and all population states θ: $f(s) = 0$ for all $s \in S$.*

Note that, unlike expediency, there is no much behavioural content underlying this property. It only asks for no expected movement in all trivial environments, precisely the sort of environments for which expediency does not apply. We next show why this property is relevant to our analysis.

Proposition 1 *Every* absolutely expedient *behaviour rule is* unbiased.

The intuition behind this result is clear. If an expedient rule were not unbiased, then it would be the case that for some trivial environment E, there exists at least one action $s \in S$ such that $f(s) < 0$. Increase slightly the payoffs of those strategies s with $f(s) < 0$ to make them the optimal strategies, i.e. the expected payoff maximising ones. By a continuity argument, in the modified environment the optimal strategies will have a negative expected movement in its probability ($f(s) < 0$). It is trivial to show that in that modified environment it is true that $g < 0$, contradicting expediency. Note the role played by the continuity assumption and the fact that the expediency property is defined over a class of environments which contains trivial environments.

Proposition 1 implies that the class of absolutely expedient rules belongs to the class of unbiased rules. Hence, it is clear that our first step towards the characterisation of expedient rules should be the characterisation of unbiased rules. This is the content of the next proposition.

Proposition 2 *An* imitative *rule L is* unbiased *if and only if there exists a function $\mathcal{B} : S \times S \to R$ such that for every $(s, x, s', y) \in S \times [0,1] \times S \times [0,1]$ with $s' \neq s$:*

$$L(s, x, s, y)(s) - \sigma(s) = 0$$

$$L(s, x, s', y)(s) - \sigma(s) = \mathcal{B}(s, s')(x - y)$$

This proposition holds for both the Common and the Independent Events case.

Remark 1 *The first formula comes directly from the imitative assumption.*

Remark 2 *The expected movement of the probability attached to strategy s for an* unbiased *imitative rule is given by*

$$f(s) = \sum_{s' \neq s} (\pi_s - \pi_{s'}) \left[\sigma(s)\theta(s')\mathcal{B}(s, s') + \sigma(s')\theta(s)\mathcal{B}(s', s) \right] \tag{1}$$

This formula holds for both the Common and the Independent Events case.

The remark 2 holds for both cases because the rule is linear in payoffs. Also from the formula in the remark can be seen that for trivial environments it is true that $f(s) = 0$.

To see why the same characterisation arises in both cases, let focus on environments in which all strategies but one receive a deterministic payoff. Note that for this class of environments, the distinction between common and independent events conditions is innocuous. It is precisely this sort of environments whichthat are used in the proof of the proposition.

Then, the main feature of unbiasedness is not only the linearity of the imitative rule in payoffs, but something a bit more demanding: *linearity in payoffs difference*. The linearity in payoffs comes from the fact that the expediency property is defined in terms of expected payoffs, and obviously, expected payoffs are linear in payoffs.[6] The linearity in payoffs difference is due to the fact that we are restricting attention to imitative rules.[7] In a setting in which the analysis is not restricted to imitative learning rules it can be conjectured that the linearity will be the distinguishing feature of this property, although more general rules will be presumably obtained.

Recall that the goal is the characterisation of the expediency property within the class of imitative rules, and note that Propositions 1 and 2 come very close to the desired result. A further step regarding the sign of the coefficients $\mathcal{B}(s, s')$ is therefore called to complete the characterisation. The next definition goes into that direction.

Definition 5 *An* unbiased imitative *rule L is* positive *if $\mathcal{B}(s, s') > 0$, for $s \neq s'$*

[6] The linearity of the expedient rules in payoffs is also obtained by Börgers et al. (2001) in their setting.

[7] The linearity in payoffs difference is also the key feature of the optimal rule in Schlag (1998).

Positivity implies that the probability attached to the played action is decreased as long as the received payoff is smaller than the sampled one; otherwise, the probability is increased. Thus, it requires the unbiased rule to consider the sampled payoff as an aspiration level for the adjustment of the probability distribution, where the update is proportional to the difference between the realised and the observed payoff.

With this definition, we can state the main result of the paper.

Theorem 1 *An* unbiased imitative *rule* L *is* expedient *if and only if it is* positive.

For grasping the intuition behind this theorem, we first interpret the positivity property in terms of the *expected* movement of the rule. Fix an environment and a population state, and perform the following mental experiment: modify the expected payoff associated to a particular action $\widehat{s}$ and look how the expected movement of the rule changes. Equation (1) yields

$$\frac{df(s)}{d\pi_{\widehat{s}}} = -\left[\sigma(s)\theta(\widehat{s})\mathcal{B}\left(s,\widehat{s}\right) + \sigma(\widehat{s})\theta(s)\mathcal{B}\left(\widehat{s},s\right)\right] < 0 \text{ for } s \neq \widehat{s} \tag{2}$$

$$\frac{df\left(\widehat{s}\right)}{d\pi_{\widehat{s}}} = \sum_{s\neq\widehat{s}} \left[\sigma(\widehat{s})\theta(s)\mathcal{B}\left(\widehat{s},s\right) + \sigma(s)\theta(\widehat{s})\mathcal{B}\left(s,\widehat{s}\right)\right] > 0 \tag{3}$$

Thus positivity implies that the expected change in the probability attached to strategy $\widehat{s}$ increases when its expected payoff is increased while the probability attached to any other strategy decreases. Hence, the rule goes in expected terms into the direction of putting more weight to the action whose expected payoff has been increased, taking away some probability from all other strategies.

With this interpretation at hand, we can understand why expediency requires positivity. Consider an expedient rule which is not positive, i.e. there exist strategies $\widehat{s}$ and s' such that $\mathcal{B}\left(\widehat{s},s\right) \leq 0$. Consider a trivial environment. As the rule is unbiased (Proposition 1), we already know that $f(s) = 0$ for all strategy $s \in S$. Increase slightly the expected payoff of action $\widehat{s}$ to make it the unique expected payoff maximising action. Note that in this new environment there are two different expected payoffs values. Focus on a collapsed population with $\theta\left(s'\right) = 1$. Then $df\left(\widehat{s}\right)/d\pi_{\widehat{s}} = \sigma(\widehat{s})\mathcal{B}\left(\widehat{s},s'\right) \leq 0$. But this clearly contradict expediency as the rule in expected terms is

taking away probability from the unique optimal strategy and increasing the probability of the sub-optimal strategies.

For the sufficiency result, consider an unbiased and positive rule. Consider a bandit problem with two different expected payoffs values. By simple inspection of equation (1), it is clear that positivity implies that the expected change in expected payoffs is positive, i.e. $g > 0$. This will be our starting point. Let $\widehat{s}$ denote a strategy belonging to the set of worst strategies. Consider the following mental experiment: modify the expected payoff associated to strategy $\widehat{s}$ and look how the expected change in expected payoffs changes. With some abuse in the notation, we can represent this as follows:

$$\frac{dg}{d\pi_{\widehat{s}}} = f(\widehat{s}) + \sum_{s \in S} \pi_s \frac{df(s)}{d\pi_{\widehat{s}}}$$

Note that positivity implies that $f(\widehat{s}) < 0$, $df(s)/d\pi_{\widehat{s}} < 0$, while $df(\widehat{s})/d\pi_{\widehat{s}} > 0$. Note however, that $\pi_{\widehat{s}}$ is the smaller expected payoff present in the bandit problem. Therefore, for any $s \neq \widehat{s}$ it is true that $\pi_s \geq \pi_{\widehat{s}}$, where the inequality is strict for some s. Hence, we can rewrite π_s as $\pi_{\widehat{s}} + \varepsilon(s)$, where $\varepsilon(s) \geq 0$ with strict inequality for some s. Then we can write the above expression as

$$\frac{dg}{d\pi_{\widehat{s}}} = f(\widehat{s}) + \pi_{\widehat{s}} \sum_{s \in S} \frac{df(s)}{d\pi_{\widehat{s}}} + \sum_{s \neq \widehat{s}} \varepsilon(s) \frac{df(s)}{d\pi_{\widehat{s}}}$$

Note that the second term is zero by definition and the third term is negative by positivity. Recalling that $f(\widehat{s})$ is negative, we find that $dg/d\pi_{\widehat{s}} < 0$.

Hence, we have found that when lowering the payoff attached to any of the worst strategies, the expected change in expected payoffs must increase. This is enough to show that unbiasedness and positivity imply expediency as we started out in an environment with $g > 0$. A more formal and elaborated proof based on this idea is developed in the Appendix.

Summarising, Theorem 1 shows that *positivity* is the basic ingredient for the *expediency* property within the class of imitative rules. We can now assess how restrictive the imitative assumption we have introduced is by comparing our results to those in Börgers et al (2001). They characterise two properties: absolutely expediency and monotonicity. A learning rule is monotone if the expected change in the probability attached to the optimal strategies is positive. They show that both monotonicity and absolutely expediency

require the positivity property.[8] They also define another property called "cross-negativity": An unbiased learning rule is cross-negative if the probability of playing tomorrow a strategy different from the one played today is non-increasing in the payoff received today. They show that an unbiased learning rule is monotone if and only if it is positive and cross-negative and that cross-negativity is sufficient for expediency although is not necessary. Noting that in our setting the positivity property and the imitative assumption imply cross-negativity, their results show that within the class of cross-negative rules, monotonicity and absolutely expediency are equivalent properties (their propositions 3 and 4). This is what happens in our paper and this is why we are able to characterise all absolutely expedient rules in our setting.

5 Conclusion

In this paper we have developed a model of learning by imitation. We have characterised a property called *expediency*. A learning rule is expedient if increases the decision maker expected payoff from one period to the next, regardless of the decision problem and the population state. The characterisation is performed within the class of *imitative* rules: In this class, the decision maker is "denied" the possibility of updating the probabilities attached to non-observed strategies.

The basic component of any expedient imitative rule is that the change in the decision maker's state is proportional to the payoffs difference between the received and the sampled payoff. A related proportional imitation component is also found in Schlag (1998) although in a quite different framework. Schlag (1998) considers *pure* strategy imitation rules in an evolving population whereas our characterization deals with *mixed* strategy imitation rules concerning one single decision maker. In addition, Schlag axiomatizes *strictly improving* rules, a property concerned with the evolution of the whole population, whereas in our setting we focus on a property concerned with the behaviour of a single individual. Improving rules imitate higher payoff strategies with a probability which is proportional to the payoff difference. In our setting absolutely expediency implies that the change in the decision maker's state is proportional to the payoff difference, although it incorporates a reinforcement component, the sampled payoff being an aspiration level, i.e.

[8] They call it "own-positivity".

the probability attached to the played strategy is increased if it gets a higher payoff than the sampled action, otherwise is decreased.

Our paper opens interesting lines of research in imitative scenarios. On the theoretical side, it might be interesting to explore the characterisation of expedient rules without imposing the imitative assumption. On the practical side, it suggests the investigation of whether experimental subjects responde in a proportional way to the payoffs experience in the way suggested by our theoretical analysis.

6 Appendix

Proof of Proposition 1. The proof is indirect. Suppose there were an environment E with $S^* = S$, a population state θ and a strategy $s \in S$ such that $f(s) \neq 0$. Let π denote the expected payoff associated with E. Let $S^+(E) = \{s \in S : f(s) \geq 0\}$ and $S^-(E) = \{s \in S : f(s) < 0\}$. Note that $S^-(E) \neq \emptyset$. Let $\tilde{E}$ denote the environment in which the payoff associated to strategies belonging to $S^-(E)$ is increased by adding $\varepsilon > 0$. Note that in this new environment, $S^* = S^-(E)$. Denote by $\tilde{g}$ the expected change in payoffs in $\tilde{E}$. Then

$$\begin{aligned}
\tilde{g} &= \sum_{s \in S^-(E)} \widetilde{f}(s)(\pi + \varepsilon) + \sum_{s \notin S^-(E)} \widetilde{f}(s)\pi \\
&= \sum_{s \notin S^-(E)} \widetilde{f}(s)\pi + \sum_{s \in S^-(E)} \widetilde{f}(s)\pi + \sum_{s \in S^-(E)} \widetilde{f}(s)\varepsilon \\
&= \pi \sum_{s \in S} \widetilde{f}(s) + \varepsilon \sum_{s \in S^-(E)} \widetilde{f}(s)
\end{aligned}$$

Note that $\sum_{s \in S} \widetilde{f}(s) = 0$. Furthermore, by continuity $\widetilde{f}(s) < 0$ for all $s \in S^-(E)$, which implies that $\tilde{g} < 0$, contradicting expediency. ■

Proof of Proposition 2.

Sufficiency: Equation (1) gives the following formula for the expected movement of the probability attached to strategy s for both the common and independent events case.

$$f(s) = \sum_{s' \neq s} (\pi_s - \pi_{s'}) \left[\sigma(s)\theta(s')\mathcal{B}(s, s') + \sigma(s')\theta(s)\mathcal{B}(s', s)\right]$$

If $\pi^s = \pi^{s'}$ for all $s' \neq s$, this expression is null, as the definition of unbiasedness required.

Necessity: In the remainder of the proof, we consider some given unbiased imitative rule and show that the imitation rule has to have the property stated in proposition 2. We proceed in two steps.

Step 1: We first show that there exists a function $\widetilde{\mathcal{B}} : S \times S \grave{\times} [0,1] \to R$ such that for all $s' \neq s$,

$$L(s,x,s',y)(s) - \sigma(s) = (x-y)\,\widetilde{\mathcal{B}}(s,s',y)$$

Let $a, y, c \in [0,1]$ with $a < y < c$. Let $\mathbb{E}$ be $\{e_1, e_2\}$. Consider an environment $\widehat{E}$ with $p(e_1) = \widehat{p}$ and $p(e_2) = 1 - \widehat{p}$. For action $\widehat{s}$ we have $\pi(\widehat{s}, e_1) = a$, $\pi(\widehat{s}, e_2) = c$ whereas for any other strategy $s \neq \widehat{s}$, $\pi(s, e_1) = \pi(s, e_2) = y$. Let $\widehat{p} = \frac{c-y}{c-a}$. It is clear that for this environment $S = S^*$ and therefore $f(s) = 0$ for all $s \in S$. Consider a collapsed population with $\theta(s') = 1$ for some $s' \neq \widehat{s}$. Then the formula for the expected change in the probability attached to strategy $\widehat{s}$ under the common events case is as follows:

$$\sigma(\widehat{s})\left[\widehat{p}L(\widehat{s},a,s',y)(\widehat{s}) + (1-\widehat{p})L(\widehat{s},c,s',y)(\widehat{s}) - \sigma(\widehat{s})\right] = 0 \tag{4}$$

For the independent events condition we have the following expression for $f(\widehat{s})$

$$\sigma(\widehat{s})\left[\begin{array}{c}\widehat{p}\widehat{p}L(\widehat{s},a,s',y)(\widehat{s}) + \widehat{p}(1-\widehat{p})L(\widehat{s},a,s',y)(\widehat{s}) + \\ (1-\widehat{p})\widehat{p}L(\widehat{s},c,s',y)(\widehat{s}) + (1-\widehat{p})(1-\widehat{p})L(\widehat{s},c,s',y)(\widehat{s}) - \sigma(\widehat{s})\end{array}\right] = 0 \tag{5}$$

Note that the second expression reduces to the first one. Therefore, both expression are the same.

Consider an alternative environment in which for all strategy s we have $\pi(s,e_1) = \pi(s,e_2) = y$. Then for this environment and the same collapsed population as before we have

$$f(\widehat{s}) = \sigma(\widehat{s})\left[L(\widehat{s},y,s',y)(\widehat{s}) - \sigma(\widehat{s})\right] = 0 \tag{6}$$

Therefore, the conclusions we reach by using equations (4) and (6) will apply to both the common and the independent events conditions.

Both equations imply

$$\widehat{p}L(\widehat{s},a,s',y)(\widehat{s}) + (1-\widehat{p})L(\widehat{s},c,s',y)(\widehat{s}) = L(\widehat{s},y,s',y)(\widehat{s})$$

Replacing $\widehat{p}$ by $\frac{c-y}{c-a}$ and rearranging,

$$\frac{L\left(\widehat{s},c,s',y\right)\left(\widehat{s}\right)-L\left(\widehat{s},b,s',y\right)\left(\widehat{s}\right)}{L\left(\widehat{s},a,s',y\right)\left(\widehat{s}\right)-L\left(\widehat{s},b,s',y\right)\left(\widehat{s}\right)}=\frac{c-y}{a-y}$$

Recalling that from equation (6) we have that $L\left(\widehat{s},y,s',y\right)\left(\widehat{s}\right)=\sigma(\widehat{s})$, it follows that

$$\frac{L\left(\widehat{s},c,s',y\right)\left(\widehat{s}\right)-\sigma(\widehat{s})}{L\left(\widehat{s},a,s',y\right)\left(\widehat{s}\right)-\sigma(\widehat{s})}=\frac{c-y}{a-y}$$

At this must be true for all a,y,c with $a<y<c$, $L\left(\widehat{s},x,s',y\right)\left(\widehat{s}\right)-\sigma(\widehat{s})$ must be of the form $(x-y)\,\widetilde{\mathcal{B}}\left(\widehat{s},s',y\right)$, as asserted.

Step 2: We next show that in fact, $\widetilde{\mathcal{B}}\left(\widehat{s},s',y\right)$ is independent of the sampled payoff y, i.e. that for all $s'\neq s$, $\widetilde{\mathcal{B}}\left(\widehat{s},s',y_1\right)=\widetilde{\mathcal{B}}\left(\widehat{s},s',y_2\right)$ for $y_1\neq y_2$.

Let $a,b\in[0,1]$ with $a\neq b$. Let $\mathbb{E}$ be $\{e_1,e_2\}$. Consider an environment E with $p\left(e_1\right)=p\left(e_2\right)=1/2$. For action $\widehat{s}$ we have $\pi\left(\widehat{s},e_1\right)=a$, $\pi\left(\widehat{s},e_2\right)=b$, for action s' we have $\pi\left(\widehat{s},e_1\right)=b$, $\pi\left(\widehat{s},e_2\right)=a$ whereas for any other strategy $s\neq\widehat{s},s'$, $\pi\left(s,e_1\right)=\pi\left(s,e_2\right)=(a+b)/2$. It is clear that for this environment $S=S^*$ and therefore $f(s)=0$ for all $s\in S$.

Consider a population state θ with $\theta\left(s'\right)=1$. We will now compute the formula for the expected change in the probability attached to action $\widehat{s}$.

For the common events condition we have

$$\sigma(\widehat{s})\left[\frac{1}{2}(a-b)\widetilde{\mathcal{B}}\left(\widehat{s},s',b\right)+\frac{1}{2}(b-a)\widetilde{\mathcal{B}}\left(\widehat{s},s',a\right)\right]=0$$

Rearranging we obtain

$$\sigma(\widehat{s})\frac{1}{2}(a-b)\left[\widetilde{\mathcal{B}}\left(\widehat{s},s',b\right)-\widetilde{\mathcal{B}}\left(\widehat{s},s',a\right)\right]=0$$

which implies $\widetilde{\mathcal{B}}\left(\widehat{s},s',b\right)=\widetilde{\mathcal{B}}\left(\widehat{s},s',a\right)$.

For the independent events condition we have

$$\sigma(\widehat{s})\frac{1}{4}\left[\begin{array}{c}(a-b)\,\widetilde{\mathcal{B}}\left(\widehat{s},s',b\right)+\\ L\left(\widehat{s},a,s',a\right)-\sigma(\widehat{s})+\\ L\left(\widehat{s},b,s',b\right)-\sigma(\widehat{s})+\\ (b-a)\,\widetilde{\mathcal{B}}\left(\widehat{s},s',a\right)\end{array}\right]=0 \tag{7}$$

Recalling that from equation (6) it is true $L\left(\widehat{s},y,s',y\right)=\sigma(\widehat{s})$, the above equation can be rewritten as follows

$$\sigma(\widehat{s})\frac{1}{4}(a-b)\left[\widetilde{\mathcal{B}}\left(\widehat{s},s',b\right)-\widetilde{\mathcal{B}}\left(\widehat{s},s',a\right)\right]=0 \tag{8}$$

which again implies $\widetilde{\mathcal{B}}(\widehat{s}, s', b) = \widetilde{\mathcal{B}}(\widehat{s}, s', a)$. As this must be true for all a, b, the proof is complete. ■

Proof of Theorem 1.

Sufficiency: The proof is indirect. Suppose there were $s, s' \in S$ with $s' \neq s$ such that $\mathcal{B}(s, s') < 0$. Let $\varepsilon, b > 0, b + \varepsilon < 1$. Let $\mathbb{E}$ be $\{e\}$. Consider an environment E with $p(e) = 1$. For action s we have $\pi(s, e) = b + \varepsilon$ whereas for any other strategy s', $\pi(s', e) = b$. Note that $S^* = \{s\}$.

The expected change in expected payoffs for this environment is given by

$$\begin{aligned} g &= \sum_{s' \in S} f(s')\pi_{s'} \\ &= f(s)\varepsilon + b\sum_{s' \in S} f(s') \\ &= f(s)\varepsilon \end{aligned}$$

Consider a population state θ with $\theta(s') = 1$. The unique strategy which the decision maker will ever sample is strategy s', and only through playing a particular action $s \neq s'$ is that the decision maker gets to update the probability attached to that particular action s. Then the expected change in the probability attached to the expected payoff maximising action is given by

$$f(s) = \sigma(s)\mathcal{B}(s, s')\varepsilon$$

which implies

$$g = \sigma(s)\mathcal{B}(s, s')\varepsilon^2$$

Therefore, $\mathcal{B}(s, s') < 0$ implies $g < 0$ contradicting expediency.

Necessity: The proof is indirect. Let L be an unbiased and positive imitative rule such that there exists some non-trivial environment E and some population state θ such that

$$g = \sum_{s \in S} f(s)\pi_s < 0 \tag{9}$$

Let $n(E)$ denote the number of different expected payoff values in environment E. If $n(E) = 2$, then the proof is trivial because it is trivial to show that positivity implies that $g > 0$. Then, we need to study the case in which there are more than 2 different expected payoffs values.

Let $\overline{S}$ denote the set actions with the smallest expected payoff in this environment. Let $\overline{\pi}$ denote such payoff. Let $\overline{\overline{S}}$ denote the set of actions with the second smallest expected payoff, and let $\overline{\overline{\pi}}$ denote such payoff. Finally, let k denote the difference between these two values, i.e. $k = \overline{\overline{\pi}} - \overline{\pi}$. Note that $k > 0$ by definition.

The expected change in the probability attached to action is given by equation (1). We write it again:

$$f(s) = \sum_{s' \neq s} (\pi_s - \pi_{s'}) A(\theta, s, s') \tag{10}$$

where $A(\theta, s, s')$ stands for $\sigma(s)\theta(s')\mathcal{B}(s, s') + \sigma(s')\theta(s)\mathcal{B}(s', s)$. Recall that positivity implies that $A(\theta, s, s') > 0$, and therefore $f(\overline{S}) < 0$.

Consider a modified environment $\widetilde{E}$ in which the expected payoff associated to strategies belonging to $\overline{S}$ is raised by an amount k. We will show that for this new environment, $\widetilde{g} < \overline{g}$, and this will be enough to prove the claim. Let $\widetilde{f}(s)$ denote the expected change in the probability attached to action s in decision problem $\widetilde{E}$. For those strategies s not belonging to the set $\overline{S}$, we have

$$\widetilde{f}(s) = f(s) - \sum_{s' \in \overline{S}} A(\theta, s, s')k \tag{11}$$

we have collected enough information to focus now on $\widetilde{g}$.

$$\widetilde{g} = \sum_{s \notin \overline{S}} \widetilde{f}(s)\pi_s + \sum_{s' \in \overline{S}} \widetilde{f}(s)(\overline{\pi} + k)$$

This can be written as

$$\widetilde{g} = \sum_{s \notin \overline{S}} \widetilde{f}(s)(\pi_s - \overline{\pi} - k)$$

where we have used that $\sum_s \widetilde{f}(s) = 0$. Rearranging

$$\begin{aligned} \widetilde{g} &= -k \sum_{s \notin \overline{S}} \widetilde{f}(s) \\ &\quad + \sum_{s \notin \overline{S}} \widetilde{f}(s)(\pi_s - \overline{\pi}) \end{aligned}$$

By using expression (11) and rearranging,

$$\begin{aligned}\widetilde{g} &= -k\sum_{s\notin\overline{S}}\widetilde{f}(s)-\sum_{s\notin\overline{S}}\sum_{s'\in\overline{S}}A(\theta,s,s')k(\pi_s-\overline{\pi})\\ &\quad+\sum_{s\notin\overline{S}}f(s)\pi_s-\overline{\pi}\sum_{s\notin\overline{S}}f(s)\end{aligned}$$

By using that $\sum_s f(s)=0$ the second line can be rewritten

$$\begin{aligned}\widetilde{g} &= -k\sum_{s\notin\overline{S}}\widetilde{f}(s)-\sum_{s\notin\overline{S}}\sum_{s'\in\overline{S}}A(\theta,s,s')k(\pi_s-\overline{\pi})\\ &\quad+\sum_{s\notin\overline{S}}f(s)\pi_s+\overline{\pi}\sum_{s\in\overline{S}}f(s)\end{aligned}$$

But note that the third line is simply g. Therefore we arrive at the following expression

$$\begin{aligned}\widetilde{g} &= g\\ &\quad-k\sum_{s\notin\overline{S}}\widetilde{f}(s)\\ &\quad-\sum_{s\notin\overline{S}}\sum_{s'\in\overline{S}}A(\theta,s,s')k(\pi_s-\overline{\pi})\end{aligned}$$

Note that the second line is negative because k is positive and $\sum_{s\notin\overline{S}}\widetilde{f}(s)$ is positive by positivity. Note that the third line is negative because k is positive, $A(\theta,s,s')$ is positive by positivity and $\pi_s-\overline{\pi}$ is positive by definition of $\overline{\pi}$. Therefore, we have proved that for environment $\widetilde{E}$, $\widetilde{g}<g<0$. What this implies for the proof? To answer this question, think of how $\widetilde{E}$ relates to E in terms of the number of different expected payoff values. In fact, $n(\widetilde{E})=n(E)-1$: This is the key. Starting at environment E with $g<0$ we can find a different environment $\widetilde{E}$ for which $\widetilde{g}<0$ but with a smaller number of different expected payoff values. By repeating the process we will eventually find an environment $\widehat{E}$ with $n(\widehat{E})=2$ such that for the imitative rule L it is true that $\widehat{g}<0$. And this is the desired contradiction. ■

References

[1] Börgers, T., Morales, A. J. and R. Sarin , Expedient and Monotone Learning Rules, (2001), mimeo.

[2] Cross, J., A Stochastic Learning Model of Economic Behavior, *Quarterly Journal of Economics* 87 (1973), 239-266.

[3] Fudenberg, D. and D. Levine, *The Theory of Learning in Games*, Cambridge and London: MIT Press, 1998.

[4] Lakshmivarahan, S. and M. Thathachar, Absolutely Expedient Algorithms for Stochastic Automata, *IEEE Transactions on Systems, Man and Cybernetics* 3 (1973), 281-286.

[5] Schlag, K., Why Imitate, and if so, How? A Boundedly Rational Approach to Multi-Armed Bandits, *Journal of Economic Theory* 78 (1998), 130-156.