Projektbereich B6
Discussion Paper No. B-361

Why Imitate, and if so, How? A Bounded Rational Approach to Multi-Armed Bandits[1]

Karl H. Schlag[2]

Febuary, 1996

---

[1] The author wishes to thank Dirk Bergemann, Jonas Björnerstedt, Georg Nöldeke, Larry Samuelson, Avner Shaked, a referee and an associate editor for helpful comments. Financial support from the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 at the University of Bonn is gratefully acknowledged.

[2] Abt. Wirtschaftstheorie III, Department of Economics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany.

Abstract

We consider the situation in which individuals in a finite population must repeatedly choose an action yielding an uncertain payoff. Between choices, each individual may observe the performance of one other individual. We search for rules of behavior with limited memory that increase expected payoffs for any underlying payoff distribution. It is shown that the rule that outperforms all other rules with this property is the one that specifies imitation of the action of an individual that performed better with a probability proportional to how much better she performed. When each individual uses this best rule, the aggregate population behavior can be approximated by the replicator dynamic.

JEL Classification Number: C72, C79

Keywords: social learning, bounded rationality, imitation, multi-armed bandit, random matching, payoff increasing, replicator dynamic.

1 Introduction

Imitation is the act of copying or mimicking the action of others, a commonly observed behavior in human decision making. This is perhaps the reason why lately, with the growing popularity of Evolutionary Game Theory, many models of social learning have emerged in which individuals select their future actions by imitating others (e.g., Banerjee [1]; Björnerstedt and Weibull [3]; Cabrales [5]; Ellison and Fudenberg [7]; Gale et al. [8]; Hofbauer [9]; Rogers [12]). We propose to answer why individuals should follow an imitative behavior in a social learning environment. Especially, we want to point out that the answer is directly related to the answer to how imitation should take place. Our approach follows the traditional line of decision and game theory in that we allow individuals to determine their behavior before we derive implications of individual behavior on the evolution of the society. At the same time we depart from traditional decision theory (Savage [16]) and suggest two alternative approaches to selecting individual behavior, a bounded rational approach and a population-oriented approach. What is interesting is that either approach leads to the same unique prescription of how to choose future actions:

- follow an imitative behavior, i.e., only change actions through imitating others,
- never imitate an individual that performed worse than oneself, and
- imitate an individual that performed better with a probability that is proportional to how much better this individual performed.

This result is not only interesting for its own sake but also for its implications on the global adjustment of a large population in which each individual follows the individually most preferred rule. In this regard, it turns out that the global process governing the change of actions in the population can be approximated by the well known replicator dynamic (Taylor [20]). In this dynamic, the growth rate of an action is equal to its relative payoff measured with respect to the average payoff in the population.

This paper generalizes an earlier work by the author (Schlag [17]) to a multi-armed bandit setting. Individuals must repeatedly choose an action from a finite set of actions A. Choosing an action yields an uncertain payoff. Payoffs are realized independently, their distribution has finite support and belongs to a bounded interval $[\alpha, \omega]$. This basic setting of multi-armed bandits has a wide range of applications to economics and behavioral sciences; choosing an action can be a synonym for the choice of a technology, of an organizational structure or of a restaurant, it may be the setting of a price (Ellison and Fudenberg [7]; Schmalensee [19]). In our model, individuals belong to a finite population and are able to learn from others. After each payoff realization each individual is able to observe (or sample) the present performance of another individual; she observes the action chosen and the payoff received by the sampled individual. Sampling occurs according to some exogenously given distribution. However, we do require that this procedure of sampling is symmetric, i.e., that individual 'c' samples individual 'd' with the same probability as individual 'd' samples individual 'c'. Random sampling will play a special role regarding aggregate population behavior.

The action chosen by an individual in the next round as a function of the information that she has acquired in the past will be called a *behavioral rule*. We want to focus on simple behavioral rules and thus limit the amount of information available to the individual about previous occurrences as follows. We assume that an individual forgets all about actions and observations from previous rounds. Hence, the rule determining play in the next round is a function of the payoffs achieved and actions taken both by oneself and the sampled individual in the present round. Moreover, we assume that an individual faces many multi-armed bandits that she can not distinguish a priori except for the fact that they all have the same set of possible actions. Hence, an individual must follow the same behavior (use the same behavioral rule) whenever she is faced with a multi-armed bandit with the same set of actions. Especially we have implicitly ruled out an individual that has enough memory capacity to be able to choose future actions by updating some prior. Similarly we have ruled out complex strategies as used in the classical multi-armed bandit problem (e.g., see Rothschild [13]). In the following consider the set of actions A and the interval $[\alpha, \omega]$ containing the realized payoffs fixed. As mentioned above, we follow two alternative approaches for selecting a behavioral rule for our setting.[3]

The first approach (formalized in Section 4.1) is from the standpoint of a boundedly rational individual. The individual enters a finite population of individuals who each face a multi-armed bandit. All multi-armed bandits yield the same payoff distributions. Entry occurs by replacing a randomly selected individual and adapting the action this individual chose in the last round. As mentioned above, the individual must use the same behavioral rule when facing each multi-armed bandit. The following assumptions are made. The individual measures her success in a given decision situation by the expected increase (or decrease) in payoffs her present choice yields in comparison to her previous choice. The individual wants to perform well in each situation in contrast to performing badly in some and very well in others. Hence, we aim to investigate if there is a rule in our setting that always leads to an increase in expected payoffs. At this point we add an additional bounded rational assumption, namely, we assume that the entering individual treats each round as if she had just entered. This means that the individual does not anticipate how her decisions affect future distributions of actions played in the population. A rule that (weakly) increases expected payoffs in any decision situation, i.e., in any state regardless of the underlying payoff distributions associated to the various actions, under the above assumption of ignorance of previous play, will be called *improving*. If an improving rule induces constant expected payoffs in all decision situations then the rule will be called *stationary*.

Our second approach to selecting behavioral rules (formalized in Section 4.2) takes a population-oriented perspective. In this scenario we will search

---

[3] Unless otherwise stated, we ignore the issue of which action to choose when the individual has no prior experience.

for a rule that, when played by each individual in a finite population, is expected to (weakly) increase average payoffs in the population over time in any decision situation. Such a rule will be called *payoff increasing*. For large (i.e., infinite) populations, this means that we search for a rule that induces a better reply dynamic (Hofbauer [9]) in all decision situations when the population is made up of replicas of the individual using this rule. In a model that includes evolutionary pressure on behavioral rules it is plausible that only payoff increasing rules can survive the entry of an alternative rule in all environments. Björnerstedt and Schlag [2] confirm this intuition, a result explained further in Appendix B.

The two approaches above have in common that a certain property must hold in all decision situations. A fundamental connection between these two approaches is that a rule is improving in the boundedly rational model if and only if it is payoff increasing in the population model. Hence, from the standpoint of either selection approach our initial aim will be to characterize the improving rules. Trivial examples of stationary improving rules are easily given, e.g., the self-explanatory rules 'never switch' and 'always switch'. We note that an improving rule is stationary if and only if it always induces constant average payoffs in the population. For this reason, we will focus special attention on the class of non-stationary improving rules.

One important aspect of our first lemma is that improving rules are imitating, i.e., that switching actions occurs by imitating the sampled individual. From this, the main theorem (Theorem 1) characterizes the set of improving rules. The criterion for an imitating rule to be improving is that the difference in the probabilities of switching when two individuals sample each other is *proportional* to the difference in the payoffs each individual achieved. The reason for this strong condition on the probabilities of switching relies on the linear structure of taking expectations. An intuitive and hence popular behavioral rule is the rule we call *'imitate if better'* (used e.g. by Ellison and Fudenberg [7] and Malawski [10]). This rule prescribes to imitate the action of the observed individual whenever she achieves a higher payoff. Our analysis reveals that this rule is not improving. Another consequence of our main theorem is that a bound on the set of feasible payoffs (in our model, payoffs are contained in $[\alpha, \omega]$) is necessary to obtain non stationary improving rules.

Our classification of the improving rules reveals a clear structure among the improving rules. This leads in both selection approaches to a natural criterion for selecting a best rule among the improving rules. In fact, in either selection approach the same behavioral rule is selected. This rule is a specific rule from the class of rules we call *proportional imitation rules* that were described at the beginning of this introduction. Especially, we obtain a unique prescription of how to choose actions in our model.

Next we investigate implications for the population adjustment process under random sampling when each individual is using our prescribed best behavioral rule. We show that the process governing the adaptation of actions in a large population can be approximated in the short run by a discrete version of the replicator dynamic (Taylor [20]) applied to this setting. Especially, with probability arbitrarily close to one, most individuals will be playing after a finite number of rounds an action that is best (measured in terms of expected payoffs) among those initially present.

In a further section we consider a two population random matching scenario. All the previous analysis and results are shown to apply to this generalized setting. Especially, individuals prefer to use a specific proportional imitation rule. Moreover, in an infinite population under random sampling in which each individual uses the optimal individual behavior, the adjustment is described by a discrete version of the replicator dynamic (Taylor [20]).

The rest of the paper is organized as follows. In Section two the basic payoff realization and sampling scenario is introduced. The feasible behavioral rules for using in this scenario are presented in Section three. Section four contains two alternative approaches to selecting a behavioral rule, each leading to the condition of improving. In Section five we present a first lemma concerning improving rules. In Section six this lemma is applied to show that 'imitate if better' is not improving. Section seven contains the central theorem of this paper, a complete characterization of an improving rule. In Section eight we select among the improving rules. Section nine contains the implications this has for the population dynamic. In Section ten we consider an alternative two population matching scenario. Section twelve contains a discussion. Appendix A contains some the proof of the main theorem. In appendix B some in sights to the evolutionary model of Björnerstedt and Schlag [2] are given.

2 The Payoff Realization and Sampling Scenario

Consider the following dynamic process of choosing actions, sampling and updating. Let P denote the underlying probability measure. Let W be a finite population (or set) of N individuals, $N \geq 2$. In a sequence of rounds, each individual must choose an action from a finite set of actions $A = \{1, .., n\}$, $n \geq 2$. Choosing the action i yields an uncertain payoff drawn from a probability distribution P_i with finite support in $[\alpha, \omega]$ where α and ω, $\alpha < \omega$, are exogenous parameters. Payoffs are realized independently of all other events. Let π_i denote the expected payoff generated by choosing action i, i.e., $\pi_i = \sum_{x \in [\alpha,\omega]} x P_i(x)$, $i \in A$. Then the tuple $\left\langle A, (P_i)_{i \in A} \right\rangle$ constitutes a *multi-armed bandit* or a game against nature. The set of all multi-armed bandits with action set A yielding payoffs in $[\alpha, \omega]$ will be denoted by $\mathcal{G}(A, [\alpha, \omega])$.

A state $s \in A^W$ of the population in a given round t is the description of the action that each individual is choosing in round t. Let $m_i = m_i(s)$ denote the number of individuals choosing the action i in state s, i.e., $m_i = |\{c \in W : s(c) = i\}|$ $(i \in A)$. Let $\Delta(A)$ be the set of probability distributions on A. For a given state s let $p \in \Delta(A)$ denote the probability distribution that is associated with randomly selecting an individual and observing the action she has chosen for this round, i.e., $p_i = m_i / N$. The set of all such probability distributions will be denoted by $\Delta^N(A)$, i.e., $p \in \Delta^N(A)$ and $i \in A$ implies $N \cdot p_i \in \mathbb{N}$. Given this notation, the average expected payoff of the population in state s, $\bar{\pi}(s)$, is given by $\bar{\pi}(s) = \sum p_i \pi_i$.

Between the rounds of payoff realization, each individual meets (or samples) another individual from the population and receives the following information. When individual 'c' samples individual 'd' $(c, d \in W)$, then individual 'c' observes the action 'd' used and the payoff 'd' achieved in the last round without observing the identity of ‘d’. For each individual ‘c’ this sampling occurs according to some exogenously given probability distribution $z_c \in \Delta(W \backslash \{c\})$ where $z_c(d)$ is the probability that individual ‘c’ samples individual ‘d’. z_c is called a *sampling rule* for individual ‘c’. The assignment of a sampling rule z_c to each individual c in the population will be called a sampling procedure. Formally, $z = (z_c)_{c \in W}$ is called a *sampling procedure* if $z_c \in \Delta(W \backslash \{c\})$ for all $c \in W$. A sampling procedure z will be called *symmetric* if for any $c, d \in W$ the probability of ‘c’ sampling ‘d’ (this event denoted by $c \rightsquigarrow d$) is the same as vice versa, i.e., $P(c \rightsquigarrow d) = P(d \rightsquigarrow c)$. In the following we will restrict attention to symmetric sampling procedures.

The above conditions restrict the variety of individual sampling procedures without specifying explicitly how the sampling rules of different individuals relate to each other. A scenario in which each individual is sampled at most once (due to time constraints) is equally feasible as one in which an individual can be sampled a multiple number of times. The sampling could be such that individuals sample independently, i.e., $P(c \rightsquigarrow d | d \rightsquigarrow c) = P(c \rightsquigarrow d)$. Similarly, we allow for a model in which individuals sample each other. In this case $c \rightsquigarrow d$ is the same event as $d \rightsquigarrow c$ for each $c, d \in W$, especially, $P(c \rightsquigarrow d | d \rightsquigarrow c) = 1$ and the sampling rule is symmetric. For example, a symmetric sampling procedure results from the following assumptions on individual sampling behavior. Individuals are located on a circle. Each individual randomly samples with equal probability among her $2m$ closest neighbors (m to the left, m to the right, $m < N/2$). Special attention will focus on the symmetric sampling procedures referred to as *random sampling*. Here each individual randomly samples an individual (with equal probability) from the population (except for herself), i.e., $P(c \rightsquigarrow d) = \frac{1}{N-1}$ for $c, d \in W$, $c \neq d$.

3 Behavioral Rules

Let A, $[\alpha, \omega]$ and N be fixed throughout the rest of the paper. The description of how an individual in our model chooses her next action whenever she faces a multi-armed bandit with action set A that realizes payoffs in $[\alpha, \omega]$ is summarized by a *behavioral rule*. Especially, the individual may not change her rule over time. We allow for an individual to use a randomizing device that generates independent events when they determine which action to choose in the next round. Following the assumptions on individual information and memory made in the introduction, a *behavioral rule* F is characterized by a function $F : A \times [\alpha, \omega] \times A \times [\alpha, \omega] \rightarrow \Delta(A)$ where $F(i, x, j, y)_k$ is the probability of choosing action k in the next round after previously choosing action i, receiving payoff x and sampling an individual who chose action j and received payoff y. For simplicity we will disregard the issue of which action to choose in the first round.

One of the simplest behavioral rules is the rule '*never switch*', formally defined by the behavioral rule F that satisfies $F(i, x, j, y)_i = 1$ for $i, j \in A$ and $x, y \in [\alpha, \omega]$. An opposite behavior is exhibited by the self explanatory rule '*always switch*'. A more plausible rule seems to be to act according to '*imitate if better*' (Ellison and Fudenberg [7]; Malawski [10]), i.e., use the rule F given by $F(i, x, j, y)_j = 1$ if $y > x$ and $F(i, x, j, y)_i = 1$ if $y \leq x$. The three rules described above belong to the class of behavioral rules that are based on imitation, i.e., either the individual does not change actions or she switches to the action used by the individual she sampled. More generally, we call a behavioral rule F *imitating* if $F(i, x, j, y)_k = 0$ when $k \notin \{i, j\}$ $(x, y \in [\alpha, \omega])$. The class of imitating rules, referred to as proportional imitation rules, described at the beginning of the introduction will play an important role in our analysis. A behavioral rule F is called a *proportional imitation rule* if F is imitating and there exists σ with $0 < \sigma \leq \frac{1}{\omega - \alpha}$ such that $F(i, x, j, y)_j = 0$ if $y \leq x$ and $F(i, x, j, y)_j = \sigma(y - x)$ if $y > x$, $i \neq j$ and $x, y \in [\alpha, \omega]$. The constant σ associated to a given proportional imitation rule will be referred to as the rate.[4]

---

[4] Switching behavior as displayed by proportional imitation rules appears in a paper by Cabrales [5] where it is justified through uniformly distibuted costs for switching actions.

Selection of Rules

The major part of our analysis concerns the issue of which behavioral rule an individual should choose in all of her future encounters of the scenario described in Section 2. The individual enters the setting by randomly replacing one of the individuals in the population (with equal probability). In her first round, the entering individual uses the action last chosen by the individual she replaced.[5] Exit occurs when the individual is replaced by a new entrant. The individual under consideration knows that she will be confronted with many such circumstances throughout her lifetime.

We present two alternative scenarios (or approaches) for analyzing which behavioral rule is best for an individual.

4.1 A Bounded Rational Approach

In the first scenario we consider a bounded rational individual. We make the following assumptions on how the individual evaluates the performance of a behavioral rule. The individual wants to avoid under all circumstances a decrease in her expected payoffs from one round to the next. When calculating expected payoffs in a given round the individual ignores the effect her behavior has on future states of the population. This means that she calculates the expected payoffs in a given state and a given round as if she just entered the population. A behavioral rule that fulfills this criterion will be called improving. This will be formalized in the following.

For a given behavioral rule F, and a given multi-armed bandit, let F_{ij}^k denote the probability of playing action k in the next round after playing action i and sampling an individual using action j $(i, j, k \in A)$. Then

$$F_{ij}^k = \sum_{x,y \in [\alpha,\omega]} F\left(i, x, j, y\right)_k P_i\left(x\right) P_j\left(y\right). \tag{1}$$

Let s^t be the state in round t. The expected increase in the payoffs, $eip\left(c, s^t\right)$,

---

[5] This assumption is made in order to simplify the presentation. It can be motivated directly by means of either of the two scenarios that will be introduced in the following.

of individual ‘c’ between rounds t and $t+1$, is given by

$$eip\left(c, s^t\right) = \sum_{r\in A} \sum_{d\in W\setminus\{c\}} P\left(c \rightsquigarrow d\right) F^r_{s^t(c)s^t(d)} \left(\pi_r - \pi_{s^t(c)}\right). \tag{2}$$

An individual’s expected increase in payoffs from rounds t to $t+1$ a priori to entering into the population in round t, denoted by $EIP_F\left(s^t\right)$, is given by

$$EIP_F\left(s^t\right) = \frac{1}{N} \sum_{c\in W} eip\left(c, s^t\right). \tag{3}$$

Due to the above ‘ignorance’ assumption, of previous play, $EIP_F\left(s^t\right)$ is also an individual’s expected increase in payoffs from rounds t to $t+1$ when she entered into the population in some round before round t. $EIP_F\left(s\right)$ will be called the *expected improvement* under F in state s. The behavioral rule F is called *improving* if $EIP_F\left(s\right) \geq 0$ for any state $s \in A^W$ and any multi-armed bandit in $\mathcal{G}\left(A, [\alpha, \omega]\right)$. An improving rule F is called *stationary* if $EIP_F\left(s\right) = 0$ for all states $s \in A^W$ and all multi-armed bandits in $\mathcal{G}\left(A, [\alpha, \omega]\right)$, otherwise it is called *non-stationary.*[6]

4.2 A Population Oriented Approach

In this alternative scenario we assume that an individual evaluates the performance of a rule in a population of replicas, i.e., in a situation where all other individuals use the same behavioral rule as she does. A population in which each individual follows the same behavior will be called a *monomorphic population.* We assume that the individual chooses a rule that causes average expected payoffs in a monomorphic population to increase in any feasible encounter. Such a behavioral rule will be called payoff increasing. A motivation for this condition through an evolutionary model of selecting

---

[6] The concept of improving is very closely related to the concept of *absolute expediency* defined by Sarin [14] in a slightly different context. Applied to our model, an absolutely expedient rule is an improving rule with the property that the expected improvement is strictly positive whenever not each action currently used in the population achieves the same expected payoff. As such this concept leads to a refinement of non-stationary improving rules.

rules in a large population is provided by Björnerstedt and Schlag [2] (see also Appendix B).

Formally, for a given round t, a given state s^t in round t, and a given multi-armed bandit, let $E_F\bar{\pi}'(s^t)$ denote the expected average payoff in round $t+1$ when all individuals are using the rule F, i.e.,

$$E_F\bar{\pi}'\left(s^t\right) = \frac{1}{N}\sum_{r\in A}\sum_{c,d\in W} P\left(c \rightsquigarrow d\right) F^r_{s^t(c)s^t(d)}\pi_r \ .$$

A behavioral rule F is called *payoff increasing* if $E_F\bar{\pi}'(s) \geq \bar{\pi}(s)$ for any state $s \in A^W$ and any multi-armed bandit in $\mathcal{G}(A,[\alpha,\omega])$.

4.3 An Equivalence Result

Rewriting (3), using (2), we obtain

$$EIP_F(s) = E_F\bar{\pi}'(s) - \bar{\pi}(s). \tag{4}$$

This leads to the following result.

Remark 1 *A behavioral rule is improving if and only if it is payoff increasing. An improving rule is stationary if and only if it induces constant average expected payoffs in a monomorphic population, i.e.,* $E_F\bar{\pi}'(s) = \bar{\pi}(s)$ *in any state* $s \in a^W$ *and in any multi-armed bandit in* $\mathcal{G}(A,[\alpha,\omega])$.

5 A First Lemma

Clearly, the rules 'never switch' and 'always switch' are stationary improving rules. Given Remark 1, an analysis of non-stationary improving rules is of interest from the viewpoint of either selection approach of Section 4. The following preliminary result contains a characterization of the improving rules that is independent of the population state. According to this result, a behavioral rule F is improving if and only if it is an imitating rule that satisfies the following condition. Consider two individuals using different actions, both following the same behavior F, that sample each other. Then it cannot be, a priori to observing the other's payoff, that the individual obtaining the strictly higher expected payoff of the two is more likely to switch actions.

Lemma 1 *The behavioral rule F is improving if and only if F is imitating and for any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$, for any $i, j \in A$, $i \neq j$,*

$$\left(F_{ij}^{j} - F_{ji}^{i}\right)(\pi_j - \pi_i) \geq 0. \tag{5}$$

The proof of the imitation property is quite intuitive. An individual will avoid switching to an action that she did not observe since it might be that the action she observed is a duplification of her own strategy (i.e., it generates the same probability distribution of payoffs) whereas all actions not observed lead necessarily to the worst outcome. Notice that imitation remains necessary to ensure the improving condition even after the event of receiving the lowest possible payoff α and sampling an individual who used the same action and also obtained α. This is because it may be that obtaining α is an unlucky event for the own action whereas it is the only outcome for any other action.

Proof. We will first show the "if" statement. Rewriting (3) for imitating rules yields

$$EIP_F(s) = \frac{1}{N} \sum_{c,d \in W} P(c \rightsquigarrow d)\, F_{s(c)s(d)}^{s(d)} \left[\pi_{s(d)} - \pi_{s(c)}\right].$$

Using the fact that the sampling rule is symmetric we obtain

$$EIP_F(s) = \frac{1}{N} \sum_{i<j} \left[\sum_{\substack{c:s(c)=i \\ d:s(d)=j}} P(c \rightsquigarrow d) \right] \left(F_{ij}^{j} - F_{ji}^{i}\right)(\pi_j - \pi_i), \tag{6}$$

which completes the proof of the "if" statement.

We will now show that improving rules are imitating. Assume that the behavioral rule F is improving. Let $x, y \in [\alpha, \omega]$, $i, j \in A$ and $r \in A \backslash \{i, j\}$ be such that $F(i, x, j, y)_r > 0$. Construct a multi-armed bandit in which $P_i(x) = P_i(y) = P_i(\omega) = \frac{1}{3}$, $P_j \equiv P_i$ and $P_k(\alpha) = 1$ for all $k \in A \backslash \{i, j\}$.

It follows that $\pi_i = \pi_j > \pi_k$. Choose $c, d \in W$ such that $P\left(c \rightsquigarrow d\right) > 0$ and consider a population state s such that $s(c) = i$, $s(d) = j$ and $m_i + m_j = N$. Then $F\left(i, x, j, y\right)_r > 0$ implies $EIP(s) < 0$ which contradicts the fact that F is improving.

Finally, we will show that an imitating rule F that violates (5) for some $i \neq j$ and some multi-armed bandit in $\mathcal{G}\left(A, [\alpha, \omega]\right)$ is not improving. Choose again $c, d \in W$ such that $P\left(c \rightsquigarrow d\right) > 0$ and consider a population state s such that $s(c) = i$, $s(d) = j$ and $m_i + m_j = N$. Since $\left(F^j_{ij} - F^i_{ji}\right)\left(\pi_j - \pi_i\right) < 0$, and following (6), $EIP_F\left(s\right) < 0$ which implies that F is not improving. ■

6 The Drawback of 'Imitate if better'

'Imitate if better' is a plausible rule. In fact, it performs well in multi-armed bandits in which uncertainty is driven solely through idiosyncratic shocks. Consider a multi-armed bandit in $\mathcal{G}\left(A, [\alpha, \omega]\right)$ with the following properties. There is a probability distribution Q with finite support and mean 0 such that $P_i\left(x\right) = \pi_i + Q\left(x\right)$ for each $i \in A$. Throughout this section, let F denote the rule 'imitate if better'. Then

$$F^j_{ij} - F^i_{ji} = \frac{1}{2}\sum_{x,y} Q\left(x\right) Q\left(y\right) \left[\begin{array}{c} F\left(i, \pi_i + x, j, \pi_j + y\right)_j - F\left(j, \pi_j + y, i, \pi_i + x\right)_i \\ +F\left(i, \pi_i + y, j, \pi_j + x\right)_j - F\left(j, \pi_j + x, i, \pi_i + y\right)_i \end{array} \right]$$

and hence, $F^j_{ij} - F^i_{ji} \geq 0$ when $\pi_j \geq \pi_i$. With (6) it follows that the expected improvement of 'imitate if better' is non negative in such a multi-armed bandit.

However, we will see 'imitate if better' generates negative expected improvement in some extremely simple multi-armed bandits; it can not distinguish between lucky payoffs and certain payoffs. Let $x \in \left(\alpha, \frac{\alpha+\omega}{2}\right)$. Consider a multi-armed bandit in which $P_1\left(x\right) = 1$, $P_2\left(\alpha\right) = \lambda$ and $P_2\left(\omega\right) = 1 - \lambda$ for some $0 < \lambda < 1$. It follows that

$$\pi_2 > \pi_1 \text{ if and only if } \lambda < \frac{\omega - x}{\omega - \alpha}.$$

n the other hand, $F^2_{12} = 1 - \lambda = F^1_{21}$ and hence,

$$F^1_{21} > F^2_{12} \text{ if and only if } \lambda > \frac{1}{2}.$$

Consequently, when $\frac{1}{2} < \lambda < \frac{\omega - x}{\omega - \alpha}$ then (5) is violated and hence 'imitate if better' is not improving.

7 A Complete Characterization

The fact that being improving is equivalent to being imitating and more likely to imitate an action with a higher expected payoff than vice versa (Lemma 1) is quite intuitive. The difficulty in finding improving rules is the fact that an individual is not able to condition her behavior on expected payoffs but must base her decision on realized payoffs. The following theorem constitutes the central result of this paper. It states a somewhat surprising characterization of the set of behavioral rules that are improving. According to this result only switching in a way that "net" switching behavior is linear in payoff differences ensures that an imitating rule is in fact improving. The consequent proof reveals that this strong characterization is due to the linear structure of taking expectations.

Theorem 1 *The behavioral rule F is improving if and only if*
i) F is imitating and
ii) for all $i, j \in A$, $i \neq j$ there exists $\sigma_{ij} = \sigma_{ji} \in \left[0, \frac{1}{\omega - \alpha}\right]$ such that

$$F(i, x, j, y)_j - F(j, y, i, x)_i = \sigma_{ij}(y - x) \text{ for all } x, y \in [\alpha, \omega]. \qquad (7)$$

From (7) we see that the rule 'imitate if better' is not improving, a fact also shown in Section 6.

Proof. (in the Appendix)

From (6) and (16) we obtain that

Corollary 1 *an improving rule is stationary if and only if $\sigma_{ij} = 0$ for all $i, j \in A$, $i \neq j$.*

It follows that non-stationary improving rules induce stochastic behavior. Moreover, non-stationary improving rules only exist if, as assumed throughout this paper, payoffs are contained in a pre-specified bounded interval $[\alpha,\omega]$. The later statement follows from the fact that σ_{ij} is bounded above by $\frac{1}{\omega-\alpha}$ (Theorem 1).

The next corollary supplements the characterization of improving rules given in Theorem 1.

Corollary 2 *Condition ii) in Theorem 1 holds if and only if the following condition holds:*

ii') for all $i,j\in A$, $i\neq j$, *either* $F(i,x,j,y)_j = F(j,y,i,x)_i$ *for all* $x,y\in[\alpha,\omega]$ *or there exists* $\sigma_{ij}=\sigma_{ji}>0$ *and a function* $g_{ij}:[\alpha,\omega]\times[\alpha,\omega]\to\mathbb{R}$ *such that for* $x,y\in[\alpha,\omega]$,

$$\begin{aligned} -\min\{x,y\} \leq g_{ij}(x,y) &\leq -\max\{x,y\}+\frac{1}{\sigma_{ij}}, \\ F(i,x,j,y)_j &= \sigma_{ij}\cdot(y+g_{ij}(x,y)) \ \textit{and} \\ F(j,y,i,x)_i &= \sigma_{ij}\cdot(x+g_{ij}(x,y)). \end{aligned}$$

Proof. The fact that ii') implies ii) follows directly. Conversely, let $i,j\in A$, $i\neq j$ and let F satisfy ii). If $\sigma_{ij}=0$ then ii) implies $F(i,x,j,y)_j = F(j,y,i,x)_i$ for all $x,y\in[\alpha,\omega]$. Assume now that $\sigma_{ij}>0$. Let $g_{ij}(\cdot,\cdot)$ be defined by $g_{ij}(x,y)=\frac{1}{\sigma_{ij}}F(i,x,j,y)_j - y$ $(x,y\in[\alpha,\omega])$. It follows that $-y\leq g_{ij}(x,y)\leq -y+\frac{1}{\sigma_{ij}}$ and $F(i,x,j,y)_j=\sigma_{ij}\cdot(y+g_{ij}(x,y))$. Together with ii) we obtain $F(j,y,i,x)_i = F(i,x,j,y)_j - \sigma_{ij}(y-x) = \sigma_{ij}\cdot(x+g_{ij}(x,y))$. This implies $-x\leq g_{ij}(x,y)\leq -x+\frac{1}{\sigma_{ij}}$ which completes the proof of condition ii'). ■

8 Selecting among Improving Rules

We now proceed with our selection of a best rule for the individual to use. In the following we show that there is a common subset of improving rules (which we will call dominant) that perform best according to the expected improvement they generate.

We will say that a behavioral rule F *dominates the improving rules* (or short, is a *dominant rule*) if there is no improving rule F', state s and multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ such that $EIP_{F'}(s) > EIP_F(s)$. A dominant rule always achieves a (weakly) higher expected improvement than any other improving rule. With (4) it follows that dominant rules are also the rules that maximize the expected increase in the average payoffs of a monomorphic population in any state among the set of payoff increasing rules. Hence, regarding either approach to selecting individual behavior, it is natural for an individual to choose a dominant rule if such a rule exists.

Following (6) and (16), the expected improvement, $EIP_F(s)$, of an improving rule F in state s is given by

$$EIP_F(s) = \left[\frac{1}{N} \sum_{i<j} \sum_{\substack{c:s(c)=i \\ d:s(d)=j}} P(c \rightsquigarrow d) \right] \sigma_{ij} (\pi_j - \pi_i)^2 \ . \qquad (8)$$

From (8) we obtain that the expected improvement of an improving rule only depends on the factors $(\sigma_{ij})_{\substack{i,j \in A \\ i \neq j}}$. Hence,

Proposition 1 *a behavioral rule is a dominant rule if and only if it is improving and for any* $i \neq j$, $\sigma_{ij} = \frac{1}{\omega - \alpha}$.

Among the set of dominant behavioral rules the proportional imitation rule with rate $\frac{1}{\omega-\alpha}$ (defined in Section 3), denoted by F^p, has the following unique properties.

Theorem 2 *i)* F^p *is the unique dominant rule that never imitates an action that achieved a lower payoff.*

ii) In any state and for any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$, F^p *minimizes the probability of switching among the set of dominant rules. Moreover,* F^p *is the unique dominant rule that has this property in all states and all bandits.*

iii) In any round t*, for any state in round* t *and for any multi-armed bandit in* $\mathcal{G}(A, [\alpha, \omega])$, F^p *minimizes the variance of the average payoffs in the monomorphic population playing* F^p *in round* $t+1$ *among the set of* *dominant improving rules. Moreover, F^p is the unique dominant rule that has this property for all states and all multi-armed bandits in $\mathcal{G}(A, [\alpha, \omega])$.*

Following part i) in Theorem 2, it can be argued that F^p is the unique dominant rule that performs best when payoffs are deterministic; realized payoffs never decrease in such degenerate bandits. Statement ii) implies that F^p is the dominant rule that changes actions the least number of times. From part iii) it follows that the monomorphic population using F^p exhibits minimal variance in average payoffs. Together with the fact that F^p is a dominant rule we conjecture that F^p maximizes the probability that average payoffs in a monomorphic population increase over time.

Proof. Statements i) and ii) follow easily from Corollary 2 since the proportional imitation rule is the unique dominant rule with $g_{ij}(x, y) = -\min\{x, y\}$. Part iii) follows from part ii) of Theorem 1 and some easy calculations. ■

The proportional imitation rule with rate $\frac{1}{\omega - \alpha}$ is improving. Moreover, it is dominant (Proposition 1), and hence always performs at least as well as any other improving rule regarding expected improvement. Finally, its unique properties among the set of dominant rules (Theorem 2) lead us naturally to strictly preferring it to any other dominant rule. We argue therefore that the proportional imitation rule with rate $\frac{1}{\omega - \alpha}$ is the best, i.e., the optimal, rule for an individual to use in our model. Especially, notice that the optimal rule does not depend on the size of the population N.

ne might mention that there is a dominant rule that requires less information than the dominant proportional imitation rule. Consider the imitating rule F that satisfies $F(i, x, j, y)_j = \frac{\omega - x}{\omega - \alpha}$ for $i, j \in A$ and $x, y \in [\alpha, \omega]$. We will call this rule the *dominant proportional reviewing rule.* [7] It can be easily shown (see Schlag [17]) for more details) that the dominant proportional reviewing rule is the unique dominant improving rule that does not depend on the sampled individual's payoff.

---

[7] Björnerstedt and Weibull (1993) and Gale et al. (1995) both use a variant of this rule in their model, the later interpret it on the basis of random aspiration levels.

9 Population Dynamics

In this section we consider the implications of the analysis in the previous sections. We have argued that any given individual should choose the dominant proportional imitation rule. If each individual in the population does so, we obtain a specific monomorphic population. Average expected payoffs in this population increase over time since everyone uses the same rule which is payoff increasing. In the following we will characterize the induced dynamic in large populations in more detail.

First we will derive a law of large numbers type of result to simplify the analysis of the dynamic. If sampling is random and independent and the population is large, then we will show that actual adjustment can be approximated by the expected adjustment, at least in the short run. In the following we will derive this approximation for any monomorphic population (i.e., all individuals use the same rule).

For a monomorphic population of size N which is in state $p^N(1) \in \Delta^N(A)$ in round 1, let $p^N(t) \in \Delta^N(A)$ be the random state in round t, $t = 2, 3, ...$ Let $\|\cdot\|$ denote the supremums norm.

Theorem 3 *Assume that sampling is random and independent. Assume that each individual is using the behavioral rule F. Then for each $\delta > 0$, $\epsilon > 0$ and $T \in \mathbb{N}$ there exists $N_0 \in \mathbb{N}$ such that for any population size $N > N_0$ and any $\tilde{p} \in \Delta^N(A)$, the event that $\left\|p^N(T) - p(T)\right\| > \delta$ occurs with a probability less than ϵ given that $p^N(1) = p(1) = \tilde{p}$ and $(p(t))_{t\in\mathbb{N}}$ satisfies*

$$p_i(t+1) = \sum_{j,r} p_j(t)\, p_r(t)\, F^i_{jr}, \ t \in \mathbb{N}. \tag{9}$$

Notice that, in fact, we approximated the stochastic adjustment in the finite population by the limit of the expected adjustment when taking the population size to infinity.

Proof. We will first prove the statement for $T = 2$. Let $i \in A$ and fix $\tilde{p} \in \Delta^N(A)$. For $c \in W$ let $z_i(c)$ be the random variable such that $z_i(c) = 1$ if individual 'c' uses action i in round two, otherwise $z_i(c) = 0$. Then

$$P(z_i(c) = 1) = \frac{m_{s(c)} - 1}{N-1} F^i_{s(c)s(c)} + \sum_{j \neq s(c)} \frac{m_j}{N-1} F^i_{s(c)j}$$

and $p_i^N(2)=\frac{1}{N}\sum_{c\in W} z_i(c)$. Then

$$E\left[p_i^N(2)\right]=\frac{N}{N-1}\sum_{j,r}\tilde{p}_j\tilde{p}_r F^i_{jr}-\frac{1}{N-1}\sum_{j}\tilde{p}_j F^i_{jj}. \tag{10}$$

Since $z_i(c)$ and $z_i(d)$ are independent variables for $c\neq d$ and $VAR(z_i(c))\leq 1$ it follows that $VAR\left(p_i^N(2)\right)\leq\frac{1}{N}$. Applying Tschebysheff's inequality we obtain that the event $\left\{\left|p_i^N(2)-E\left[p_i^N(2)\right]\right|>\frac{\delta}{2}\right\}$ occurs with a probability of less than $\frac{4}{N\delta^2}$. Let N_0 be such that $\frac{4}{N_0\delta^2}<\epsilon$ and $\left|E\left[p_i^N(2)\right]-\sum_{j,r}\tilde{p}_j\tilde{p}_rF^i_{jr}\right|<\frac{\delta}{2}$ for $N>N_0$. Then $\left\{\left|p_i^N(2)-\sum_{j,r}\tilde{p}_j\tilde{p}_rF^i_{jr}\right|>\delta\right\}$ occurs with probability less than ϵ when $N>N_0$. Since N_0 has been chosen independent of $\tilde{p}$ the proof for $T=2$ is complete.

We will now prove the statement for $T=3$ by iterating the proof for $T=2$. Let $\delta>0$ and $\epsilon>0$ be given. Let $f:\Delta(A)\times\Delta(A)\rightarrow\mathbb{R}$ be defined by $f(p)_i=\sum_{j,r}p_jp_rF^i_{jr}$, $i\in A$. Let $p^N(t,\tilde{p})$ be the random state in round t given state $\tilde{p}\in\Delta^N(A)$ in round one $(t>1)$. Since f is a continuous function on a compact space there exists $\beta\in\left(0,\frac{\delta}{2}\right)$ such that $\|f(w)-f(w')\|<\frac{\delta}{2}$ if $\|w-w'\|<\beta$. Let μ be such that $(1-\mu)^2=1-\epsilon$. Following the proof for $T=2$ there exists N_0 such that for $N>N_0$ and $\tilde{p}\in\Delta^N(A)$, $P\left(\left\|p^N(2,\tilde{p})-f(\tilde{p})\right\|<\beta\right)>1-\mu$. For $N>N_0$ we therefore obtain that

$$\begin{aligned}
& P\left(\left\|p^N(3,\tilde{p})-f(f(\tilde{p}))\right\|\leq\delta\right)\\
= & \sum_{w\in\Delta^N(A)}P\left(\left\|p^N(2,w)-f(f(\tilde{p}))\right\|\leq\delta\right)P\left(p^N(2,\tilde{p})=w\right)\\
\geq & \sum_{w\in\Delta^N(A):\|w-f(\tilde{p})\|<\beta}P\left(\left\|p^N(2,w)-f(w)\right\|\leq\frac{\delta}{2}\right)P\left(p^N(2,\tilde{p})=w\right)\\
\geq & (1-\mu)^2=1-\epsilon
\end{aligned}$$

which completes the proof for $T=3$. The proof for the more general case of $T>3$ follows similarly using induction. ■

According to the above theorem the adjustment of a monomorphic population can be approximated by the deterministic process $\left(p^t\right)_{t\in\mathbb{N}}$ that satisfies (9). If the underlying rule F is improving then, using (16), (9) simplifies to

$$p_i^{t+1}=p_i^t+p_i^t\sum_{i,j\in A}\sigma_{ij}p_j^t\cdot(\pi_i-\pi_j). \tag{11}$$

onsequently, if F is improving with underlying $\sigma_{ij} > 0$ for all $i \neq j$, then all individuals in an infinite monomorphic population under random sampling will in the long run play actions achieving maximal expected payoff among those that were initially present, i.e., actions in $\arg\max_{i \in A} \{\pi_i, p_i^1(1) > 0\}$. In fact, it can easily be shown that the converse of this statement is also true (use Lemma 1 and (19)). Especially, if all indivduals in an infinite population under random sampling use the rule 'imitate if better' then they will eventually all be choosing the inefficient action in the gamble of Section 6 when $\frac{1}{2} < \lambda < \frac{\omega - x}{\omega - \alpha}$ and $p_1^1 \in (0,1)$.

If F is a dominant improving rule (e.g., the dominant proportional imitation rule), then we obtain

$$p_i^{t+1} = p_i^t + \frac{1}{\omega - \alpha} \left[\pi_i - \bar{\pi}\left(p^t\right)\right] \cdot p_i^t, \qquad (12)$$

where $\bar{\pi}(p) = \sum_i \pi_i \cdot p_i$ is the average payoff in the state $p \in \Delta(A)$. Hence, if each individual uses her optimal rule then the dynamic adjustment of the population is approximated in the short run by (12) — a discrete version of the replicator dynamic (Taylor [20]) applied to multi-armed bandits.

10 A Two-Population Random Matching Setting

Previously we assumed that the payoff distribution generated by an action is stationary over time. In this section we consider a scenario in which individuals face non stationary multi-armed bandits; individuals obtain payoffs by being randomly matched in pairs to play a game. We will extend the selection approaches from Section 4 and show that the dominant proportional imitation rule is still the unique optimal rule. The resulting aggregate behavior of an infinite population in which each individual is using the optimal rule will be described by the a discrete version of the replicator dynamic (Taylor [20]).

onsider two finite, disjoint populations W_1 and W_2, each of size N, also referred to as population one and two. In a sequence of rounds each individual must choose an action and is then matched against an individual from the opposite population. Let A_i be the finite set of actions available to an individual in population i, $i = 1, 2$. When an individual in population one using action $i \in A_1$ is matched with an individual in population two using action $j \in A_2$, the individual in population k achieves an uncertain payoff drawn from a given, independent probability distribution P_{ij}^k, $k = 1, 2$. Associating player i to being an individual in population i, the tuple $\left\langle A_1, A_2, \left(P_{ij}^1\right)_{\substack{i\in A_1\\ j\in A_2}}, \left(P_{ij}^2\right)_{\substack{i\in A_1\\ j\in A_2}} \right\rangle$ defines an *asymmetric two player normal form game*. In the following we will restrict attention to the class of asymmetric two player normal form games, denoted by $\mathcal{G}\left(A_1, A_2, [\alpha_1, \omega_1], [\alpha_2, \omega_2]\right)$, in which player k has action set $A_k, k = 1, 2$, where P_{ij}^1 has finite support in $[\alpha_1, \omega_1]$ and P_{ij}^2 has finite support in $[\alpha_2, \omega_2]$ for all $i \in A_1$ and $j \in A_2$; $\alpha_1 < \omega_1$ and $\alpha_2 < \omega_2$ are given. For a given asymmetric game, let $\pi_1(\cdot)$ and $\pi_2(\cdot)$ be the bilinear functions on $\Delta(A_1) \times \Delta(A_2)$ where $\pi_k(i, j)$ is the expected payoff to player k when player one is using action i and player two is using action j, i.e., $\pi_k(i, j) = \sum_{x\in[\alpha_k,\omega_k]} x P_{ij}^k(x)$, $k = 1, 2$.

Individuals are matched at random, formally this means the following for an individual in population one. Let $s_1 \in (A_1)^{W_1}$ be the current state in population one and let $p \in \Delta^N(A_1)$ be the associated population shares. Similarly let $s_2 \in (A_2)^{W_2}$ and $q \in \Delta^N(A_2)$ be the corresponding expressions for population two. Then an individual in population one is matched with an individual in population two using action $j \in A_2$ with probability q_j. Since we consider random matching, $\pi_1(i, q)$ specifies the expected payoff of an individual in population one using action $i \in A_1$ and $\pi_1(p, q)$ specifies the average payoff in population one in this state. Especially, given the current state, each individual in population one is facing a multi-armed bandit $\left\langle A_1, (P_i')_{i\in A} \right\rangle$ in $\mathcal{G}(A_1, [\alpha_1, \omega_1])$ where $P_i'(x) = \sum_{j\in A} q_j P_{ij}^1(x)$ for $x \in [\alpha_1, \omega_1]$.

Sampling occurs within a population and is otherwise performed as in the multi-armed bandit setting.

A *behavioral rule* F for an individual in population k is a function $F : \Delta(A_k) \times [\alpha_k, \omega_k] \times \Delta(A_k) \times [\alpha_k, \omega_k] \to \Delta(A_k)$, $k = 1, 2$.

We will now analyze the problem of selecting a behavior for an individual in population one. onsider first the bounded rational approach of Section 4.1. Notice that in any given state the game appears as a multi-armed bandit. An individual might not recognize that she is being matched to play an asymmetric game. She might also ignore the possibility that individuals in the opposite population have of changing their action. Under either of these assumptions the best behavioral rule for an individual in population one remains as in the multi-armed bandit setting the proportional imitation rule with rate $\frac{1}{\omega_1-\alpha_1}$.

onsider now the approach of Section 4.2 where an individual evaluates her behavior in a population of replicas. Let F denote this individual's behavioral rule. Let $(s_1, s_2) \in (A_1)^{W_1} \times (A_2)^{W_2}$ be the state in round t. Then the switching probabilities are given by

$$\begin{aligned} F_{jr}^{i}(s) &= \sum_{u} \frac{n_u (n_u - 1)}{N (N-1)} F(j, \pi_1(j,u), r, \pi_1(r,u))_i \\ &\quad + \sum_{u \neq v} \frac{n_u n_v}{N(N-1)} F(j, \pi_1(j,u), r, \pi_1(r,v))_i \end{aligned}$$

where $n_k = |\{c \in W_2 : s_2(c) = k\}|$ for $k \in A_2$ $(i, j, r \in A_1)$. If each individual in population one is using F then the expected proportion of individuals choosing action $i \in A_1$ in round $t+1$, denoted by $E_F p_i'(s)$ given state $s = (s_1, s_2) \in (A_1)^{W_1} \times (A_2)^{W_2}$ in round t is given by

$$E_F p_i'(s) = \frac{1}{N} \sum_{j,r \in A_1} \sum_{\substack{c,d \in W_1 \\ s_1(c)=j, s_1(d)=r}} P(c \rightsquigarrow d) \cdot F_{jr}^{i}(s) \ \ , i \in A_1. \qquad (13)$$

We assume that an individual chooses a rule that is payoff increasing in any state and in any asymmetric game when the actions played in the opposite population remain unchanged. In other words, she chooses a rule that is expected to produce a better reply to the present state in each state and bandit. Formally, we will say that a behavior rule F of an individual in population one *induces better replies* if for all asymmetric games in $\mathcal{G}(A_1, A_2, [\alpha_1, \omega_1], [\alpha_2, \omega_2])$ and all states $s = (s_1, s_2) \in (A_1)^{W_1} \times (A_2)^{W_2}$,

$$\sum_{i \in A_1} \pi_1(i, q(t)) \cdot E_F p_i'(s) - \sum_{i \in A_1} \pi_1(i, q(t)) \cdot p_i(t) \ \ \geq 0 \ . \qquad (14)$$

In Section B we motivate the above condition as a necessary condition for the survival of a rule in an evolutionary model when the population is large.

Following practically the same arguments as in Section 4.2 we obtain that a rule induces better replies if and only if it is improving. Especially, if each individual in population one is using the same rule that induces better replies, then (13) simplifies to

$$E_F p_i'(s) = p_i + \frac{1}{N} \sum_{i,j \in A_1} \sum_{\substack{c,d \in W_1 \\ s_1(c)=i, s_1(d)=j}} P(c \rightsquigarrow d)\, \sigma_{ij} \left[\pi_1(i,q) - \pi_1(j,q)\right] \ , \ i \in A_1.$$

Hence it follows that an improving rule is dominant if and only if the left hand side in (14) is maximized for all states and all asymmetric games in $\mathcal{G}(A_1, A_2, [\alpha_1, \omega_1], [\alpha_2, \omega_2])$.

Finally, analogue to property iii) in Theorem 2, the dominant proportional imitation rule for population one is the unique dominant rule that always minimizes the variance in the adjustment in population one when both populations are finite.

Following the above, one may argue that the proportional imitation rule with rate $\frac{1}{\omega_1 - \alpha_1}$ is the optimal rule for an individual in population one in this random matching setting in both the individually rational and the population-oriented selection approach.

We now consider population behavior under random sampling when each individual follows her optimal rule. Using a law of large numbers type of argument or following the direct approach as in Theorem 3 (see Schlag [17]) the short run behavior of a large population is approximated by the dynamic adjustment in an infinite population. This process is given by

$$\begin{aligned} p_i^{t+1} &= p_i^t + \frac{1}{\omega_1 - \alpha_1} \left[\pi_1\left(i, q^t\right) - \pi_1\left(p^t, q^t\right)\right] p_i^t, \ i \in A_1, \\ q_j^{t+1} &= q_j^t + \frac{1}{\omega_2 - \alpha_2} \left[\pi_2\left(p^t, j\right) - \pi_2\left(p^t, q^t\right)\right] q_j^t, \ j \in A_2 \ . \end{aligned} \tag{15}$$

Notice that (15), which is the two population analogue to (12), is a discrete version of the replicator dynamic (Taylor [20]).

11 Discussion

In this section we discuss some our assumptions and relate our work to the existing literature.

An important assumption concerning the selection of rules in our model is that we search for rules that perform well in any situation. hoosing a rule that is not improving will lead to a decrease in expected payoffs in some situations. Hence, such a choice implicitly means that the individual considers good performance in some situations enough to offset bad performance in others. We ruled out any such a priori bias of the individual and hence only improving rules came into question. Of course, one might consider a model in which an individual has a priori more information about the setting. Similarly, individuals might have better memory capabilities to be able to acquire information about the bandit while playing. In such alternate models, our analysis can be considered a benchmark, especially the proportional imitation rule is a useful rule to use while gathering information needed for a more elaborate rule. Sarin [14] incorporates the same idea of performing well in any situation into one of the axioms he postulates in a model of individual learning in two person games. Sarin calls this axiom *absolute expediency.* Applied to our setting, a rule satisfying this axiom is a payoff increasing rule such that expected payoffs increase strictly when not all actions present in the population achieve the same expected payoff. With Theorem 1 and (8) it follows that a rule is absolutely expedient under random sampling if and only if it is improving with underlying switching rates $\sigma_{ij} > 0$ for all $i \neq j$.

The central theme of our analysis is the selection of an individual's behavioral or learning rule, the description of what to do whenever a decision must be made. We search for behavioral rules that perform well in each situation. Our notion of performing well leads to the condition of *improving*, a rule performing better than any other improving rule in any situation is called *dominant.* A rule selected among the dominant rules is called *optimal.* An individual's decision is based on the information available about the specific situation. Naturally, different informational assumptions lead to the selection of different behavioral rules.

In our model, the information of the individual is extremely limited, especially she only observes the performance of one other individual in between rounds. Here the proportional imitation rule is a dominant rule, this rule is argued to be the unique optimal rule. When each individual uses the optimal rule we obtain the replicator dynamic. Our model is the first to reveal a derivation of the replicator dynamic from a model in which individual behavior is chosen optimally. Others have been able to construct individual behavior rules that lead to the replicator dynamic (Björnerstedt and Weibull [3]; Cabrales [5]; Gale et al. [8]), however they did not choose to justify individual behavior. Axiomatizations of learning rules in slightly different contexts have also lead to the replicator dynamic (Easley and Rustichini [6]; Sarin [14], in combination with the paper by Börgers and Sarin [4]). However, it should be noted that the basic approach in these models differs fundamentally from our approach — the former models contain axioms concerning the functional form of a desirable learning rule whereas the selection of rules in our model is based entirely on individual information and induced performance.

The existence of dominant rules in our setting is quite a surprising result. In a recent investigation we expand our model and assume that an individual may observe two individuals between rounds (Schlag [18]). Here dominant rules no longer exist. However, we find a simple rule (a modification of the proportional imitation rule) that is best at performing better than the optimal rule based on one observation. Especially, when each individual in an infinite population is using this simple rule, aggregate behavior follows an aggregate monotone dynamic as defined by Samuelson and Zhang [15].

Consider alternatively the situation where the individual has perfect information about the setting and about the current distribution of the actions played in the population(s). A behavioral rule now becomes a function that includes this additional information. In this alternative setting, it follows immediately that playing a best response is the unique dominant rule. The resulting adjustment process of the population, when each individual uses a dominant rule, is trivial in the multi-armed bandit setting; all individuals immediately adapt an action that achieves the highest expected payoff. In the two person game setting of Section 10 the adjustment follows the best response dynamic (Matsui [11]). omparing this result to ours (see (15)) we see that the replicator dynamic and the best response dynamic compromise extreme points in the class of adjustment dynamics based on individually optimal behavior.

A sort of intermediate case regarding informational assumptions is a scenario where individuals know the expected payoffs of both the action they use and of the action they sample. Although this assumption is difficult to motivate it is quite popular in the literature (e.g., see Björnerstedt and Weibull [3]; Hofbauer [9]). Under these informational assumptions, the rule 'imitate if better' is the unique dominant rule. An infinite population in which each individual uses the dominant rule eventually learns which of the actions initially present achieves the highest expected payoff. In this context, one might argue that our analysis reveals that the dominant proportional imitation rule comes under minimal information requirements closest to the performance of 'imitate if better' in the model where expected payoffs are observable.

There are some alternative models that analyze why individuals might imitate. Banerjee [1] presents a model in which individuals imitate for hope that the observed individual has more information. Rogers [12] presents an example of a changing environment in which individuals imitate in order to evade search costs incurred for finding the best action on their own. The evolutionarily stable proportions of individual learning (i.e., of individuals that search for the best action on their own) and of social learning (i.e., of individuals that adapt their action by imitating others) are computed. In contrast to our model, individual payoffs are not observable in the model of Rogers [12].

Finally, we want to mention that Malawski [10] conducted experiments in the two population random matching setting of Section 10. He refuted an imitation hypothesis because of the high proportion of individuals switching to actions other than the one observed in the last round (over 30%). Instead, the data is partially explained with aspiration level learning, a model that disregards the observed performance of others. In the mean time, Malawski and Schlag have informally reviewed the data from this experiment and discovered that observations of the performance of others, in fact, differences between others and own performance, does influence switching behavior. An extensive reevaluation of the data of the experiment of Malawski has therefore been planned.

References

[1] A. V. Banerjee, A Simple Model of Herd Behavior, *Quart. J. Econ.* **107** (1992), 797-818.

[2] J. Björnerstedt, and K. H. Schlag, "The Evolution of Imitative Behavior," mimeo, University of Bonn and Stockholm University, 1995.

[3] J. Björnerstedt, and J. Weibull, "Nash Equilibrium and Evolution by Imitation," In K. Arrow and E. Colombatto (eds.), Rationality in Economics, New York: Macmillan (forthcoming), 1993.

[4] T. Börgers, and R. Sarin, "Learning Through Reinforcement and Replicator Dynamics," Disc. Paper No. 93-19, University College of London, 1993.

[5] A. Cabrales, "Stochastic Replicator Dynamics," mimeo, University of California, San Diego, 1993.

[6] D. Easley, and A. Rustichini, "Choice Without Beliefs," mimeo, Cornell University and C.O.R.E., 1995.

[7] G. Ellison, and D. Fudenberg, Word-Of-Mouth Communication and Social Learning, *Quart. J. Econ.* **440** (1995), 93-125.

[8] J. Gale, K. G. Binmore, and L. Samuelson, Learning to be Imperfect: the Ultimatum Game, *Games Econ. Beh.* **8** (1995), 56-90.

[9] J. Hofbauer, "Imitation Dynamics for Games," University of Vienna, mimeo, 1995.

[10] M. Malawski, “Some Learning Processes in Population Games,” Inaugural-Dissertation, University of Bonn, 1989.

[11] A. Matsui, Best Response Dynamics and Socially Stable Strategies, *J. Econ. Theory* **57** (1992), 343-362.

[12] A. Rogers, Does Biology Constrain Culture? *Amer. Anthropol.* **90** (1989), 819-831.

[13] M. Rothschild, “A Two-Armed Bandit Theory of Market Pricing,” *J. Econ. Theory* **9** (1974), 185-202.

[14] R. Sarin, “An Axiomatization of the Cross Learning Dynamic,” mimeo, University of California, San Diego, 1993.

[15] L. Samuelson, and J. Zhang, Evolutionary Stability in Asymmetric Games, *J. Econ. Theory* **57** (1992), 363-391.

[16] L. J. Savage, *The Foundations of Statistics*, John Wiley and Sons, New York, 1954.

[17] K. H. Schlag, “Why Imitate, and if so, How? Exploring a Model of Social Evolution,” University of Bonn, Disc. Paper **B-296**, Bonn, 1994.

[18] K. H. Schlag, “Which One Should I Imitate?,” University of Bonn, mimeo, 1996.

[19] R. Schmalensee, “Alternative Models of Bandit Selection,” *J. Econ. Theory* **10** (1975), 333-342.

[20] P. Taylor, Evolutionarily Stable Strategies With Two Types of Players, *J. Applied Prob.* **16** (1979), 76-83.

A The Proof of Theorem 1

Proof. We will first show that conditions i) and ii) are sufficient. Let F be an imitating behavioral rule that satisfies condition ii). (1) and (7) imply

$$F_{ij}^{j} - F_{ji}^{i} = \sigma_{ij} \left(\pi_j - \pi_i \right). \tag{16}$$

So together with Lemma 1 it follows that F is improving.

We will now prove the necessity of conditions i) and ii). Let F be improving and fix $i, j \in A$ with $i \neq j$. Let $g(x,y) := F(i,x,j,y)_j - F(j,y,i,x)_i$ for $x, y \in [\alpha, \omega]$. For given $x, y \in [\alpha, \omega]$, consider the multi-armed bandit where $P_i(x) = P_j(y) = 1$. Then $F_{ij}^j = F(i,x,j,y)_j$ and hence following Lemma 1,

$$g(x,y) \geq 0 \text{ and } g(y,x) \leq 0 \text{ whenever } y > x. \tag{17}$$

Moreover, using arguments involving symmetry it follows that $g(x,x) = 0$ for all $x \in [\alpha, \omega]$. Next we will show that

$$\frac{g(x,y)}{y-x} = \frac{g(x,z)}{z-x} \; \forall y < x < z. \tag{18}$$

Given $y < x < z$, consider a multi-armed bandit where $P_i(x) = 1$, $P_j(y) = \lambda$ and $P_j(z) = 1 - \lambda$, $0 \leq \lambda \leq 1$. Then $\pi_j > \pi_i$ if and only if $\lambda < \frac{z-x}{z-y} =: \lambda^*$. It follows from Lemma 1 that

$$\begin{aligned} F_{ij}^j - F_{ji}^i = \lambda g(x,y) + (1-\lambda) g(x,z) &\geq 0 \text{ if } \lambda < \lambda^* \text{ and} \\ \lambda g(x,y) + (1-\lambda) g(x,z) &\leq 0 \text{ if } \lambda > \lambda^* \end{aligned}$$

Therefore, $\lambda^* g(x,y) + (1-\lambda^*) g(x,z) = 0$, which, after simplification, shows that (18) is true.

Following (18) there exists $\sigma_{ij} : (\alpha, \omega) \to \mathbb{R}_0^+$ such that $g(x,y) = \sigma_{ij}(x) \cdot (y - x)$ for all $x, y \in (\alpha, \omega)$. Exchanging the roles of i and j implies that $\sigma_{ij} = \sigma_{ji}$ is a constant. Hence, we have shown (7) for all $x, y \in (\alpha, \omega)$. Looking back at the above proof we see that the explicit values of α and ω did not enter the argument. Hence, (7) holds for all $x, y \in [\alpha, \omega]$.

Finally, $\sigma_{ij}(\omega - \alpha) = g(\alpha, \omega) \leq F_{ij}^j \leq 1$ implies $\sigma_{ij} \leq \frac{1}{\omega - \alpha}$. ■

B Some Notes on an Evolutionary Model

The payoff increasing condition in Section 4.2 was partially motivated through evolutionary arguments, thereby citing Björnerstedt and Schlag [2]. In the following we will give some more insights to this argument, however, for the precise arguments we must refer directly to the paper by Björnerstedt and Schlag [2].

Björnerstedt and Schlag [2] consider an infinite population under random sampling in the setting of Sections 2 and 3 for $A = \{1, 2\}$. Behavioral rules are under selected infrequently based on the replicator dynamic. In the following we demonstrate why all individuals will not adapt the same rule if this rule is not improving. Consider a behavioral rule F that is not improving. Then there is a multi-armed bandit in which the induced switching probabilities contradict Lemma 1. This means, w.l.o.g., that there is a gamble in which $\pi_1 > \pi_2$ and either $F_{11}^2 > 0$ or $F_{12}^2 > F_{21}^1$ holds. When each individual is using F, then

$$p_1^{t+1} = \left(p_1^t\right)^2 \left(1 - F_{11}^2\right) + p_1^t p_2^t \left(F_{21}^1 - F_{12}^2\right) + \left(p_2^t\right)^2 F_{22}^1 \ , \tag{19}$$

and hence not all individuals will adapt the better action, i.e., action one, in the long run when starting from a completely mixed initial population state. Now consider an alternative rule F' that prescribes to play action one regardless of the observations. If most individuals are playing F and only few are playing F' then the individuals using F start adapting the worse action as they would do if all individuals were using F. Consequently, the individuals using F perform on average strictly worse than those using F' and hence the proportion using F decreases.

Björnerstedt and Schlag [2] show that this can not happen when F is a non-stationary improving rule. When most individuals are using F then from any initial distribution of actions played in the population the individuals using F adapt the better action and thus perform equally well as F' before the alternative rule F' takes over a substantial proportion of the population.

Concerning what happens in an evolutionary model in the two population random matching setting of Section 10 we can no longer refer to Björnerstedt and Schlag [2] who only deal with multi-armed bandits. Never-the-less, the arguments made in the previous paragraph imply that a necessary condition for a rule to survive in an evolutionary framework in this more general setting is that it does not decrease expected payoffs when all individuals in the opposite population play an action that maximizes their expected payoff.

Although this condition is weaker than the ‘induce better replies’ condition of Section 10, rules that have this property in all games must never-the-less be improving.