JOURNAL OF ECONOMIC THEORY **78**, 130–156 (1998)
ARTICLE NO. ET972347

Why Imitate, and If So, How?

A Boundedly Rational Approach to Multi-armed Bandits[1]

Karl H. Schlag

Economic Theory III, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany

Received December 23, 1994; revised June 19, 1997

Individuals in a finite population repeatedly choose among actions yielding uncertain payoffs. Between choices, each individual observes the action and realized outcome of *one* other individual. We restrict our search to learning rules with limited memory that increase expected payoffs regardless of the distribution underlying their realizations. It is shown that the rule that outperforms all others is that which imitates the action of an observed individual (whose realized outcome is better than self) with a probability proportional to the difference in these realizations. When each individual uses this best rule, the aggregate population behavior is approximated by the replicator dynamic. *Journal of Economic Literature* Classification Numbers: C72, C79, D83.

1. INTRODUCTION

Imitation, as opposed to innovation, is the act of copying or mimicking the action of others. Imitation is a commonly observed behavior of human decision making.[2] We ask why individuals should imitate, and what sort of imitation rule they should adopt. First we identify a uniquely optimal individual rule and then derive implications for societies where each individual uses this rule. Optimality is determined according to two different perspectives: that of a boundedly rational individual and that of a social planer. Both approaches lead to the same unique prescription of how to choose future actions:

[1] This paper developed out of earlier unpublished work (Schlag, 1994). 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] Many recent models of social learning consider individuals who select future actions by imitating others (e.g., Banerjee [1]; Björnerstedt and Weibull [4]; Cabrales [7]; Ellison and Fudenberg [9]; Gale *et al.* [11]; Helbing [12]; Hofbauer [13]; Rogers [17]).

0022-0531/98 $25.00

- follow an imitative behavior, i.e., change actions only through imitating others

- never imitate an individual that performed worse than you

- imitate an individual that performed better with a probability that is proportional to how much better this individual performed.

Rules meeting these three criteria are called *Proportional Imitation Rules* herein. When each individual in a large society adopts this optimal rule then the stochastic process governing learned choices throughout society is approximated in the short run by the replicator dynamic (Taylor [25]).

The basic decision problem is modelled as a *multi-armed bandit*. An individual must repeatedly choose an action from a finite set of actions A. Actions yield uncertain payoffs. Payoffs are realized independently, their distribution has finite support, and belongs to a bounded interval $[\alpha, \omega]$. Multi-armed bandits have wide application in economics and behavioral sciences; the arm chosen can be, e.g., choice of technology or managerial structure within industries, setting prices under uncertain demand, or visit of a restaurant of uncertain quality.[3]

In our model, identical individuals belong to a finite population in which, periodically, new individuals replace (some) existing ones. Each individual is equally likely to be replaced regardless of prior durations. Individuals in the population face, independently and repeatedly, the same multi-armed bandit. Individuals do not know the probability distributions governing payoffs realized by the arms. Instead, they gather information from each other in the following way. On entry an individual observes the previous choice and realized payoff of the individual replaced. Before each payoff realization each individual observes (or *samples*) the previous choice and realized payoff of one other individual. Sampling is independent of actions or realized payoffs.

In the classical multi-armed bandit setting, an individual has infinite memory and constantly updates a subjective prior over possible payoff distributions (Rothschild [18]). We restrict attention to simpler individual behavior by assuming that an individual forgets all information she acquired before the last payoff realization. Hence, the *behavioral rule*, the rule determining an individual's next choice, is a function of the payoffs achieved and actions taken both that individual (or by the replaced individual) and by the individual sampled in the previous round.[4] Each individual must commit to a behavioral rule before entering the population.

[3] (Ellison and Fudenberg [9]; Schmalensee [24])

[4] We ignore the issue of which action individuals choose at the beginning of time when there is no one to replace.

We will determine which of these rules is 'optimal' from two distinct perspectives.

The first approach (formalized in Section 5.1) assumes boundedly rational individuals. Here individuals are myopic, only interested in how rules perform upon first encounter of the bandit. Thus, the entering individual acts as if she were to exit the population after one round. The description of the action which each individual in the population chooses in a given round is called a *population state*; *entry state* is the population state at an individual's entry. The performance of a behavioral rule depends on the entry state and the payoff distribution of the commonly experienced multi-armed bandit. It also depends on the realization of 'objective'[5] uncertainty, i.e., point of entry, sample and payoff. Individuals are assumed to be risk neutral towards objective uncertainty, i.e., lotteries induced by the realization of objective uncertainty are compared based on expected payoff.[6]

One feasible behavioral rule, *Never Switch*, is to forever pull the arm last chosen by the individual you replaced. The expected payoff of this rule will reflect the information accumulated in the population about the bandit. Some population states and bandits may exist in which other rules perform better (worse) than Never Switch. Classic decision theory (Savage [21]) demands that an individual determines an estimate (a subjective prior) of the likelihood of each bandit and population state and then selects a rule that maximizes subjective expected payoffs. Our boundedly rational approach does not utilize subjective priors. We assume that an individual wants to perform well in each situation, in particular, never worse than the 'baseline' rule Never Switch. Hence, individuals restrict attention to *improving* rules that sometimes yield higher, and never lower, (objective) expected payoff than Never Switch upon first encounter of the bandit in any entry state and any bandit (with action set A that yields payoffs in $[\alpha, \omega]$).

In our second approach to selecting behavioral rules (formalized in Section 5.2) a social planner determines a rule for common use that yields the best performance for the entire population. For a specific multi-armed bandit, a rule is called *payoff increasing* if it generates a population dynamic where average expected payoffs will weakly increase (i.e., not decrease) over time for each initial state. This concept is compatible with the evolutionary game theory literature (e.g. Weibull [26]) where similar conditions on population dynamics are postulated.

Our social planner limits attention to rules that are *payoff increasing in each multi-armed bandit* with action set A that yields payoffs in $[\alpha, \omega]$.

[5] Savage [21] distinguishes between objective and subjective (or personal) uncertainty.

[6] Risk neutrality is assumed for simplicity. More general risk preferences can be incorporated as follows: individuals observe payoffs, translate them into utilities and then apply their rule to the utilities.

Uncertainty regarding the payoff distribution of the bandit, or rare, unobservable changes in the payoff distributions during an individual's life time motivate this criterion. An explicit analytic justification for the social planner's objective can be found in an evolutionary model of Björnerstedt and Schlag [3] where rare mutations affect rules and payoff distributions.

A first result establishes that a behavioral rule is improving if and only if it is payoff increasing in each bandit. Hence, both the boundedly rational individual and the social planner will select among improving rules. In fact, when an individual receives a rule from the social planner her expected payoff calculated a priori to her entry will weakly increase over time.

Simple improving rules are easily found, e.g., the rule Never Switch and the self-explanatory rule *Always Switch*. Our first goal is to characterize the entire set of improving rules. A first lemma shows that improving rules are imitating, i.e., an individual using an improving rule changes actions only through imitating others. The main theorem (Theorem 1) completes the characterization. Thereby, an imitating rule is improving if and only if, when two individuals using different actions happen to sample each other, the difference in the probabilities of switching is *proportional* to the difference in their realized payoffs—the individual realizing the lower payoff being more likely to switch. This relationship between switching probabilities and realized payoffs results from the linear structure of taking expectations. There are many rules with this property, e.g., *Proportional Imitation Rules* as defined above. On the other hand, the rule, *Imitate if Better*[7], which only (and always) allows imitation of individuals with higher payoff than self is *not* improving. We also show that improving rules would perform just like Never Switch were the set of obtainable payoffs not bounded.

The severe restrictions on the switching behavior of improving rules simplifies selection among them dramatically. Under various criteria and for either bounded rationality or social planning we find the same (unique) rule to be optimal. This rule is a Proportional Imitation Rule with a specific proportionality constant that depends on the payoff interval $[\alpha, \omega]$ (see Theorem 2).

Next we make some predictions about a large population in which individuals use our optimal rule and sample randomly and independently. Here, the stochastic process governing the choices made in the population over time can be approximated in the short run by a discrete version of the replicator dynamic (Taylor [25]). In particular, for any initial state in which each action is present, with probability arbitrarily close to one, provided the population is sufficiently large, most individuals will be choosing an expected payoff maximizing action after a finite number of rounds.

[7] (Ellison and Fudenberg [9]; Malawski [14])

In a further section we consider a more general two population random matching scenario. In each round two types of individuals are matched to play a normal form game. Selection of a behavioral rule using generalizations of the previous concepts yields the same optimal rule. In a large population under random and independent sampling with each individual using the optimal rule, short run adjustment is again approximated by the discrete replicator dynamic.

The paper is organized as follows. In Sections 2 and 3 the basic payoff realization and sampling scenario are introduced. Feasible behavioral rules are presented in Section 4. Section 5 contains two alternative approaches to selecting a behavioral rule, each leading to the condition of improving. In Section 6 we present a first lemma on improving rules. In Section 7 this lemma is used to illustrate why Imitate if Better is not improving. Section 8 contains the main theorem completely characterizing improving rules. In Section 9 we select an optimal rule. Section 10 deals with the implications of optimal behavior for aggregate population adjustment. In Section 11 previous findings are generalized to a game playing scenario. Section 12 contains a discussion. The Appendix contains a corollary on improving rules.

2. THE PAYOFF REALIZATION SCENARIO

In the following three sections we describe a dynamic process of choosing actions, sampling and updating. First we establish how payoffs are realized. Let W be a finite population (or set) of N individuals, $N \geqslant 2$. In a sequence of rounds, each individual in the population must choose an action (or arm) from a finite set of actions A, $|A| \geqslant 2$. Choosing action i yields an uncertain payoff drawn from a given probability distribution P_i with *finite* support in $[\alpha, \omega]$, where α and ω, $\alpha < \omega$, are exogenous parameters. π_i denotes the expected payoff generated by choosing action i, i.e., $\pi_i = \sum_x xP_i(x)$, $i \in A$. Payoffs are realized independently of all other events. The tuple $\langle A, (P_i)_{i \in A} \rangle$, which specifies the set of actions together with a payoff distribution for each action, will be called a *multi-armed bandit* (or game against nature). $\mathcal{G}(A, [\alpha, \omega])$ denotes the set of all such multi-armed bandits.[8]

Let A, $[\alpha, \omega]$ and N be fixed throughout the rest of the paper. A *population state* $s \in A^W$ in a given round t is the description of the action which each individual chooses in round t. Let $m_i = m_i(s)$ denote the number of

[8] Alternatively, one might say that each arm can be one of an infinite number of types, the true type of an arm i being associated with a specific underlying payoff distribution P_i. The set of feasible types of arm i is then the set of probability distributions with finite support on $[\alpha, \omega]$. In our notation, a multi-armed bandit is the *realization* of a type for each arm, denoted by $\langle A, (P_i)_{i \in A} \rangle$, not the *collection* of feasible types of each arm, denoted by $\mathcal{G}(A, [\alpha, \omega])$.

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 is choosing, i.e., $p_i = m_i/N$ for $i \in A$. 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 $N \cdot p_i \in \mathbb{N} \cup \{0\}$. Given this notation, the *average expected payoff* in the population in state s, $\bar{\pi}(s)$, is given by $\bar{\pi}(s) = \sum_i p_i \pi_i$.

Individuals do not remain in the population W forever. Periodically a new individual appears who randomly replaces one of the individuals in the population; replacement occurs after a payoff realization, $1/N$ is the probability of replacing a given individual, the replaced individual exits the population. It will not be necessary for the analysis that follows to explicitly specify the process governing when new individuals appear. An individual's *entry state* is the population state of the round in which this individual enters the population.

3. INFORMATION ABOUT OTHERS

An entering individual learns the last choice and payoff realized by the individual she replaces.

Once in the population an individual receives information about the play of other individuals according to the following sampling scenario. After a round of payoff realization, each individual meets (or samples) one other 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. Sampling does not depend on realized payoffs nor on the population state and occurs independent of previous events. Who gets to sample whom with which probability is determined by the *sampling procedure*. Given $Z = \{f \in W^W : f(c) \neq c \ \forall c \in W\}$, let $f \in Z$ denote the event in which individual c samples individual $f(c)$, $c \in W$. A *sampling procedure* is an exogenously given distribution z over the events $f \in Z$, i.e., $z \in \Delta(Z)$.

For $c, d \in W$, $c \neq d$, let $c \rightsquigarrow d$ denote the event of c sampling d and let $\Pr(c \rightsquigarrow d)$ denote the probability of this event, i.e., $\Pr(c \rightsquigarrow d) = \sum_{f: f(c) = d} z(f)$. In the following we will restrict attention to *symmetric sampling procedures*, i.e., for any $c, d \in W$, $c \neq d$, the probability of c sampling d is the same as vice versa, i.e., $\Pr(c \rightsquigarrow d) = \Pr(d \rightsquigarrow c)$.

Symmetric sampling procedures can have a variety of different characteristics, e.g., regarding the way information is obtained. One may want to assume that individuals exchange information. In our setting this means

that individuals sample each other. Here, $c \rightsquigarrow d$ is the same event as $d \rightsquigarrow c$ for each $c, d \in W$, $c \neq d$. We also allow for settings in which individuals obtain information without necessarily revealing their own. Such a situation arises when individuals *sample independently*, i.e., when $\Pr(c \rightsquigarrow d \cap d \rightsquigarrow c) = \Pr(c \rightsquigarrow d) \cdot \Pr(d \rightsquigarrow c)$ for all $c, d \in W$, $c \neq d$.

Symmetric sampling procedures may differ according to the number of different samples an individual may obtain. E.g., a symmetric sampling procedure obtains from the following story. 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, $m \leqslant N/2$). In the extreme case, *random sampling*, each individual randomly samples (with equal probability) from the entire population, i.e., $\Pr(c \rightsquigarrow d) = 1/(N-1)$ for $c, d \in W$, $c \neq d$.

4. BEHAVIORAL RULES

A *behavioral rule* is the formal description of how an individual chooses her next action as a function of past experience. We focus on behavior of individuals entering after the very first round of the model. How the first N individuals choose their actions is not modelled. Choice will depend on (i) individual knowledge, (ii) information and memory, and (iii) tools available.

(i) An individual knows the action set A, the set of feasible payoffs $[\alpha, \omega]$ but not the underlying payoff distributions in the bandit. She knows the symmetric sampling procedure and the entry and exit mechanism.

(ii) An individual forgets all information she obtained prior to the previous round. She does not condition play on the current round number. In particular, in her first encounter of the bandit, she treats the previous choice and realized payoff of the individual she replaced as if it were her own.[9]

(iii) An individual has access to a randomizing device that generates independent events.

The extreme limitation posed by (ii) is a means to focus on simple behavior. Given the above assumptions, a *behavioral rule* F is characterized by a function

$$F: A \times [\alpha, \omega] \times A \times [\alpha, \omega] \to \Delta(A), \tag{1}$$

[9] Relaxing this assumption will have no effect on optimal behavior.

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.

Given a behavioral rule F and a multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$, let F_{ij}^k be the probability of playing action k after playing action i and sampling an individual using action j calculated a priori to realization of payoffs ($i, j, k \in A$), i.e.,

$$F_{ij}^k = \sum_{x,\, y} F(i, x, j, y)_k \, P_i(x) \, P_j(y). \tag{2}$$

$(F_{ij}^k)_{i,j,k \in A}$ are called *induced switching probabilities.*

One of the simplest behavioral rules, *Never Switch*, is the 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 rule, *Always Switch*, where $F(i, x, j, y)_j = 1$ for $i, j \in A$ and $x, y \in [\alpha, \omega]$. A more plausible rule seems to be to act according to *Imitate if Better* (Ellison and Fudenberg [9]; Malawski [14]), 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 $y \leqslant x$. 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 *Proportional Imitation Rule*, is the imitating rule F where there exists $\sigma \in (0, 1/(\omega - \alpha)]$ such that $F(i, x, j, y)_j = 0$ if $y \leqslant x$ and $F(i, x, j, y)_j = \sigma(y - x)$ if $y > x$, $i \neq j$ and $x, y \in [\alpha, \omega]$. The associated constant σ is called the *switching rate.*[10]

5. SELECTION OF RULES

Each individual must commit to a behavioral rule before she enters the population. The major part of our analysis is concerned with finding an optimal rule. We present two alternative scenarios (or approaches) for determining the notion of optimality.

5.1. *A Boundedly Rational Approach*

In the first scenario we consider boundedly rational individuals. Here, individuals are myopic and evaluate rules according to performance in their first encounter of the bandit. This performance depends on entry state and payoff distributions $(P_i)_{i \in A}$ of the commonly experienced bandit. It

[10] Switching behavior as displayed by Proportional Imitation Rule appears in papers by Cabrales [7] and Helbing [12], the former intuitively justifying such behavior through uniformly distibuted costs for switching actions.

also depends on realization of, 'objective' (*sensu* Savage [21]) uncertainty implicit in the model, i.e., point of entry, sample and payoff. Individuals are assumed to be risk neutral towards objective uncertainty. Thus, if an individual were to know entry state and payoff distributions of the bandit, comparing two given rules she would prefer the one yielding higher expected payoff in her first encounter. In the following, the term 'performance' refers to objective expected payoff in the first encounter.

Given the entry state and the multi-armed bandit encountered the selected behavioral rule might perform poorly and it might perform excellently. We assume that an individual prefers a rule that performs well whatever circumstances she enters into. This criterion does not make sense until we calibrate the measurement of performance. Following the rule Never Switch means to perform as well as the average individual performed in the previous round. Thus, the expected payoff of this rule will reflect the information accumulated in the population about the bandit. Never Switch will be our a baseline for analyzing the performance of a rule. Hence, "to perform well" will mean to perform at least as good as Never Switch. Rules that perform at least as good as Never Switch under any circumstances will be called improving, a condition to be formalized in the following.

Consider an individual that is about to enter in round t a population in state s^t. Remember that an entering individual adapts all attributes of the individual she replaces. Hence, the individual's expected payoff in round $t+1$, denoted by $E_{F,s^t}\pi'$, is given by

$$E_{F,s^t}\pi' = \frac{1}{N} \sum_{c \in W} \sum_{d \in W \setminus \{c\}} \Pr(c \rightsquigarrow d) \sum_{r \in A} F^r_{s^t(c)\, s^t(d)} \pi_r.$$

The expected payoff of Never Switch in round $t+1$ is equal to the average expected payoff in the population in round t, $\bar{\pi}(s^t)$. Let $EIP_F(s^t)$ denote the difference between the performance of F and Never Switch, i.e.,

$$EIP_F(s^t) = E_{F,s^t}\pi' - \bar{\pi}(s^t). \tag{3}$$

$EIP_F(s)$ is called the *expected improvement* under F in state s. The behavioral rule F is called *improving* (given A and $[\alpha,\omega]$) if $EIP_F(s) \geqslant 0$ for any state $s \in A^W$ and any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$.[11] The improving rule F is *degenerate* if $EIP_F(s) = 0$ for any state $s \in A^W$ and any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$.

[11] The concept of improving is very closely related to the concept of *absolute expediency* defined by Sarin [19] 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 improving rules.

5.2. *A Social Planer's Approach*

In this alternative scenario, a social planner determines the behavioral rule each individual follows. Equal treatment of the identical individuals calls for the social planner to prescribe the same rule to each individual. The aim of the planner is to select the rule that is best (to be specified) for society.

When each individual follows the same behavioral rule we obtain a *monomorphic population*. The *initial state* is the population state in the very first round of the model. Given initial state $s^1 \in A^W$, multi-armed bandit $b \in \mathcal{G}(A, [\alpha, \omega])$ and behavioral rule F, the monomorphic population induces a Markov process on A^W that describes the change of the population state over time. If s^t is the population state in round t then the expected proportion of individuals in round $t+1$ using action i, $E_F p_i'(s^t)$, calculated a priori to the payoff realizations in round t, is given by

$$E_F p_i'(s^t) = \frac{1}{N} \sum_{c \in W} \sum_{d \in W \setminus \{c\}} \Pr(c \rightsquigarrow d)\, F^i_{s^t(c)\, s^t(d)}. \tag{4}$$

$E_F \bar{\pi}'(s^t) = \sum_i E_F p_i'(s^t) \cdot \pi_i$ is the average expected payoff in the population in round $t+1$. Notice that

$$E_F \bar{\pi}'(s^t) = E_{F, s^t} \pi'. \tag{5}$$

The behavioral rule F is called *payoff increasing* in the bandit b if average expected payoff weakly increases over time for any initial state, i.e., $E_F \bar{\pi}'(s) \geqslant \bar{\pi}(s)$ for any $s \in A^W$.

The social planner finds the population already "in action" when he first prescribes a rule. Lack of information about bandit and current state or rare, unobservable, changes in payoff distributions make the planner prefer a rule that performs well in each situation. Here, the social planer selects among the rules that are payoff increasing in each bandit in $\mathcal{G}(A, [\alpha, \omega])$.

As in the bounded rational setting, as of yet a formal justification why a rule should be payoff increasing in each bandit is missing. The story of a social planner makes it easy to describe the payoff increasing condition. However, this condition also plays an important role without social planner when rules are under selection pressure. Consider a *large* population in which successful rules propagate. If success of a rule is determined by average payoff in a given state then a successful rule must be able to find the expected payoff maximizing action. Otherwise an alternative rule with a bias towards the action maximizing expected payoff will have a selective advantage. At the same time, two rules that are both able to learn which action is best among those present will eventually eliminate selection pressure between them and hence both survive. Consequently, in an

evolutionary setting in which bandits are subject to rare, arbitrary, and unobservable changes, it seems that only a rule that is payoff increasing in each bandit can be successful. Björnerstedt and Schlag [3] confirm this intuition in an evolutionary analysis of an infinite population facing our matching and sampling scenario.[12]

5.3. *Comparing Approaches*

Combining (3) and (5), we obtain that

Remark 1. A behavioral rule is improving if and only if it is payoff increasing in any multi-armed bandit belonging to $\mathscr{G}(A, [\alpha, \omega])$.

Boundedly rational individuals modelled in Section 5.1 restrict attention to performance in their first encounter. Thus, for each individual, performance can be calculated independently of rules used by others. How does an individual perceive her performance in later rounds if she receives her rule from a social planner? The social planner prescribes the same improving rule to each individual. Hence, realization of future states does not depend on which individuals are replaced by whom. In any round (and not only the round in which the individual enters) and for any state in this round the individual expects to be equally likely in the position of each of the individuals $c \in W$. Hence,

Remark 2. For any (improving) rule prescribed by the social planner, an individual's expected payoff calculated a priori to her entry weakly increases over time for any entry state and any bandit.

Given above, both approaches select among improving rules. Once we have characterized improving rules it will become clear which improving rule a boundedly rational individual or a social planer will regard as optimal.

6. A FIRST LEMMA

Clearly, the rule Never Switch is improving. Matching being symmetric causes Always Switch to be improving too. Either rule is not a very good candidate for optimal behavior since they both leave average expected payoff constant over time, i.e., each of them is a degenerate improving rule.

The following preliminary result characterizes improving rules in a way that does not depend on the population state. Hereby, a behavioral rule F

[12] Typically an individual that selects a behavioral rule according to a subjective prior will perform worse. In this sense, this is an instance (similar to Robson [16]) in which successful behavior in an evolutionary setting does not comply with the fundaments of rational decision making.

is improving if and only if it is an imitating rule that satisfies the following condition. Consider two individuals choosing different actions, using the same rule F, that sample each other. Then, before observing each other's payoff, the individual with the lower expected payoff is more likely to switch actions.

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

$$(F^j_{ij} - F^i_{ji})(\pi_j - \pi_i) \geqslant 0. \tag{6}$$

The proof of the imitation property is quite intuitive. An individual does not switch to an action she did not observe since she fears this action achieves the lowest and all other actions the highest expected payoff. 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. Calculating expected improvement for imitating rules yields

$$EIP_F(s) = \frac{1}{N} \sum_{c \in W} \sum_{d \in W \setminus \{c\}} \Pr(c \rightsquigarrow d)\, F^{s(d)}_{s(c)\, s(d)} [\pi_{s(d)} - \pi_{s(c)}].$$

Using the fact that the sampling procedure is symmetric we obtain

$$EIP_F(s) = \frac{1}{N} \sum_{i<j} \left[\sum_{\substack{c:\, s(c)=i \\ d:\, s(d)=j}} \Pr(c \rightsquigarrow d) \right] (F^j_{ij} - F^i_{ji})(\pi_j - \pi_i), \tag{7}$$

which completes the proof of the "if" statement.

Next we will 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 $r \in A \setminus \{i, j\}$ be such that $F(i, x, j, y)_r > 0$. Consider a multi-armed bandit belonging to $\mathcal{G}(A, [\alpha, \omega])$ with $P_i(x) = P_i(y) = P_i(\omega) = \frac{1}{3}$, $P_j \equiv P_i$ and $P_k(\alpha) = 1$ for all $k \in A \setminus \{i, j\}$. It follows that $\pi_i = \pi_j > \pi_k$. Choose $c, d \in W$ such that $\Pr(c \rightsquigarrow d) > 0$ and consider a population state s such that $s(c) = i$, $s(d) = j$ and $m_i + m_j = N$. Then $F(i, x, j, y)_r > 0$ implies $EIP(s) < 0$ which contradicts the fact that F is improving.

Finally, we will show that an imitating rule F that violates (6) for some $i \neq j$ and some multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ is not improving. Choose again $c, d \in W$ such that $\Pr(c \rightsquigarrow d) > 0$ and consider a population

state s such that $s(c)=i, s(d)=j$ and $m_i+m_j=N$. Since $(F^j_{ij}-F^i_{ji})(\pi_j-\pi_i)<0$, following (7), $EIP_F(s)<0$ which implies that F is not improving. ∎

In the social planer's approach we restricted attention to rules that are payoff increasing in any bandit. One might wish to impose the following weaker condition. Assume that only those actions are expected to increase that perform at least as well as some other action present, i.e.,

$$\forall s \in A^W,\ i \in A : E_F p'_i(s) \geqslant p_i(s) \Rightarrow \exists c \in W : \pi_i \geqslant \pi_{s(c)}. \tag{8}$$

Our condition (8) is weaker than most necessary conditions postulated in evolutionary game theory for reasonable dynamics in infinite populations.[13] Never-the-less it is sufficient to drive our results.

Remark 3. A behavioral rule induces a monomorphic population dynamic that satisfies (8) in all multi-armed bandits contained in $\mathcal{G}(A, [\alpha, \omega])$ if and only if it is improving.

The statement in Remark 3 is easily verified using the proof of the imitation property in Lemma 1 and the fact that (8) is equivalent to payoff increasing when $|A|=2$.

7. 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 (*sensu* Ellison and Fudenberg [9]) shocks. Consider a multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ with the following properties. There is a probability distribution Q with finite support and mean 0 such that $P_i(x)=Q(x-\pi_i)$ for each $i \in A$ and $x \in [\alpha, \omega]$. Throughout this section, let F denote the rule Imitate if Better. Then

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

and hence, $F^j_{ij}-F^i_{ji}\geqslant 0$ when $\pi_j \geqslant \pi_i$. With (7) it follows that the expected improvement of Imitate if Better is non negative in such multi-armed bandits.

However, we will see that Imitate if Better generates negative expected improvement in some extremely simple multi-armed bandits; it can not distinguish between lucky and certain (or highly probable) payoffs. Let

[13] (E.g., compatibility, also known as payoff monotonicity, and weak compatibility, Friedman [10]; payoff positivity, Weibull [26])

$x \in (\alpha, (\alpha+\omega)/2)$. Consider a multi-armed bandit in which $P_1(x)=1$, $P_2(\alpha)=\lambda$ and $P_2(\omega)=1-\lambda$ for some $0<\lambda<1$. It follows that

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

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

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

Consequently, when $\frac{1}{2}<\lambda<(\omega-x)/(\omega-\alpha)$ then (6) is violated and hence Imitate if Better is not improving.

8. 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 that an individual is not able to condition her behavior on expected payoffs but must base her decision on realized payoffs. The following theorem contains the central result of this paper, 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 $\sigma_{ij}=\sigma_{ji} \in [0, 1/(\omega-\alpha)]$ such that*

$$F(i, x, j, y)_j - F(j, y, i, x)_i = \sigma_{ij}(y-x) \qquad \text{for all} \quad x, y \in [\alpha, \omega]. \quad (9)$$

From (9) we see immediately that Imitate if Better is not improving, confirming our findings from Section 7.

Proof. We will first show that conditions (i) and (ii) are sufficient. Let F be an imitating behavioral rule that satisfies condition (ii). (2) and (9) imply

$$F^j_{ij} - F^i_{ji} = \sigma_{ij}(\pi_j - \pi_i). \quad (10)$$

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 $g_{ij}(x, y) := F(i, x, j, y)_j - F(j, y, i, x)_i$ for $x, y \in [\alpha, \omega]$. First we will show that

$$\frac{g_{ij}(x, u)}{u - x} = \frac{g_{ij}(x, z)}{z - x} \qquad \forall u < x < z. \tag{11}$$

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

$$F_{ij}^j - F_{ji}^i = \lambda g_{ij}(x, u) + (1 - \lambda)\, g_{ij}(x, z) \geqslant 0 \qquad \text{if} \quad \lambda < \lambda^* \text{ and} \tag{12}$$

$$\lambda g_{ij}(xc, u) + (1 - \lambda)\, g_{ij}(x, z) \leqslant 0 \qquad \text{if} \quad \lambda > \lambda^* \tag{13}$$

Therefore, $\lambda^* g_{ij}(x, u) + (1 - \lambda^*)\, g_{ij}(x, z) = 0$, which, after simplification, shows that (11) is true.

Since the left hand side in (11) is independent of z, so is the right hand side. Given $x \in (\alpha, \omega)$, let $\sigma_{ij}(x) = (g_{ij}(x, z))/(z - x)$ for some $z > x$. Following (11), $g_{ij}(x, u) = \sigma_{ij}(x) \cdot (u - x)$ for all $u < x$ and $g_{ij}(x, z) = \sigma_{ij}(x) \cdot (z - x)$ for all $z > x$. Hence, for all $x, y \in (\alpha, \omega)$, $x \neq y$,

$$g_{ij}(x, y) = F(i, x, j, y)_j - F(j, y, i, x)_i = \sigma_{ij}(x)(y - x), \tag{14}$$

or equivalently,

$$F(j, y, i, x)_i - F(i, x, j, y)_j = \sigma_{ij}(x)(x - y). \tag{15}$$

Exchanging the variables i and j and the variables x and y in (14) implies

$$F(j, y, i, x)_i - F(i, x, j, y)_j = \sigma_{ji}(y)(x - y). \tag{16}$$

From (15) and (16) it follows that $\sigma_{ij} = \sigma_{ji}$ is a constant. Setting $\lambda = 0$ in (12) it follows that this constant is non-negative. Hence, we have shown (9) for all $x, y \in (\alpha, \omega)$, $x \neq y$.

Looking back at the above proof we see that the explicit values of α and ω did not enter the argument. Hence, (9) holds for all $x, y \in [\alpha, \omega]$, $x \neq y$.

Assume that $g_{ij}(x, x) > 0$ for some $x \in [\alpha, \omega)$. Consider a multi-armed bandit where $P_i(x) = 1 - \lambda$, $P_i(\omega) = \lambda$ and $P_j(x) = 1$, $0 < \lambda < 1$. Then $\pi_j < \pi_i$ and $F_{ij}^j - F_{ji}^i = (1 - \lambda)\, g_{ij}(x, x) + \lambda \sigma_{ij} \cdot (x - \omega) > 0$ for λ sufficiently small contradicts the fact that F is improving. Similarly, $g_{ij}(x, x) < 0$ leads to a contradiction. Hence, $g_{ij}(x, x) = 0$ for all $x \in [\alpha, \omega)$. The proof of $g(\omega, \omega) = 0$ is analogue. This completes the proof of (9).

Finally, $\sigma_{ij}(\omega - \alpha) = g_{ij}(\alpha, \omega) \leqslant F_{ij}^j \leqslant 1$ implies $\sigma_{ij} \leqslant 1/(\omega - \alpha)$. ∎

From (7) and (10) we obtain that

COROLLARY 1. *An improving rule is degenerate if and only if* $\sigma_{ij}=0$ *for all* $i, j\in A$, $i\neq j$.[14]

Thus, non degenerate improving rules induce stochastic behavior. Moreover, since σ_{ij} is bounded above by $1/(\omega-\alpha)$ (Theorem 1), all improving rules would be degenerate if payoffs were not contained in a bounded interval.

In the appendix we include a corollary that gives a more precise characterization of improving rules.

9. SELECTING AMONG IMPROVING RULES

We now proceed with our search for an optimal behavioral rule. First we will show that there are improving rules that perform better than all others. A behavioral rule *F dominates the improving rules* (or short, is a *dominant rule*) if it *always* generates weakly higher expected improvement than any other improving rule, i.e., for any improving rule F', state s and multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$, $EIP_F(s)\geqslant EIP_{F'}(s)$ holds. With (3) and (5) it follows that dominant rules are also the rules that maximize the increase in average expected payoffs of a monomorphic population in any given state and bandit among the set of improving rules. Hence, both the boundedly rational individual and the social planer will select a dominant rule if such a rule exists.

Following (7) and (10),

$$EIP_F(s)=\left[\frac{1}{N}\sum_{i<j}\sum_{\substack{c:\,s(c)=i\\ d:\,s(d)=j}}\Pr(c\rightsquigarrow d)\right]\sigma_{ij}(\pi_j-\pi_i)^2. \tag{17}$$

Given (17), the expected improvement of an improving rule depends only 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* $i\neq j$, $\sigma_{ij}=1/(\omega-\alpha)$.

Next we demonstrate three unique properties of the Proportional Imitation Rule with switching rate $1/(\omega-\alpha)$ (defined in Section 4, denoted

[14] In this context, notice that a rule is absolutely expedient (Footnote 11) if and only if it is improving with $\sigma_{ij}>0$ for all $i, j\in A$, $i\neq j$.

by F^p). These properties will cause us to select it as the unique optimal rule for our model.

THEOREM 2. *F^p is the unique dominant rule that satisfies any one of the following properties.*

(i) *It never prescribes to imitate an action that achieved a lower payoff.*

(ii) *In each multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ and state it minimizes the probability of switching among dominant rules.*

(iii) *For each multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ and current state it minimizes the variance of the average payoff in a monomorphic population in the next round among dominant rules.*

Proof of Theorem 2. Statements (i) and (ii) follow easily from Corollary 2 stated in the appendix since F^p 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. ∎

Following (i) in Theorem 2, it can be argued that F^p is the best dominant rule under deterministic payoffs; realized payoffs never decrease when actions yield certain payoffs. (ii) implies that F^p is the dominant rule that changes actions the least number of times. Given (iii), among improving rules, F^p maximizes increase in average expected payoffs using minimal variance. This leads us to conjecture that F^p maximizes the probability that average payoffs realized in a monomorphic population increase over time.

The Proportional Imitation Rule with switching rate $1/(\omega - \alpha)$ is improving. 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 dominant rules (Theorem 2) lead us to argue that it is the *optimal rule* in either selection approach. Notice that the optimal rule does not depend on the size of the population N.

Remark 4. One should mention that there is a dominant rule that utilizes less information than the dominant Proportional Imitation Rule. The dominant *Proportional Reviewing Rule* is the imitating rule F where $F(i, x, j, y)_j = (\omega - x)/(\omega - \alpha)$ for $i \neq j$, $i, j \in A$ and $x, y \in [\alpha, \omega]$.[15] It can be easily shown (see Schlag [22] 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.

[15] Björnerstedt and Weibull ([4]) and Gale *et al.* ([11]) both use a variant of this rule in their model, the later interpret it on the basis of random aspiration levels.

10. POPULATION DYNAMICS

In this section we investigate how much individuals learn about the bandit when each of them uses the optimal rule. For this we analyze the stochastic process governing the evolution of the population state over time. Attention is restricted to random and independent sampling in large populations.

First we derive a 'law of large numbers' type of result for a monomorphic population based on an arbitrary behavioral rule. We identify a state $s \in A^W$ with the associated distribution, $p \in \Delta^N(A)$, of actions chosen. 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, $t = 2, 3, \ldots$ Let $\|\cdot\|$ denote the supremums norm. For large populations, our result shows that the stochastic adjustment can be approximated in the short run by the deterministic adjustment that would take place if the size of the population were infinite.

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

$$p_i(t+1) = \sum_{j,\, r} p_j(t)\, p_r(t)\, F^i_{jr}, \qquad t \in \mathbb{N}. \tag{18}$$

Proof. We will first prove the statement for $T = 2$. Fix $i \in A$ and $\tilde{p} \in \Delta^N(A)$. For $c \in W$ let $w_i(c)$ be the random variable such that $w_i(c) = 1$ if individual c uses action i in round two, otherwise $w_i(c) = 0$. Then

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

and $p^N_i(2) = (1/N) \sum_{c \in W} w_i(c)$. Since $w_i(c)$ and $w_i(d)$ are independent variables for $c \neq d$ and $VAR(w_i(c)) \leqslant 1$ it follows that $VAR(p^N_i(2)) \leqslant 1/N$. Applying Tschebysheff's inequality we obtain that the event $\{|p^N_i(2) - E[p^N_i(2)]| > \delta/2\}$ occurs with a probability of less than $4/N\delta^2$. Given

$$E[p^N_i(2)] = \frac{N}{N-1} \sum_{j,\, r} \tilde{p}_j \tilde{p}_r F^i_{jr} - \frac{1}{N-1} \sum_j \tilde{p}_j F^i_{jj}, \tag{19}$$

there exists N_0 such that $4/N\delta^2 < \varepsilon$ and $|E[p^N_i(2)] - \sum_{j,\, r} \tilde{p}_j \tilde{p}_r F^i_{jr}| < \delta/2$ for $N > N_0$. Then $\{|p^N_i(2) - \sum_{j,\, r} \tilde{p}_j \tilde{p}_r F^i_{jr}| > \delta\}$ occurs with probability less

than ε when $N > N_0$. Since N_0 can be 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 $\varepsilon > 0$ be given. Let $f: \Delta(A) \to \Delta(A)$ be defined by $f(p)_i = \sum_{j,r} p_j p_r F^i_{jr}$, $i \in A$. 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 $\beta \in (0, \delta/2)$ such that $\|f(w) - f(w')\| < \delta/2$ if $\|w - w'\| < \beta$. Let μ be such that $(1-\mu)^2 = 1 - \varepsilon$. Following the proof for $T = 2$ there exists N_0 such that for $N > N_0$ and $\tilde{p} \in \Delta^N(A)$, $\Pr(\|p^N(2, \tilde{p}) - f(\tilde{p})\| < \beta) > 1 - \mu$. For $N > N_0$ it follows that

$$
\begin{aligned}
&\Pr(\|p^N(3, \tilde{p}) - f(f(\tilde{p}))\| \leqslant \delta) \\
&\quad = \sum_{w \in \Delta^N(A)} \Pr(\|p^N(2, w) - f(f(\tilde{p}))\| \leqslant \delta) \cdot \Pr(p^N(2, \tilde{p}) = w) \\
&\quad \geqslant \sum_{w \in \Delta^N(A) : \|w - f(\tilde{p})\| < \beta} \Pr\left(\|p^N(2, w) - f(w)\| \leqslant \frac{\delta}{2}\right) \cdot \Pr(p^N(2, \tilde{p}) = w) \\
&\quad \geqslant (1-\mu)^2 = 1 - \varepsilon,
\end{aligned}
$$

which completes the proof for $T = 3$. The proof for $T > 3$ follows similarly using induction. ∎

Theorem 3 makes a statement about the short run adjustment of large populations. First the time horizon and precision of the approximation is set, then we choose the population size to be sufficiently large. Why is it necessary to keep the time horizon fixed? For any given population size, long run behavior can differ quite dramatically from the behavior of (18).[16] The following is easily verified.

Remark 5. Consider a monomorphic population of size N, based on a non degenerate improving rule, facing a two-armed bandit, i.e., $A = \{1, 2\}$. Typically, $F^2_{12} > 0$ and $F^1_{21} > 0$.[17] Then for any initial interior state (i.e., $0 < p^N_1(1) < 1$), eventually each individual will be playing the same action. There is a positive probability that all individuals will be playing the worse action after a finite number of rounds. This can not happen in an infinite population as we will see below.

Given Theorem 3, understanding adjustment of the infinite population helps understand short run adjustment of a large but finite population. An infinite monomorphic population induces a deterministic process $(p(t))_{t \in \mathbb{N}}$

[16] For further reading, see (Boylan [6]) and Binmore *et al.* [2]).

[17] i.e., unless $\mathrm{Supp}(1) \cap \mathrm{Supp}(2) = \varnothing$ where $\mathrm{Supp}(i) = [\min\{y: P_i(y) > 0\}, \max\{y: P_i(y) > 0\}]$.

that satisfies (18). If the underlying rule F is improving then, using (10), (18) simplifies to

$$p_i(t+1) = p_i(t) + p_i(t) \sum_{j \in A} \sigma_{ij} \cdot p_j(t) \cdot (\pi_i - \pi_j). \tag{20}$$

Consequently, if F is improving with underlying $\sigma_{ij} > 0$ for all $i \neq j$, then in the long run all individuals in an infinite monomorphic population will choose actions achieving maximal expected payoff among those that were initially present, i.e., $\lim_{t \to \infty} p_i(t) = 0$ for $i \notin \arg\max_{j \in A} \{\pi_j, p_j(1) > 0\}$. It is easy to show that the converse of this statement is also true (use Lemma 1 and (18)). In particular, if all individuals in an infinite population use Imitate if Better then eventually they will all be choosing the inefficient action in the bandit from Section 7 if $\frac{1}{2} < \lambda < (\omega - x)/(\omega - \alpha)$ and $p_1(1) \in (0, 1)$.

If F is a dominant improving rule (e.g., the dominant Proportional Imitation Rule), then

$$p_i(t+1) = p_i(t) + \frac{1}{\omega - \alpha} [\pi_i - \bar{\pi}(p(t))] \cdot p_i(t), \tag{21}$$

where $\bar{\pi}(p) = \sum_i \pi_i \cdot p_i$ is the average payoff instate $p \in \Delta(A)$. Hence, if each individual uses the optimal rule then dynamic adjustment of a large but finite population is approximated in the short run by (21)—a discrete version of the replicator dynamic (Taylor [25]) applied to multi-armed bandits.

This leads to the following result about what typically happens in a large population of individuals using the optimal rule (compare to Remark 5). Loosely speaking, it is highly probable that most individuals will choose the best action after some finite time provided all actions are initially present.

Remark 6. Consider a finite population of individuals, using the dominant Proportional Imitation Rule, facing a given bandit in $\mathcal{G}(A, [\alpha, \omega])$. Let $M = \arg\max_{i \in A} \{\pi_i\}$. Then for any $\gamma \in (0, 1/|A|)$, $\delta > 0$ and $\varepsilon > 0$ there exits $T, N_0 \in \mathbb{N}$ such that the event $\{|1 - \sum_{i \in M} p_i^N(T)| > \delta\}$ occurs with probability less than ε given that $N > N_0$ and $p_i^N(1) > \gamma$ for all $i \in A$.

This statement follows directly from Theorem 3 and (21).

11. A GAME PLAYING SETTING

Above we derived optimal behavioral rules for stationary multi-armed bandits. In the following we extend our approach to the classic evolutionary game theoretic model of interacting individuals. Here, individuals are

repeatedly randomly matched to play a one shot game. This means that individuals repeatedly face a non-stationary multi-armed bandit where changes in payoff distributions result from changes in play of matched opponents. In order to simplify presentation we restrict attention to two person games. More general results for games with any given finite number of players are easily derived.

Consider two finite, disjoint populations W_1 and W_2, each of size N, referred to as *population one* and *two*. In a sequence of rounds each individual must choose an action and is then matched with 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 receives an uncertain payoff drawn from the probability distribution P_{ij}^k, $k=1, 2$. Payoffs are realized independently. Associating player i to being an individual in population i, the tuple $\langle A_1, A_2, (P_{ij}^1)_{\substack{i \in A_1 \\ j \in A_2}}, (P_{ij}^2)_{\substack{i \in A_1 \\ j \in A_2}} \rangle$ defines an *asymmetric two player normal form game*. We restrict attention to the class of asymmetric two player normal form games, $\mathcal{G}(A_1, A_2, [\alpha_1, \omega_1], [\alpha_2, \omega_2])$, in which P_{ij}^k has finite support in $[\alpha_k, \omega_k]$, $i \in A_1$, $j \in A_2$ and $k=1, 2$; $\alpha_1 < \omega_1$ and $\alpha_2 < \omega_2$ are given. For a given asymmetric game, let $\pi_1(\cdot, \cdot)$ and $\pi_2(\cdot, \cdot)$ 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., $\pi_k(i, j) = \sum_{\{x \in [\alpha_k, \omega_k] : P_{ij}(x) > 0\}} x P_{ij}^k(x)$, $k=1, 2$.

In each round, the *population state* (s_1, s_2) is an element of $(A_1)^{W_1} \times (A_2)^{W_2}$. Individuals are randomly matched in pairs, each individual being equally likely to be matched with each individual of the opposite population. Given the population state s, let $p(s) \in \Delta^N(A_1)$ be the vector of proportions of each action chosen in population one. Similarly let $q(s) \in \Delta^N(A_2)$ be the corresponding expression for population two. Then $\pi_1(i, q(s))$ specifies the expected payoff of an individual in population one using action $i \in A_1$ and $\pi_1(p(s), q(s))$ specifies the average expected payoff in population one in this state. For a given current state, each individual in population one is facing a multi-armed bandit $\langle A_1, (P_i')_{i \in A} \rangle$ in $\mathcal{G}(A_1, [\alpha_1, \omega_1])$ where $P_i'(x) = \sum_{j \in A} q_j(s) \cdot P_{ij}^1(x)$ for $x \in [\alpha_1, \omega_1]$.

Sampling occurs within each population according to a sampling procedure as described in Section 3.

A *behavioral rule* F for an individual in population k $(k=1, 2)$ is a function

$$F: A_k \times [\alpha_k, \omega_k] \times A_k \times [\alpha_k, \omega_k] \to \Delta(A_k).$$

Switching probabilities now depend on the population state. For a given behavioral rule F of an individual in population one, the induced switching

probabilities $(F^i_{jr}(s))_{i,\,j,\,r\in A_1}$ in state $s=(s_1, s_2)\in (A_1)^{W_1}\times (A_2)^{W_2}$ are given by

$$F^i_{jr}(s)=\sum_u \frac{n_u(n_u-1)}{N(N-1)}F(j, \pi_1(j, u), r, \pi_1(r, u))_i$$

$$+\sum_{u\neq v}\frac{n_u n_v}{N(N-1)}F(j, \pi_1(j, u), r, \pi_1(r, v))_i,$$

where $n_k=|\{c\in W_2 : s_2(c)=k\}|$ for $k\in A_2$ $(i, j, r\in A_1)$.

11.1. *Optimal Behavior*

Which behavioral rule should an individual entering into population one use? Consider a boundedly rational individual. In any given state the game appears as a multi-armed bandit. However, in contrast to the multi-armed bandit setting, underlying payoff distributions are no longer stationary. The best choice of an action in the next round depends on how opponents' adjust. We assume that an individual does not anticipate how the play of her opponents changes. Instead, she evaluates performance in her first encounter according to the play of population two in her entry state. Thus, the individual acts as if she were going to face a stationary bandit. Here, as in the multi-armed bandit setting, the Proportional Imitation Rule with switching rate $1/(\omega_1-\alpha_1)$ is the optimal rule.

Consider now a social planer selecting individual behavior, prescribing the same behavior to individuals belonging to the same population. If each individual in population one is using the rule F and s is the population state in round t then the expected proportion of individuals choosing action $i\in A_1$ in round $t+1$, denoted by $E_F p'_i(s)$, is given by

$$E_F p'_i(s)=\frac{1}{N}\sum_{c,\,d\in W_1,\,c\neq d}\Pr(c\rightsquigarrow d)\cdot F^i_{s_1(c)\,s_1(d)}(s), \qquad i\in A_1. \tag{22}$$

We say that F is *expected to induce a weak compatible dynamic in population one* if for each round and state, average expected play in the next round is a better reply to the state of the previous round, i.e., if

$$\sum_{i\in A_1}\pi_1(i, q(s))\cdot E_F p'_i(s)-\pi_1(p(s), q(s))\geqslant 0 \tag{23}$$

holds for all states $s\in (A_1)^{W_1}\times (A_2)^{W_2}$.[18] (23) replaces the 'payoff increasing' condition in Section 5.2.

[18] Definition adapted from the concept of *weak compatibility* for infinite populations (Friedman [10]).

The social planer chooses a rule for individuals in population one that is expected to induce a weak compatible dynamic (in population one) in each asymmetric game in $\mathcal{G}(A_1, A_2, [\alpha_1, \omega_1], [\alpha_2, \omega_2])$. These are precisely the rules that are improving for bandits in $\mathcal{G}(A_1, [\alpha_1, \omega_1])$. Further selection as in the multi-armed bandit setting (maximize left hand side in (23) with minimal variance) reveals the Proportional Imitation Rule with switching rate $1/(\omega_1 - \alpha_1)$ as the unique optimal rule. Symmetric arguments apply to population two.

11.2. *Population Dynamics in Games*

Assume that each individual uses the optimal Proportional Imitation Rule for her population. How does the population state evolve under random and independent sampling? Using the same law of large numbers type of argument as in Theorem 3[19] behavior of a large but finite population is approximated in the short run by the deterministic dynamic $(p^t, q^t)_{t=1,2,3,\ldots}$ that satisfies

$$
\begin{aligned}
p_i^{t+1} &= p_i^t + \frac{1}{\omega_1 - \alpha_1} [\pi_1(i, q^t) - \pi_1(p^t, q^t)] \cdot p_i^t, \qquad i \in A_1, \\
q_j^{t+1} &= q_j^t + \frac{1}{\omega_2 - \alpha_2} [\pi_2(p^t, j) - \pi_2(p^t, q^t)] \cdot q_j^t, \qquad j \in A_2, \quad t \in \mathbb{N}.
\end{aligned} \tag{24}
$$

Notice that (24), the two population analogue of (21), is a discrete version of the replicator dynamic defined by Taylor [25].

12. DISCUSSION

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

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*. For this discussion, let *optimal population adjustment* refer to an infinite population in which each individual uses the optimal rule.

[19] Small adjustments in the proof need to be made (see Schlag [22]) since switching probabilities are no longer (completely) independent due to the fact that individuals are matched in pairs.

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, individual information is extremely limited—an individual only observes the performance of *one* other individual between rounds. The Proportional Imitation Rule is argued to be the unique optimal rule, optimal population adjustment follows 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 adaptive rules that lead to the replicator dynamic (Björnerstedt and Weibull [4]; Cabrales [7]; Gale *et al.* [11], Helbing [12]), however they did not choose to analytically justify individual behavior. Axiomatizations of learning rules in slightly different contexts have also lead to the replicator dynamic (Easley and Rustichini [8]; Sarin [19], in combination with the paper by Börgers and Sarin [5]). However, their basic approach differs fundamentally from ours—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 surprising. In a recent investigation we expand our model and assume that an individual samples *two* individuals between rounds (Schlag [23]). Here dominant rules no longer exist. However, we find a simple, optimal, rule (a modification of the Proportional Imitation Rule) that is best at performing better than any improving rule based on a single sample. Optimal population adjustment is described by an aggregate monotone dynamic (Samuelson and Zhang [20]).

When individuals in our setting have perfect information, playing a best response would be the unique dominant rule. Optimal population adjustment becomes 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 (Section 11) optimal adjustment follows a version of the best response dynamic (Matsui [15]). Comparing this result to ours, we see that the replicator dynamic and the best response dynamic compromise extreme points in the class of adjustment dynamics based on individually optimal myopic behavior.

An intermediate case regarding informational assumptions is a scenario where individuals observe expected payoffs of action used and action sampled. Although this assumption is difficult to motivate it is quite popular in the literature (e.g., see Björnerstedt and Weibull [4]; Hofbauer [13]).[20]

[20] Repeated (i.e., finitely many) pulls of the same arm between samples does not generate this situation since unlucky draws will distort information.

Here, Imitate if Better is the unique dominant rule and hence optimal. Optimal population adjustment in the multi-armed bandit setting leads to the state in which each individual chooses the best action among those initially present. We show that the dominant Proportional Imitation Rule has the same property under much less severe informational requirements.

Two alternative justifications for why individuals may choose to imitate under similar circumstances should be mentioned. Rogers [17] presents an example of a changing environment in which individuals imitate in order to evade search costs. The evolutionarily stable proportions of individual learning (i.e., individuals incur a cost and learn the currently best action) and social learning (i.e., individuals imitate without observing payoffs) are computed. Banerjee [1] presents a model in which rational individuals imitate for hope that the observed individual has more information.

Finally, we want to mention Malawski's [14] experiments in the game playing setting of Section 11. In this investigation an imitation hypothesis is refuted due to the high proportion of individuals switching to actions other than the one previously observed (over 30%). The data is partially explained with aspiration level learning, a model that entirely ignores information obtained through sampling. 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 from the experiment of Malawski and the conduction of new experiments has therefore been planned.

APPENDIX A: A COROLLARY ON IMPROVING RULES

COROLLARY 2. *Condition* (ii) *in Theorem* 1 *holds if and only if the following conditions 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* $x, y \in [\alpha, \omega]$,

$$-\min\{x, y\} \leqslant g_{ij}(x, y) \leqslant -\max\{x, y\} + \frac{1}{\sigma_{ij}},$$

$$F(i, x, j, y)_j = \sigma_{ij} \cdot (y + g_{ij}(x, y)) \quad \textit{and}$$

$$F(j, y, i, x)_i = \sigma_{ij} \cdot (x + g_{ij}(x, y)).$$

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 $g_{ij}(x, y) = (1/\sigma_{ij}) F(i, x, j, y)_j - y \quad (x, y \in [\alpha, \omega])$. It follows that $-y \leqslant g_{ij}(x, y) \leqslant -y + (1/\sigma_{ij})$ 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 $-x \leqslant g_{ij}(x, y) \leqslant -x + (1/\sigma_{ij})$ which completes the proof of condition (ii′). ∎

REFERENCES

1. A. V. Banerjee, A simple model of herd behavior, *Quart. J. Econ.* **107** (1992), 797–818.
2. K. G. Binmore, L. Samuelson, and R. Vaughan, Musical chairs: Modeling noisy evolution, *Games Econ. Beh.* **11** (1995), 1–35.
3. J. Björnerstedt and K. H. Schlag, "On The Evolution of Imitative Behavior," Discussion Paper No. **B-378**, University of Bonn, 1996.
4. J. Björnerstedt and J. Weibull, Nash equilibrium and evolution by imitation, "The Rational Foundations of Economic Behaviour," Proc. IEA Conference, (K. Arrow *et al.*, Eds.), pp. 155–171, MacMillan, London, 1996.
5. T. Börgers and R. Sarin, "Learning Through Reinforcement and Replicator Dynamics," Discussion Paper No. **93-19**, University College of London, 1993.
6. R. T. Boylan, Laws of large numbers for dynamical systems with randomly matched individuals, *J. Econ. Theory* **57** (1992), 473–504.
7. A. Cabrales, "Stochastic Replicator Dynamics," mimeo, University of California, San Diego, 1993.
8. D. Easley and A. Rustichini, "Choice Without Beliefs," mimeo, Cornell University and C.O.R.E., 1995.
9. G. Ellison and D. Fudenberg, Word-of-mouth communication and social learning, *Quart. J. Econ.* **440** (1995), 93–125.
10. D. Friedman, Evolutionary games in economics, *Econometrica* **59** (1991), 637–666.
11. J. Gale, K. G. Binmore, and L. Samuelson, Learning to be imperfect: The ultimatum game, *Games Econ. Beh.* **8** (1995), 56–90.
12. D. Helbing, Interrelations between stochastic equations for systems with pair interactions, *Physica A* **181** (1992), 29–52.
13. J. Hofbauer, "Imitation Dynamics for Games," University of Vienna, mimeo, 1995.
14. M. Malawski, "Some Learning Processes in Population Games," Inaugural-Dissertation, University of Bonn, 1989.
15. A. Matsui, Best response dynamics and socially stable strategies, *J. Econ. Theory* **57** (1992), 343–362.
16. A. Robson, A biological basis for expected and non-expected utility, *J. Econ. Theory* **9** (1996), 397–424.
17. A. Rogers, Does biology constrain culture? *Amer. Anthropol.* **90** (1989), 819–831.
18. M. Rothschild, A two-armed bandit theory of market pricing, *J. Econ. Theory* **9** (1974), 185–202.
19. R. Sarin, "An Axiomatization of the Cross Learning Dynamic," mimeo, University of California, San Diego, 1993.
20. L. Samuelson and J. Zhang, Evolutionary stability in asymmetric games, *J. Econ. Theory* **57** (1992), 363–391.
21. L. J. Savage, "The Foundations of Statistics," Wiley, New York, 1954.
22. K. H. Schlag, "Why Imitate, and if so, How? Exploring a Model of Social Evolution," Discussion Paper **B-296**, University of Bonn, 1994.

23. K. H. Schlag, "Which One Should I Imitate?," Discussion Paper **B-365**, University of Bonn, 1996.
24. R. Schmalensee, Alternative models of bandit selection, *J. Econ. Theory* **10** (1975), 333–342.
25. P. Taylor, Evolutionarily stable strategies with two types of players, *J. Applied Prob.* **16** (1979), 76–83.
26. J. Weibull, "Evolutionary Game Theory," MIT Press, Cambridge, MA, 1995.