Which one should I imitate?[1]

Karl H. Schlag[2]

Projektbereich B
Discussion Paper No. **B-365**
March, 1996

[1] I wish to thank Avner Shaked 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 model of social learning by Schlag [5]. Individuals must repeatedly choose an action in a multi-armed bandit. We assume that each indivdiual observes the outcomes of two other individuals' choices before her own next choice must be made – the original model only allows for one observation. Selection of optimal behavior yields a variant of the proportional imitation rule – the optimal rule based on one observation. When each individual uses this rule then the adaptation of actions in an infinite population follows an aggregate monotone dynamic.

JEL classification numbers: C72, C79.

Keywords: social learning, multi-armed bandit, imitation, payoff increasing, proportional imitaiton rule, aggregate monotone dynamic.

1 Introduction

In this paper we consider a variant of a model by Schlag [5]. Schlag considers a model of social learning in which individuals repeatedly face a multi-armed bandit. Between their choices each individual may observe the performance of one other individual, a situation referred to in the following as *single sampling.* Individuals forget about observations in the past. Two alternative approaches to selecting an optimal individual behavior, a bounded rational approach and a population-oriented approach are suggested. Either approach leads to the same unique prescription, the so-called *proportional imitation rule*, of how to choose future actions:

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

ii) never imitate an individual that performed worse than oneself, and

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

In this paper we analyze how the above result changes when an individual is allowed to observe the performance of two other individuals between her choices. This situation will be referred to as *double sampling.* In contrast to the single sampling setting it turns out that there is no behavior that is better than all other behavioral rules (according to either of the selection approaches for individual behavior). However, there is a best way of performing better than under single sampling. This can be achieved by modifying the proportional imitation rule, the resulting rule we call the *adjusted proportional imitation rule.* This variant of the proportional imitation rule specifies additionally to i) to iii) above,

iv) to be more likely to imitate the individual in the sample who realized the higher payoff, and

v) to be more likely to imitate one of the two sampled individuals the lower the payoff of the other one is, especially not to ignore a sampled individual that realized a lower payoff even though he will never be imitated.

Its simple functional form and its performance lead us to selecting the adjusted proportional imitation rule as the optimal rule under double sampling.

Where aggregate behavior of an infinite population of individuals using the optimal rule under single sampling followed the replicator dynamic (Taylor [6]), under double sampling it follows an aggregate monotone dynamic (as defined by Samuelson and Zhang [4]).

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 this setting are presented. In section three we select among the behavioral rules. Section four contains the implications optimal behavior has for the population dynamics. In Section five we consider an alternative two population matching scenario. The Appendix contains the proof of the main theorem which is stated in Section three.

2 The Setting

Consider the following dynamic process of choosing actions, sampling and updating. Let W be a finite population (or set) of N individuals, $N \geq 3$. In a sequence of rounds, each individual must choose an action from a finite set of actions $A = \{1, 2, .., n\}$ where $n \geq 2$. 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 $\Delta(A)$ be the set of probability distributions on A. For a given state s let $p = p(s) \in \Delta(A)$ denote the probability distribution that is associated with randomly selecting an individual and observing the action she has chosen for this round, i.e., $p_i(s) = \frac{1}{N} |\{c \in W : s(c) = i\}|$ $(i \in A)$. 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 $N \cdot p_i \in \mathbb{N}$. Given this notation, the average expected payoff of the population in state s, $\bar{\pi}(s)$, is given by $\bar{\pi}(s) = \sum p_i(s)\,\pi_i$.

Before each round of payoff realization, each individual meets (or samples) two other individuals from the population and observes the payoff each of them received, together with the associated action, in the previous round. Given three different individuals $c, d, e \in W$, the probability that individual 'c' samples individuals 'd' and 'e' is denoted by $P(c \rightsquigarrow \{d, e\})$. In the following we will assume that sampling is symmetric, i.e., that $P(c \rightsquigarrow \{d, e\}) = P(d \rightsquigarrow \{c, e\}) = P(e \rightsquigarrow \{c, d\})$ for all $c, d, e \in W$. The situation in which samplings occurs by choosing two individuals randomly from the population will be called *random sampling*, in this case $P(c \rightsquigarrow \{d, e\}) = \frac{1}{(N-1)(N-2)}$ for all $c, d, e \in W$.

The description of how an individual chooses her next action in a multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$ based on her previous observations is summarized by a behavioral rule. We allow for the individual to use a randomizing device that generates independent events when making this choice. We restrict attention to behavioral rules where observations prior to her last payoff realization do not influence her next choice, i.e., essentially an individual forgets these observations. Hence, a *behavioral rule* is a function $F : A \times [\alpha, \omega] \times \{A \times [\alpha, \omega] \times A \times [\alpha, \omega]\} \to \Delta(A)$ where $F(i, x, \{j, y, k, z\})_r$ is the probability of playing action r after obtaining payoff x with action i and sampling individuals using action j and action k that obtained payoff y and payoff z respectively. For $i, j, k, r \in A$, let

$$F_{ijk}^r := \sum_{x,y,z \in [\alpha,\omega]} F(i, x, \{j, y, k, z\})_r \, P_i(x)\, P_j(y)\, P_k(z)\,,\ i, j, k, r \in A,$$

be the so-called switching probabilities.

A class of behavioral rules of special importance in our analysis will be the class of imitating rules. A behavioral rule F is called *imitating* if $F(i, x, \{j, y, k, z\})_r = 0$ when $r \notin \{i, j, k\}$. For an imitating behavioral rule F let $F(i, x, \{j, y, k, z\})_{jk}$ denote the probability of switching actions (to either action j or k), i.e., $F(i, x, \{j, y, k, z\})_{jk} = 1 - F(i, x, \{j, y, k, z\})_i$. Similarly, let F_{ijk}^{jk} be the associated switching probabilities, i.e., $F_{ijk}^{jk} = 1 - F_{ijk}^i$.

3 Examples

In the following we present some examples of behavioral rules.

Behavioral rules under single sampling can be embedded in the class of behavioral rules under double sampling by randomly selecting one of the two sampled individuals and applying the single sampling rule. More specifically, the behavioral rule f under single sampling, i.e., $f : A \times [\alpha, \omega] \times A \times [\alpha, \omega] \to \Delta(A)$, is associated to the behavioral rule F^f under double sampling defined by $F^f(i, x, \{j, y, k, z\})_r = \frac{1}{2} f(i, x, j, y)_r + \frac{1}{2} f(i, x, k, z)_r$, $i, j, k, r \in A$, $x, y, z \in [\alpha, \omega]$. Behavioral rules constructed in this way will be called *single sampling rules.*

An important behavioral rule under single sampling rule is the imitating rule f^p that satisfies $f^p(i, x, j, y)_j = \frac{1}{\omega - \alpha}[y - x]_+$ where $[x]_+ = x$ when $x > 0$ and $[x]_+ = 0$ when $x \leq 0$. Schlag [5] argues that this so-called *proportional imitation rule with rate* $\frac{1}{\omega - \alpha}$ is the unique optimal rule under single sampling. The associated single sampling rule will be denoted by F^p.

The behavioral rule of importance in the present model of double sampling is the rule we refer to as the adjusted proportional imitation rule. Let $\sigma^* : [\alpha, \omega] \to \mathbb{R}^+$ be the linearly decreasing function such that $\sigma^*(\alpha) = \frac{2}{\omega - \alpha}$ and $\sigma^*(\omega) = \frac{1}{\omega - \alpha}$, i.e.,

$$\sigma^*(x) = \frac{1}{\omega - \alpha} + \frac{\omega - x}{(\omega - \alpha)^2} \text{ for } x \in [\alpha, \omega].$$

Consider the behavioral rule $\hat{F}$ such that $\hat{F}(i, x, \{j, y, k, z\})_j = \frac{1}{2}\sigma^*(z)[y - x]_+$, $\hat{F}(i, x, \{j, y, k, z\})_k = \frac{1}{2}\sigma^*(y)[z - x]_+$ and $\hat{F}(i, x, \{j, y, j, z\})_j = \frac{1}{2}\sigma^*(z)[y - x]_+ + \frac{1}{2}\sigma^*(y)[z - x]_+$, $|\{i, j, k\}| = 3$. In order for $\hat{F}$ to be in fact a behavioral rule we must show that $\hat{F}(i, x, \{j, y, k, z\})_{jk} \leq 1$ when $x < y \leq z$; this is true since

$$\hat{F}(i, \alpha, \{j, y, k, z\})_{jk} = \frac{1}{2}\frac{y - \alpha}{\omega - \alpha}\left(1 + \frac{\omega - z}{\omega - \alpha}\right) + \frac{1}{2}\frac{z - \alpha}{\omega - \alpha}\left(1 + \frac{\omega - y}{\omega - \alpha}\right).$$

We will call $\hat{F}$ the *adjusted proportional imitation rule (based on* $[\alpha, \omega]$*).* Notice that an individual following this rule will be more likely to imitate the individual in the sample that realized the higher payoff. He will never imitate an individual that realized a lower payoff, will never-the-less use the payoff of such an individual to determine how likely to switch to the other individual. Some extreme situations for how the payoff of the one individual influences the probability of switching to the other are as follows: $\hat{F}\left(i,x,\{j,y,k,\alpha\}\right)_j = f^p\left(i,x,j,y\right)_j$ and $\hat{F}\left(i,x,\{j,y,k,\omega\}\right)_j = \frac{1}{2}f^p\left(i,x,j,y\right)_j$.

A popular rule under single sampling is the imitating rule 'imitate if better', where the individual adapts the action of the observed individual if and only if it achieved a higher payoff. In the literature this rule is extended to the framework of multiple sampling in the following two different ways. '*Imitate the best*' (Axelrod [1]) is the imitating behavioral rule F that satisfies: $F\left(i,x,\{j,y,k,z\}\right)_j = 1$ if $y > \max\{x,z\}$ and $F\left(i,x,\{j,y,k,z\}\right)_i = 1$ if $x \geq \max\{y,z\}$, $i,j,k \in A$, $x,y,z \in [\alpha,\omega]$. '*Imitate the best average*' (Bruch [2]; Ellison and Fudenberg [3]) is the imitating behavioral rule F that satisfies $F\left(i,x,\{j,y,j,z\}\right)_j = 1$ if $\frac{1}{2}(y+z) > x$ and 0 otherwise, $F\left(i,x,\{i,y,j,z\}\right)_j = 1$ if $z > \frac{1}{2}(x+y)$ and 0 otherwise, $F\left(i,x,\{j,y,k,z\}\right)_j = 1$ if $y > \max\{x,z\}$, $F\left(i,x,\{j,y,k,y\}\right)_j = F\left(i,x,\{j,y,k,y\}\right)_k = \frac{1}{2}$ if $y > x$ and $F\left(i,x,\{j,y,k,z\}\right)_i = 1$ if $x \geq \max\{y,z\}$, $i,j,k \in A$ with $|\{i,j,k\}| = 3$, $x,y,z \in [\alpha,\omega]$.

4 Selection Among the Rules

The so-called *expected improvement* $EIP_F(s)$ in state s is given by the following expression:

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

Individuals are assumed to prefer so-called improving behavioral rules, these are rules that always generate non negative expected improvement. Formally, a behavioral rule F is called *improving* if $EIP_F(s) \geq 0$ for all $s \in \Delta^N(A)$ and all multi-armed bandits in $\mathcal{G}(A,[\alpha,\omega])$. Schlag [5] gives two alternative scenarios that cause an individual to choose an improving rule.

1) Individuals are boundedly rational. They enter the population by replacing a random individual in the population. They adapt the action last

chosen by this individual. In each round an individual evaluates the performance of her behavior as if she just entered. Individuals prefer a rule that always increases expected payoffs in any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$.

2) Individuals evaluate the performance of their behavior in a population of replicas. An individual considers a population in which each individual is using her behavioral rule. She prefers a rule that is expected to increase average payoffs in each state and each multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$.

Schlag [5] characterizes the set of improving rules under single sampling. Especially it turns out that the proportional imitation rule with rate $\frac{1}{\omega-\alpha}$ is improving and that the rule 'imitate if better' is not improving. Clearly, an improving behavioral rule f in the single sampling setting is associated to a single sampling rule F^f (see Section 3) that is improving in the present double sampling setting. The following theorem characterizes the entire set of improving behavioral rules under double sampling.

Theorem 1 *The behavioral rule F is improving if and only if F is imitating and for all subsets $\{i, j, k\} \subset A$ with $|\{i, j, k\}| > 1$ there exists a function $\sigma_{\{i,j,k\}} : [\alpha, \omega] \to \mathbb{R}_0^+$ such that*

$$\begin{aligned} & F(i, x, \{j, y, k, z\})_{jk} - F(j, y, \{i, x, k, z\})_i - F(k, z, \{i, x, j, y\})_i \\ = \; & \frac{1}{2}\sigma_{\{i,j,k\}}(z)(y - x) + \frac{1}{2}\sigma_{\{i,j,k\}}(y)(z - x), \end{aligned} \tag{1}$$

if $i \notin \{j, k\}$.

Proof. (in the Appendix)

Theorem 1 and its proof give little insight as to which functions $\sigma_{\{i,j,k\}}(\cdot)$ are associated to an improving rule. Of course, the right hand side of (1) must be bounded above by 1, especially $\sigma_{\{i,j,k\}}(y) \leq \max\left\{\frac{1}{y-\alpha}, \frac{1}{2(\omega-\alpha)}\right\}$ for all $y \in [\alpha, \omega]$.

However, Theorem 1 enables us to verify whether a behavioral rule is improving or not. Consider for example the rules 'imitate the best average' and 'imitate the best'. (1) implies that neither of these rules is improving. In the following we show this statement using a counterexample in order to explicitly illustrate how these two rules fail to be improving.

Fix $x \in \left(\alpha, \frac{2}{3}\alpha + \frac{1}{3}\omega\right)$. Consider a multi-armed bandit in which $P_1(x) = 1$, $P_2(\alpha) = \lambda$ and $P_2(\omega) = 1-\lambda$ for some $0 < \lambda < 1$. Then $\pi_1 > \pi_2$ if and only if $\lambda < \frac{\omega - x}{\omega - \alpha}$. Notice that the rule 'imitate the best' and the rule 'imitate the best average' induce the same switching probabilities $F_{122}^2 = 1 - F_{212}^1 = 1 - \lambda^2$ and $F_{211}^1 = 1 - F_{121}^2 = \lambda$. Especially, $\lambda > \frac{2}{3}$ implies $F_{211}^1 - 2F_{121}^2 > 0$ and $F_{122}^2 - 2F_{212}^1 < 0$. Following (8), this leads to negative expected improvement if only action 1 and action 2 are played in the population, with positive probability some individual using action 1 observes some individual using action 2 and if $\frac{2}{3} < \lambda < \frac{\omega - x}{\omega - \alpha}$. Hence we see that neither 'imitate the best' nor 'imitate the best average' is improving.

Under the single sampling rules, Schlag [5] shows that the proportional imitation rule F^p never achieves a lower expected improvement than any other improving single sampling rule. Hence, we say that F^p dominates the single sampling rules. More generally, let $\mathcal{F}$ be a set of behavioral rules. We say that a behavioral rule F *dominates* the set of behavioral rules $\mathcal{F}$ if $EIP_F(s) \geq EIP_{F'}(s)$ for all $F' \in \mathcal{F}$, for any state s and for any multi-armed bandit in $\mathcal{G}(A, [\alpha, \omega])$. Consequently, if $\mathcal{F}$ contains an improving rule and F dominates $\mathcal{F}$ then F is improving.

In the following we will show that improving rules under double sampling with constant switching rates $\sigma_{\{i,j,k\}}(\cdot)$ are of no advantage compared to the single sampling scenario. As mentioned above, the highest expected improvement is realized by the proportional imitation rule F^p. Following (8),

$$EIP_{F^p}(s) = \frac{1}{2N(\omega - \alpha)} \sum_{c,d,e \in W} P(c \rightsquigarrow \{d, e\}) \left(\pi_{s(d)} - \pi_{s(c)}\right)^2 . \qquad (2)$$

Since $\sigma_{\{i,j,k\}}(\omega) \leq \frac{1}{\omega - \alpha}$, following (2) and (8), an improving rule under double sampling with constant switching rates $\sigma_{\{i,j,k\}}$ never achieves a higher expected improvement than F^p.

The advantage of double sampling lies in the fact that switching rates of improving rules must no longer be constant. The following theorem states that, unlike under single sampling, under double sampling there is no behavioral rule that dominates all other improving rules. However, we show that following the adjusted proportional imitation rule $\hat{F}$ is the best way of performing better than the proportional imitation rule F^p.

Theorem 2 *Let $\mathcal{F}_1$ be the set of single sampling rules that are improving. Let $\mathcal{F}_2$ be the set of rules that dominate $\mathcal{F}_1$. Then the adjusted proportional imitation rule $\hat{F}$ dominates $\mathcal{F}_2$.*

There is no behavioral rule that dominates the set of improving rules.

In the following, let $\mathcal{F}_3$ be the set of rules that dominate $\mathcal{F}_2$.

Proof. Consider an improving rule $F' \in \mathcal{F}_2$, let $\sigma'_{\{i,j,k\}}$ be the associated switching rates. We will first show that $\sigma'_{\{i,j,k\}}(\omega) = \frac{1}{\omega-\alpha}$. As mentioned above, $\sigma'_{\{i,j,k\}}(\omega) \leq \frac{1}{\omega-\alpha}$. Consider the multi-armed bandit in which $P_i(\alpha) = 1$ and $P_j(\omega) = P_k(\omega) = 1$. Consider a population with one individual using i, one using j and the rest using k. Using the fact that $a^j_{ijk} = a^k_{ijk} = \sigma'_{\{i,j,k\}}(\omega)$ it follows that $EIP_{F'} = \frac{1}{N}\sigma'_{\{i,j,k\}}(\omega)(\omega-\alpha)^2 \leq \frac{1}{N}\frac{1}{\omega-\alpha}(\omega-\alpha)^2 = EIP_{F^p}$. Since F' dominates F^p and $F^p \in \mathcal{F}_1$ we obtain that $\sigma'_{\{i,j,k\}}(\omega) = \frac{1}{\omega-\alpha}$.

Notice that

$$\frac{1}{2}\sigma'_{\{i,j,k\}}(\omega)(y-\alpha) + \frac{1}{2}\sigma'_{\{i,j,k\}}(y)(\omega-\alpha) \leq F'(i,\alpha,\{j,y,k,\omega\})_{jk} \leq 1$$

implies

$$\sigma'_{\{i,j,k\}}(y) \leq \frac{1}{\omega-\alpha}\left[2-(y-\alpha)\sigma'_{\{i,j,k\}}(\omega)\right] = \frac{1}{\omega-\alpha} + \frac{\omega-y}{(\omega-\alpha)^2} = \sigma^*(y). \tag{3}$$

Hence, (3) and (8) imply $EIP_{\hat{F}} \geq EIP_{F'}$ for any state s and any multi-armed bandit in $\mathcal{G}(A,[\alpha,\omega])$ which means that $\hat{F} \in \mathcal{F}_3$. Especially, it follows that $\sigma_{\{i,j,k\}}(y) = \sigma^*(y)$ for any rule $F \in \mathcal{F}_3$.

We will now construct a rule that is not dominated by any rule in $\mathcal{F}_3$. This will show that there is no rule that dominates all other improving rules. Let $\tilde{F}$ be the behavioral rule that is constructed like $\hat{F}$ using the function $\tilde{\sigma}$ where $\tilde{\sigma}(y) = \frac{2}{\omega-\alpha}$ when $y \leq \frac{\alpha+\omega}{2}$ and $\tilde{\sigma}(y) = 0$ for $y > \frac{\alpha+\omega}{2}$. It follows that $\tilde{F}(i,\alpha,\{j,y,k,z\})_{jk} \leq 1$ and hence that $\tilde{F}$ is in fact a behavioral rule. Moreover, by construction, $\tilde{F}$ is improving and $\tilde{\sigma}(y) > \sigma^*(y)$ for all $\alpha < y \leq \frac{\alpha+\omega}{2}$. Hence, $\tilde{F}$ is not dominated by any rule in $\mathcal{F}_3$. ■

One can argue that an individual will choose an improving rule that dominates the improving rules under single sampling, i.e., a rule in $\mathcal{F}_2$. She might as well choose a rule that is best at doing this, i.e., a rule in $\mathcal{F}_3$. We presented such a rule, the adjusted proportional imitation rule, that additionally never imitates lower payoffs and has a simple form. This leads us to selecting this rule as the optimal rule under double sampling.

5 Population Dynamics

In this section we consider the aggregate behavior of a population in which each individual uses the optimal rule. We will restrict attention to random sampling. Moreover, we will consider adjustment in infinite populations as an approximation of the short run adjustment of a large population. Schlag [5] specifies the exact meaning of this approximation for the single sampling setting. In an infinite population, random sampling means that the probability that an individual observes action i is equal to the proportion of individuals using this action. In this sense, a description of the proportions p_i using action i for each $i \in A$ is sufficient to determine the population adjustment. Hence we will identify the state of a population with $p = (p_i)_{i\in A} \in \Delta(A)$. Straightforward calculations show that the adjustment process $(p^t)_{t\in\mathbb{N}}$ of a monomorphic population (each individual is following the same behavior) in which the underlying rule is improving, given an initial state $p^1 \in \Delta(A)$, is given by

$$p_i^{t+1} = p_i^t + p_i^t \sum_{j,k} \frac{1}{2} p_j^t p_k^t \left[a_{ijk}^j (\pi_i - \pi_k) + a_{ijk}^k (\pi_i - \pi_j) \right] ,$$

for $i \in A$ and $t \in \mathbb{N}$. If, in addition, the underlying rule is the adjusted proportional imitation rule $\hat{F}$, we obtain

$$p_i^{t+1} = p_i^t + \left[\frac{1}{\omega - \alpha} + \frac{\omega - \bar{\pi}(p^t)}{(\omega - \alpha)^2} \right] \left(\pi_i - \bar{\pi}\left(p^t\right) \right) \cdot p_i^t , \qquad (4)$$

where $\bar{\pi}(p) = \sum_{i\in A} p_i \pi_i$.

6 A Two Population Matching Scenario

What about a setting in which the multi-armed bandit is not stationary over time? We will consider a popular example for such a situation; individuals will be randomly matched to play a game. Consider two finite, disjoint sets (populations) of individuals W_1 and W_2, each of size N, also referred to as population one and two. Let A_i be the finite set of actions available to an individual in population i, $i = 1, 2$. Payoffs are realized by matching individuals from different populations. 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*. 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 of opposite populations are matched at random in pairs, for an individual in population one this means the following. 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 defined 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, each individual in population one is facing a multi-armed bandit $\left\langle A_1, (P_i')_{i\in A}\right\rangle \in \mathcal{G}\left(A_1, [\alpha_1, \omega_1]\right)$ that depends on the population shares in population two; $P_i'(x) = \sum_{j \in A} q_j P_{ij}^1(x)$ for $x \in [\alpha_1, \omega_1]$ and $i \in A$.

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

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

Schlag [5] gives two scenarios in which an individual prefers to use the same rule in this population matching setting as in the former multi-armed bandit setting:

1) It might be that individuals do not realize that the multi-armed bandit is non stationary or that they simply ignore this fact.

2) An individual might choose her rule according to its performance in a population of replicas and prefers a rule that is expected to increase average payoffs whenever all individuals in the opposite population do not change their action.

Hence, we consider the adjusted proportional imitation rule based on $[\alpha_i, \omega_i]$ to be the optimal rule for an individual in population i in this population matching setting. In the following we consider the aggregate behavior of the two populations under random sampling when each individual uses her optimal rule. As in Section 5 we consider the limit of this adjustment as the population size N tends to infinity and apply a law of large numbers type of argument. Analogue to (4), the resulting adjustment process $(p^t, q^t)_{t \in \mathbb{N}}$ is given by

$$\begin{aligned} p_i^{t+1} &= p_i^t + \left[\frac{1}{\omega_1 - \alpha_1} + \frac{\omega_1 - \pi_1(p^t, q^t)}{(\omega_1 - \alpha_1)^2}\right]\left[\pi_1\left(i, q^t\right) - \pi_1\left(p^t, q^t\right)\right] \cdot p_i^t, \quad (5) \\ q_j^{t+1} &= q_j^t + \left[\frac{1}{\omega_2 - \alpha_2} + \frac{\omega_2 - \pi_2(p^t, q^t)}{(\omega_2 - \alpha_2)^2}\right]\left[\pi_2\left(p^t, j\right) - \pi_2\left(p^t, q^t\right)\right] \cdot q_j^t\ , \end{aligned}$$

for $i \in A_1, j \in A_2$ and $t \in \mathbb{N}$. According to Samuelson and Zhang [4], (5) is called an aggregate monotone dynamic. Under single sampling the adjustment generated when each individual is using her optimal rule (i.e., the proportional imitation rule with rate $\frac{1}{\omega_i - \alpha_i}$ for population i) is approximated by the following discrete version of the replicator dynamic (Taylor [6]):

$$\begin{aligned} p_i^{t+1} &= p_i^t + \frac{1}{\omega_1 - \alpha_1}\left[\pi_1\left(i, q^t\right) - \pi_1\left(p^t, q^t\right)\right] \cdot p_i^t, \\ q_j^{t+1} &= q_j^t + \frac{1}{\omega_2 - \alpha_2}\left[\pi_2\left(p^t, j\right) - \pi_2\left(p^t, q^t\right)\right] \cdot q_j^t \ . \end{aligned} \tag{6}$$

Comparing this to (5) we see that the advantage of double sampling for individuals using their optimal rule $\hat{F}$ is greatest when average payoffs in their own population are low.

References

[1] R. M. Axelrod, *The Evolution of Cooperation,* Basic Books, New York, 1984.

[2] E. Bruch, "Evolution von Kooperation in Netzwerken", Diplomarbeit, University of Bonn, 1993.

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

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

[5] K. H. Schlag, "Why Imitate, and if so, How? A Bounded Rational Approach to Multi-Armed Bandits," University of Bonn, Disc. Paper **B-361**, Bonn, 1996.

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

A The Proof of Theorem 1

Proof. For $i, j, k \in A$ and $s \in A^W$ let

$$p_{ijk}(s) = \sum_{\substack{c,d,e \in W \\ \{s(c),s(d),s(e)\}=\{i,j,k\}}} P\left(c \rightsquigarrow \{d,e\}\right) \ .$$

We will first show the 'if' statement.

$$\begin{aligned}
EIP_F(s) &= \frac{1}{N}\sum_{\substack{c,d,e\in W\\ s(d)=s(e)}} P\left(c \rightsquigarrow \{d,e\}\right)\left[F^{s(d)}_{s(c)s(d)s(e)}\left(\pi_{s(d)}-\pi_{s(c)}\right)\right] \\
&\quad +\frac{1}{N}\sum_{\substack{c,d,e\in W\\ s(d)\neq s(e)}} P\left(c \rightsquigarrow \{d,e\}\right) \\
&\quad \left[F^{s(d)}_{s(c)s(d)s(e)}\left(\pi_{s(d)}-\pi_{s(c)}\right)+F^{s(e)}_{s(c)s(d)s(e)}\left(\pi_{s(e)}-\pi_{s(c)}\right)\right] \\
&= \frac{1}{3N}\sum_{i,j\in A} p_{ijj}(s)\left(F^j_{ijj}-2F^i_{jij}\right)\left(\pi_j-\pi_i\right) \qquad (7) \\
&\quad +\frac{1}{3N}\sum_{\substack{i,j,k\in A\\ |\{i,j,k\}|=3}} p_{ijk}(s)\begin{bmatrix} \left(F^j_{ijk}-F^i_{jik}\right)\left(\pi_j-\pi_i\right) \\ +\left(F^k_{ijk}-F^i_{kij}\right)\left(\pi_k-\pi_i\right) \\ +\left(F^k_{jik}-F^j_{kij}\right)\left(\pi_k-\pi_j\right)\end{bmatrix}
\end{aligned}$$

Consider actions $i,j,k\in A$ such that $i\notin\{j,k\}$. Then

$$\begin{aligned}
& F^{jk}_{ijk}-F^i_{jik}-F^i_{kij} \\
=\ & \sum_{x,y,z} P_i(x)\,P_j(y)\,P_k(z) \\
& \left[F\left(i,x,\{j,y,k,z\}\right)_{jk}-F\left(j,y,\{i,x,k,z\}\right)_i-F\left(k,z,\{i,x,j,y\}\right)_i\right] \\
=\ & \sum_{x,y,z} P_i(x)\,P_j(y)\,P_k(z)\left[\frac{1}{2}\sigma_{\{i,j,k\}}(z)\,(y-x)+\frac{1}{2}\sigma_{\{i,j,k\}}(y)\,(z-x)\right] \\
=\ & \frac{1}{2}\left[\sum_z P_k(z)\,\sigma_{\{i,j,k\}}(z)\right]\left(\pi_j-\pi_i\right)+\frac{1}{2}\left[\sum_y P_j(y)\,\sigma_{\{i,j,k\}}(y)\right]\left(\pi_k-\pi_i\right)
\end{aligned}$$

and, given $a^l_{ijk}=\sum_y P_l(y)\,\sigma_{\{i,j,k\}}(y)$, $l\in\{i,j,k\}$, we obtain

$$\begin{aligned}
& -\left[\left(F^j_{ijk}-F^i_{jik}\right)\left(\pi_j-\pi_i\right)+\left(F^k_{ijk}-F^i_{kij}\right)\left(\pi_k-\pi_i\right)+\left(F^k_{jik}-F^j_{kij}\right)\left(\pi_k-\pi_j\right)\right] \\
=\ & \left(F^{jk}_{ijk}-F^i_{jik}-F^i_{kij}\right)\pi_i+\left(F^{ik}_{jik}-F^j_{ijk}-F^j_{kij}\right)\pi_j+\left(F^{ij}_{kij}-F^k_{ijk}-F^k_{jik}\right)\pi_k \\
=\ & \left[\frac{1}{2}\left(\pi_k-\pi_i\right)a^j_{\{i,j,k\}}+\frac{1}{2}\left(\pi_j-\pi_i\right)a^k_{\{i,j,k\}}\right]\pi_i \\
& +\left[\frac{1}{2}\left(\pi_i-\pi_j\right)a^k_{\{i,j,k\}}+\frac{1}{2}\left(\pi_k-\pi_j\right)a^i_{\{i,j,k\}}\right]\pi_j \\
& +\left[\frac{1}{2}\left(\pi_i-\pi_k\right)a^j_{\{i,j,k\}}+\frac{1}{2}\left(\pi_j-\pi_k\right)a^i_{\{i,j,k\}}\right]\pi_k \\
=\ & -\frac{1}{2}\left[\left(\pi_i-\pi_j\right)^2 a^k_{\{i,j,k\}}+\left(\pi_i-\pi_k\right)^2 a^j_{\{i,j,k\}}+\left(\pi_j-\pi_k\right)^2 a^i_{\{i,j,k\}}\right]\leq 0.
\end{aligned}$$

Hence, (7) simplifies to

$$EIP_F(s) = \frac{1}{2N} \sum_{c,d,e \in W} P(c \rightsquigarrow \{d,e\}) \left(\pi_{s(d)} - \pi_{s(c)}\right)^2 a^{s(e)}_{\{s(c),s(d),s(e)\}} \tag{8}$$

and it follows that $EIP_F \geq 0$.

We now come to the proof of the 'only if' statement. In order to simplify the presentation of the proof we will assume that $P(c \rightsquigarrow \{d,e\}) > 0$ for all $c,d,e \in W$ ($|\{c,d,e\}| = 3$). The proof can be easily adjusted to the more general case.

The fact that F is imitating follows just like under single sampling (Schlag [5]). Assume that $F(i,x,\{j,y,k,z\})_r > 0$ for some $r \notin \{i,j,k\}$ and $x,y,z \in [\alpha,\omega]$. Consider a multi-armed bandit in which $P_i(x) = P_i(y) = P_i(\omega) = \frac{1}{3}$, $P_j \equiv P_k \equiv P_i$ and $P_l(\alpha) = 1$ for all $l \notin \{i,j,k\}$. In a state s which only actions in $\{i,j,k\}$ are being played it follows that $EIP_F(s) < 0$.

Next we will show (1) for $i \neq j = k$. Consider a population state in which one individual is playing action i and the rest are playing action j. Then

$$EIP_F(s) = \frac{1}{N}\left(F^j_{ijj} - 2F^i_{jij}\right)(\pi_j - \pi_i). \tag{9}$$

Let

$$g(x,y,z) = F(i,x,\{j,y,j,z\})_j - F(j,y,\{i,x,j,z\})_i - F(j,z,\{i,x,j,y\})_i,$$

$x,y,z \in [\alpha,\omega]$. Let $y = z$. We now follow the same arguments as in the proof of Theorem 1 in Schlag [5] to shows that there exists $\sigma_{ijj} : [\alpha,\omega] \to \mathbb{R}^+_0$ such that

$$g(x,y,y) = \sigma_{ijj}(x)(y-x) \text{ for all } x,y \in [\alpha,\omega]. \tag{10}$$

For given $x,y \in [\alpha,\omega]$, consider the multi-armed bandit where $P_i(x) = P_j(y) = 1$. Then $F^j_{ijj} - 2F^i_{jij} = g(x,y,y)$ and hence following (9),

$$g(x,y) \geq 0 \text{ and } g(y,x) \leq 0 \text{ whenever } y > x. \tag{11}$$

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

$$\frac{g(x,y,y)}{y-x} = \frac{g(x,z,z)}{z-x} \; \forall y < x < z. \tag{12}$$

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 $\lambda < \frac{z-x}{z-y} =: \lambda^*$. Again following (9), we obtain

$$\begin{aligned} F^j_{ijj} - 2F^i_{jij} = \lambda g(x,y) + (1-\lambda) g(x,z) &\geq 0 \text{ if } \lambda < \lambda^* \text{ and} \\ \lambda g(x,y) + (1-\lambda) g(x,z) &\leq 0 \text{ if } \lambda > \lambda^* \end{aligned}$$

Therefore, $\lambda^* g(x,y) + (1-\lambda^*) g(x,z) = 0$, which, after simplification, shows that (12) is true.

Following (12) there exists $\sigma_{ijj} : (\alpha,\omega) \to \mathbb{R}^+_0$ such that $g(x,y,y) = \sigma_{ijj}(x) \cdot (y-x)$ 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, (10) holds for all $x, y \in [\alpha,\omega]$.

Consider now a multi-armed bandit with $P_i(x) = 1$, $P_j(y) = \lambda$ and $P_j(z) = 1-\lambda$ for $y < x < z$ and $0 \leq \lambda \leq 1$. Following (9) and given

$$\begin{aligned} I(\lambda) &= F^j_{ijj} - 2F^i_{jij} \qquad (13) \\ &= \lambda^2 \sigma_{ijj}(y)(y-x) + 2\lambda(1-\lambda) g(x,y,z) + (1-\lambda)^2 \sigma_{ijj}(z)(z-x) \end{aligned}$$

we obtain that $I \geq 0$ if and only if $\pi_j \geq \pi_i$. Hence $I = 0$ if and only if $\pi_j = \pi_i$ if and only if $\lambda = \frac{z-x}{z-y} =: \lambda^*$. Since $I(\lambda^*) = 0$, $-(z-x)\sigma_{ijj}(y) + 2g(x,y,z) + (x-y)\sigma_{ijj}(z) = 0$ and hence

$$g(x,y,z) = \frac{1}{2}\sigma_{ijj}(z)(y-x) + \frac{1}{2}\sigma_{ijj}(y)(z-x). \qquad (14)$$

We will now derive $g(x,y,z)$ for $y \leq z < x$. Consider a multi-armed bandit with $P_j(x) = 1$, $P_i(y) = \lambda$, $P_i(z) = \mu$ and $P_i(z') = 1-\lambda-\mu$ for $z' > x$. Then

$$\begin{aligned} I = \; & \lambda^2 \sigma_{ijj}(y)(y-x) + \mu^2 \sigma_{ijj}(z)(z-x) + 2\lambda\mu g(x,y,z) \qquad (15) \\ & + \lambda(1-\lambda-\mu)\left[\sigma_{ijj}(z')(y-x) + \sigma_{ijj}(y)(z'-x)\right] \\ & + \mu(1-\lambda-\mu)\left[\sigma_{ijj}(z')(z-x) + \sigma_{ijj}(z)(z'-x)\right] \\ & + (1-\lambda-\mu)^2 \sigma_{ijj}(z')(z'-x) \; . \end{aligned}$$

As before, $I = 0$ if and only if $\pi_i = \pi_j$ if and only if $\lambda(x-y) + \mu(x-z) = (1-\lambda-\mu)(z'-x)$. Setting $\lambda(x-y) + \mu(x-z) = (1-\lambda-\mu)(z'-x)$, together with (15) implies that (14) also holds for $y \leq z < x$.

Repeating such calculations for the remaining values of (x, y, z) not yet considered finally yields that (14) holds for all $x, y, z \in [\alpha, \omega]$. This completes the proof of the 'only if' statement for $j = k$.

We now proceed with the case where $j \neq k$. Consider a population in which one individual is playing i, one is playing k and the rest are playing j. Consider a multi-armed bandit in which $\pi_j = \pi_k$. Then

$$\begin{aligned} EIP_F(s) &= \frac{1}{3N} p_{ijj}(s) \left(F_{ijj}^{j} - 2F_{jij}^{i} \right) (\pi_j - \pi_i) \\ &\quad + \frac{1}{3N} p_{ijk}(s) \left[F_{ijk}^{jk} - F_{jik}^{i} - F_{kij}^{i} \right] (\pi_j - \pi_i) \end{aligned}$$

Since $F_{ijj}^{j} - 2F_{jij}^{i} = 0$ if $\pi_i = \pi_j$ it follows that $F_{ijk}^{jk} - F_{jik}^{i} - F_{kij}^{i} = 0$ if and only if $\pi_i = \pi_j$ must hold. Following the same arguments as in the proof where $j = k$ we obtain that there exists $\sigma_{ijk} \equiv \sigma_{ikj} : [\alpha, \omega] \to \mathbb{R}^{+}$ such that

$$\begin{aligned} & F(i, x, \{j, y, k, z\})_{jk} - F(j, y, \{i, x, k, z\})_i - F(k, z, \{i, x, j, y\})_i \\ = & \frac{1}{2} \sigma_{ijk}(z)(y - x) + \frac{1}{2} \sigma_{ijk}(y)(z - x) \end{aligned}$$

holds for all $x, y, z \in [\alpha, \omega]$.

The only thing remaining to show is that σ_{ijk} is independent of a permutation of i, j and k. Consider a multi-armed bandit with $P_i(x) = P_j(y) = P_k(z) = 1$ and a population in which one individual is playing i, one is playing k and the rest are playing j. Following the calculations when proving the 'if' statement, we obtain

$$\begin{aligned} EIP_F(s) &= \frac{1}{3N} p_{ijj}(s) \sigma_{ijj}(y)(y - x)^2 \\ &\quad + \frac{1}{3N} p_{kjj}(s) \sigma_{kjj}(y)(y - z)^2 \\ &\quad - \frac{1}{3N} p_{ijk}(s) \left[\begin{array}{c} x \left[\frac{1}{2} \sigma_{ijk}(z)(y - x) + \frac{1}{2} \sigma_{ijk}(y)(z - x) \right] \\ + y \left[\frac{1}{2} \sigma_{jik}(z)(x - y) + \frac{1}{2} \sigma_{jik}(x)(z - y) \right] \\ + z \left[\frac{1}{2} \sigma_{kij}(x)(y - z) + \frac{1}{2} \sigma_{kij}(y)(x - z) \right] \end{array} \right] \end{aligned} \tag{16}$$

Setting $y = z$ this simplifies to

$$\begin{aligned} EIP_F(s) &= \frac{1}{3N} p_{ijj}(s) \sigma_{ijj}(y)(y - x)^2 \\ &\quad + \frac{1}{3N} p_{ijk}(s) \left[x \sigma_{ijk}(y) - \frac{1}{2} y \sigma_{jik}(y) - \frac{1}{2} y \sigma_{kij}(y) \right] \end{aligned}$$

which implies, when setting $y = x \neq 0$, that $\sigma_{ijk}(x) - \frac{1}{2}\sigma_{jik}(x) - \frac{1}{2}\sigma_{kij}(x) = 0$ for any $x \neq 0$. Similarly, setting $x = y$ in (16) leads to $\sigma_{kij}(x) - \frac{1}{2}\sigma_{ijk}(x) - \frac{1}{2}\sigma_{jik}(x) = 0$ for any $x \neq 0$. Together this means that $\sigma_{kij}(x) = \sigma_{ijk}(x) = \sigma_{jik}(x)$ for all $x \neq 0$. The special case of $x = 0$ is easily shown using more general multi-armed bandits and hence the proof is complete. ■