Updating Strategies Through Observed Play

-

Optimization Under Bounded Rationality [1]

R. Cressman
Department of Mathematics
Wilfrid Laurier University
Waterloo, Ontario, N2L 3C5
Canada

K.H. Schlag
Economic Theory III
University of Bonn
Adenauerallee 24-26
D-53113 Bonn, Germany
www.econ3.uni-bonn.de
schlag@econ3.uni-bonn.de

University of Bonn Discussion Paper No. **B-432**
April, 1998

[1]We wish to thank Dorothea Herreiner for helpful comments. Financial support from the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 303 at the University of Bonn and from the Natural Sciences and Engineering Research Council of Canada is gratefully acknowledged.

Abstract

Individuals repeatedly face a multi-decision task with unknown payoff distributions. They have minimal memory and update their *strategy* by observing previous *play* (and not strategy) of someone else. We select behavior rules that increase average payoffs as often as possible in a large population where all use the same rule. Here imitation generalizes to a pasting procedure. When decisions within the task are unrelated, individuals eventually learn the efficient strategy but the underlying dynamic is not monotone. However, when choices influence which decisions are subsequently faced in the task, play may not be efficient in the long run as it approaches a Nash equilibrium of the agent normal form.

Keywords: Multi-Armed Bandit, improving, undominated behavioral rule, play-wise imitating, replicator dynamic, monotone dynamics.

JEL classification: C72, C79

1 Introduction

People are constantly faced with choices, that affect their well-being, without a complete knowledge of the consequences of their possible actions. In a society, people in these situations often make their choice through observing (or *sampling*) actions and their consequences of other individuals. As argued by Samuelson (1997), they develop rules of thumb, based on their limited information, which their experience suggests will lead to their best choice. We call these *behavioral rules*.

In extensive-form games, strategies are typically very complex, consisting of a plan of action at each decision node. The ability to observe the entire strategy of other players is particularly unnatural in large extensive-form games. To relax this assumption and to assume that only previous *play* (i.e., choices made on the outcome path) is observable has led to investigations by Fudenberg and Levine (1993) and Nöldeke and Samuelson (1993). In both of these models, individuals are still rather sophisticated, being completely informed about the game's payoff and able to calculate best replies. Models of boundedly rational behavior with more constraints on individual rationality (e.g., imitation models) typically appeal to the normal form when analyzing extensive-form games (e.g., Cressman, 1997; Gale et al., 1995). When motivated by models of imitation, this means that *strategies* are sampled. Only recently have attempts been made to investigate models of imitation based on sampled *play* (Cooper, 1997; Ruebeck, 1997). Results for the repeated Prisoners' Dilemma are shown to be drastically altered. However, these investigations rely on specific individual behavior which leaves open the possibility that no change in prediction arises under appropriately chosen alternative behavior. We set out to investigate whether universal rules of behavior exist that can cope with this imperfect information in an optimal way. For this analysis, we restrict attention to *Multi-Decision Tasks* which are one-player extensive-form games against nature with chance nodes.[1] A strategy describes a plan of action in each node (or decision) and play describes the decisions faced and actions chosen during one task.

We consider a large population of identical individuals, each individual independently and repeatedly facing the same Multi-Decision Task. Each decision in the task has a finite number of choices, play in the task yields payoffs in a known bounded interval. This

[1] Technically, the extensive form game against nature is a perfect information game since each decision is a singleton information set. However, the phrase "perfect information" is not to be confused with the incomplete information inherent in our tasks where the explicit payoff consequences of action(s) chosen is not known.

is all an individual initially knows about the task. Each time play in the task is over, an individual receives information about the play of others. He samples at random an individual from the population and observes last play (not the strategy) and payoff of the sampled individual. Then, given his behavioral rule, the individual determines which strategy to use the next time he faces the task.

Our assumptions of bounded rationality are as follows. We assume that an individual forgets all occurrences prior to the last time he faced the task. Thus, a behavioral rule depends only own last strategy plus the play and payoff it generated in the last task together with the play and payoff of the sampled individual. We assume that each individual uses the same behavioral rule.

The first purpose of this paper is to find a good rule of thumb that can be applied to arbitrary Multi-Decision Tasks. This rule should cause average payoffs to increase as often as possible and it should be simple to implement. Our choice of properties of a good rule are guided both by the boundedly rational optimization approach found in Schlag (1998b) and by existing dynamic properties in the Evolutionary Game Theory literature summarized in Remark 1. Two guidelines will aid our systematic search.

First of all, we will define a partial order for comparing the performance of different rules. The partial order on the set of rules is defined as follows. The more often a rule leads to a (weak) increase in average payoffs in the population the better.[2] Two rules that increase average payoffs in the same situations will be compared if one always generates a higher increase than the other. A maximal element under this partial order in a particular set of rules will be called *undominated* in this set. *Improving* rules always increase average payoffs. Since Never Switch Strategies is improving, a rule that is undominated among all *feasible* rules is improving (but need not be unique). An improving rules is *strictly improving* if average payoffs increase strictly whenever all strategies are present.

Secondly, we will restrict attention to rules that behave in complex tasks as they would in simple ones whenever current information is *consistent* with the simpler setting. That is, consistency means that an individual behaves as he would in the simplest task in which the present observations could occur. This condition considerably simplifies our analysis. It turns out that we only need to search for a best rule in three kinds of tasks: tasks that consist of a single decision, tasks in which decisions are independent, and tasks with two decisions in sequence.

[2]The implicit risk neutrality in the population consisting of identical individuals can be replaced by more general risk preferences: translate payoffs into utilities before applying the rule.

Tasks with a single decision (Section 2) are called *Multi-Armed Bandits* (Rothschild, 1974; Berry and Fristedt, 1985). Here, notions of play and strategy coincide. As shown by Schlag (1998b), undominated rules for Multi-Armed Bandits are strictly improving and *imitative* (i.e. either maintain own last strategy or switch to the strategy used by the sampled individual).

Multi-Decision Tasks consisting of unrelated or independent decisions (Section 3) are called *Parallel Bandits.* In these tasks there is an initial move by nature that selects one of a finite number of decisions (each modelled by a Multi-Armed Bandit, Figure 1). The decision last faced may differ from the decision in which play is sampled. The specific case where there are only two decisions with one decision containing only one choice is called a *Simple Parallel Bandit* (see Figure 2). As in Multi-Armed Bandits, undominated rules for Simple Parallel Bandits are strictly improving and imitative. These rules ignore own last payoff in *cross-decision* learning situations where sampled play and own last play differ. To find a rule for more general Parallel Bandits we consider only those rules consistent with undominated behavior in Simple Parallel Bandits. Again, undominated rules are strictly improving. However, here imitation is *play-wise* in that one only switches actions by adopting sampled play. This contrasts markedly with recent treatments in the literature (e.g. Cooper, 1997; Ruebeck, 1997; and related to Harrington, 1996) where individuals use behavioral rules that ignore one's own last play at unsampled decisions.

The task consisting of two interrelated decisions (Section 4) is called a *Centipede Bandit.* This task contains two decisions for the individual where actions chosen at the first decision influence whether the second decision is faced or not (see Figure 3). These tasks are too complex for rules to exist that are strictly improving (Theorem 7). Rules undominated among the consistent rules change strategies through play-wise imitation. Among these rules, the *Centipede Proportional Imitation Rule With Maximal Switch* (CPIR With Maximal Switch) stands out due to its simple implementation. It minimizes switching at the crucial initial decision and imitates sampled play at decisions not reachable under own last strategy.

Further analysis of consistent rules for more general Multi-Decision Tasks is not necessary. Behavior of consistent and undominated rules are uniquely determined by their behavior in Parallel and Centipede Bandits (see Conclusion in Section 5). Thus, CPIR With Maximal Switch becomes the most preferred rule in any Multi-Decision Task.

The second purpose of the paper is to study population dynamics. Following Schlag (1998b), undominated rules for Single-Decision Tasks lead to the discrete replicator dy-

namic which is the original monotone selection dynamic borrowed from the biological perspective of Evolutionary Game Theory (Taylor, 1979). That is, the boundedly rational approach of (Schlag, 1998b) provides a justification for the use of the replicator dynamic over any other myopic adjustment process in Evolutionary Game Theory applied to Sociology and Economics. Individuals are lead in the long run to the *best* (or *efficient*) *strategy* (i.e., the one that maximizes expected payoffs).

In Parallel Bandits, a monotonic increase of play of the best strategy (such as the replicator dynamic) cannot be universally enforced. However, the adjustment of play in any given decision under an undominated consistent rule follows the replicator dynamic. Thus, similar to Single-Decision Tasks, an undominated consistent rule leads the population to the best strategy (even though its proportion in the population is not always monotonically increasing).

In Centipede Bandits, the most preferred rule CPIR with Maximal Switch induces a population dynamic in which average payoffs eventually increase in each round (Theorem 14). Nevertheless, the population may not reach the best strategy in the long run (Theorem 13) since it is only guaranteed that play converges to a Nash equilibrium of the agent normal form game (Remark 5). Thus, an individual behaves in the long run as if he were making his choices at the various decisions independently. The prediction that best strategies will be learned in the long run is recovered by including rare mutations in the learning model.

In Section 5, we summarize the basic insights of the paper. The Appendix contains a note summarizing dynamic properties used in the literature along with the proofs of results in Sections 3 and 4 that are not given in the main text.

2 Single-Decision Tasks

We embed our analysis of learning from observed play in the framework of the single-decision model of Schlag (1998b).

Consider an infinite population of identical individuals. In each round, each individual faces the same *Single-Decision Task* where he must choose among a fixed finite number of actions belonging to the set S with $|S| = n \geq 2$. Each choice of action $i \in S$ realizes a payoff that is drawn independently from a probability distribution P_i that has finite support in a given finite interval $[\alpha, \omega]$. π_i denotes the expected payoff generated by choosing action i. The model so far forms the basis for a *Multi-Armed* (in fact, n-armed)

Bandit, denoted by $\langle (i, P_i)_{i \in S} \rangle$, that is used extensively in the literature (Rothschild, 1974; Berry and Fristedt, 1985). The set of Multi-Armed Bandits with action set S and payoffs generated in $[\alpha, \omega]$ will be denoted by $MAB\,(S, [\alpha, \omega])$.

An individual prefers the action that generates the highest expected payoff; such an action is called *best* (or *efficient*). Initially all an individual knows is that he is facing a task belonging to $MAB\,(S, [\alpha, \omega])$. After each choice an individual receives information about the performance of others who faced the same task.

For instance, if an individual has unlimited possibilities to observe what other individuals chose in the last round, then he could choose an action that maximized average payoffs in the previous round. If each individual uses this rule and if each action is chosen by some individuals in the first round, then all individuals would chose a best action in all later rounds.

We consider assumptions on information (and memory) more in the spirit of bounded rationality. Specifically, after each choice an individual is informed about the previous play (i.e., action) and payoff of a randomly selected individual from the population. Individual memory is also limited. An individual forgets all occurrences before his own last play. Thus, an individual may only condition his choice of an action for the next round on his own last play and payoff as well as that of an individual sampled at random from the population.[3] This assumption of bounded rationality (or close variants that underlie such updating rules as Imitate if Better[4]) is common in both theoretical treatments (Ellison and Fudenberg, 1995) and simulation models (Nowak and May, 1992). Individual next choice as a function of information gathered is formally described by a *behavioral rule* F. To be *feasible* for our Multi-Armed Bandit, $F = F(i, x, j, y)_k$ is a map that specifies the probability of choosing action k in the next round if last action was i with payoff x and the observed action is j with payoff y (where $x, y \in [\alpha, \omega]$).[5] We are not concerned with

[3] An individual using action i is sampled with probability equal to the frequency p_i of individuals in the population that used this action in the previous round. The event of who an individual samples together with the event of who the sampled individual samples is assumed to be independent of all other events. The two most natural scenarios are (i) individuals sample independently, and (ii) individuals are independently grouped in pairs that sample each other. This model is easily extended to include the possibility that not all individuals sample in each round.

[4] That is, switch to the sampled action if and only if the sampled payoff is higher than your previous payoff, otherwise do not change actions.

[5] We allow individuals to use a randomizing device to aid their decision of which action to choose next as in a one-shot mixed strategy. Important is the fact that an individual remembers his action chosen in the previous round, not the mixed strategy.

how an individual chooses an action in the first round and simply assume that each action is chosen by some individuals.

We focus on populations in which each individual uses the same rule. Our objective is to find a behavioral rule for which population average payoff increases as often as possible. That is, in keeping with the argument of Samuelson (1997) that people apply the same "rules of thumb" to games that they perceive as similar, we are searching for good rules of thumb for Multi-Armed Bandits where all that is known is the set of actions and an interval containing the payoffs.

Imitating behavior or variants thereof will play a special role in our model. A behavioral rule F is *imitating* if, whenever an individual switches his strategy, he adopts the action of the individual he sampled. Formally, a behavioral rule F is *imitating* if $F(i,x,j,y)_k = 0$ whenever $k \notin \{i,j\}$. A necessary condition to employ an imitating behavior (other than *Never Switch Strategies)* is that sampled play reveals the strategy used by the opponent. Although this condition is clearly true for Single-Decision Tasks considered in this section, it is generally not true in Multi-Decision Tasks as considered in Sections 3 and 4. It is worth mentioning here that many feasible rules are not imitating (e.g., the rule F given by $F(i,x,j,y)_k = 1/|S|$ that randomizes uniformly among all possible strategies).

Some additional notation is useful here. For a given Multi-Armed Bandit with set of actions S and for a given round, the *state p of the population* describes the proportions of the population choosing the various actions in this round (notice that the dependence of p on time is denoted p^t but this dependence is often not made explicit). Thus $p \in \Delta(S) \equiv \{p \in \mathbb{R}^n \mid \sum_i p_i = 1, 0 \leq p_i \leq 1 \ \forall i \in S\}$. $C(p) = \{i \in S | p_i > 0\}$ is the *support* of the state p. State p is *interior* if $C(p) = S$. The state p^0 in round 0 is called the *initial state*. $\pi(p) = \sum_i p_i \pi_i$ denotes the average payoff in state p. Under our above assumption on information gathering, $p_i \cdot P_i(y)$ is the probability of observing in state p an individual who chose strategy i and received a payoff y.

Each behavioral rule F induces a discrete-time deterministic dynamic $(p^t)_{t=0,1,2,..}$ on $\Delta(S)$ given the interior initial state p^0 through the assumptions that all individuals use the same rule and that realized payoffs and observations are independent and at random.

This population adjustment process satisfies[6]

$$p_k^{t+1} = p_k^{t+1}(F) = \sum_{i,j \in S} \sum_{x,y \in [\alpha,\omega]} p_i^t p_j^t F(i,x,j,y)_k P_i(x) P_j(y) \quad . \tag{1}$$

Remark 1 *In the Evolutionary Game Theory literature, the adjustment process for the population is frequently postulated to have certain "short-run" (or myopic) properties instead of deducing them by aggregating individual behavior as in (1). It is commonly assumed that the dynamics are* regular selection *whose most important consequence for us is that extinct strategies remain extinct. Important subclasses are those dynamics that are* monotone *(which requires a positive correlation between relative growth rates and expected payoffs), and those that are* sign preserving *(which requires a strategy to have a positive growth rate if and only if it yields an above average payoff)*[7]*. Our work will center around the concept of improving, as defined in the following paragraph, since it can be justified on the basis of individual optimization (see also Schlag (1998b)). Improving is a weaker condition than requiring either monotonicity or sign preserving to hold in each Multi-Armed Bandit in* $MAB(S,[\alpha,\omega])$ *(see Remark 7 in the Appendix).*

In order to guide our systematic search for a good rule of thumb, we use the following partial order $\preceq$ that is based on a lexico-graphic preference relationship. Change in average payoffs, more specifically, increase in average payoffs in the population plays a central role. The first order preference in this lexico-graphic ordering is that a rule F is preferred to a rule G if F induces an increase in average payoffs more often than G does. For formal definitions, let $I(F')$ be the circumstances where average payoffs increase under the rule F'. Here, stated for Single-Decision Tasks,

$$I(F') = \left\{ \left(p^t, \left\langle (i,P_i)_{i \in S} \right\rangle\right) \in \Delta(S) \times MAB(S,[\alpha,\omega]) : \pi\left(p^{t+1}(F')\right) \geq \pi\left(p^t\right) \right\}.$$

Then $G \preceq F$ if $I(G) \subsetneq I(F)$. A rule that always induces an increase in average payoffs is called *improving* (Schlag, 1998b).[8] Only two rules F and G that are equally good according to the above ordering (i.e., $I(G) = I(F)$) are compared according to the size

[6] Discrete probability distributions P_i are assumed for simplicity, the entire analysis is easily extended to continuous distributions by taking appropriate expected values.

[7] Both these subclasses were introduced by Nachbar (1990) who used the term *relative monotone* instead of monotone. The precise definitions as well as their interrelations are described in more detail in Remark 7.

[8] An improving rule induces in each n-armed bandit a *myopic* dynamic (Swinkels, 1993) in the sense that $\pi\left(p^{t+1}\right) \geq \pi\left(p^t\right)$ holds for any p^t.

of the increase as follows. Specifically, for second order preference, F is preferred to G if F always leads to a higher increase in average payoffs than G. Combining these two criteria, we obtain the following partial order on the set of behavioral rules:

$$G \preceq F :\Leftrightarrow [I(G) \subsetneq I(F)] \vee \left\{ [I(G) = I(F)] \ \wedge \ \left[\pi\left(p^{t+1}(G)\right) \leq \pi\left(p^{t+1}(F)\right) \right] \right\}.$$

A rule that is strictly preferred to the improving rule Never Switch Strategies is improving and sometimes leads to a strict increase in average payoffs. An improving rule that leads to a strict increase in average payoffs whenever not all actions present achieve the same expected payoff is called *strictly improving* (Schlag, 1998a; called absolutely expedient by Börgers et al., 1997). The partial order $\preceq$ provides enough cutting power to select dynamically equivalent good rules of thumb when the task consists of a single decision or of multiple independent decisions (see part (iii) of Theorem 1 and Corollary 2 below respectively). In particular, these rules will be strictly improving. However, when decisions are faced in sequence in the task, we will need to extend the partial order to include a third order preference relation (see Section 4.3 below).

Maximal elements within a given set of rules D of this partial ordering will be called *undominated* in D. F is called *undominated* if D is the set of all feasible behavioral rules. So a rule F is undominated (in D) if there is no rule G such that $G \succ F$ (with $F, G \in D$) where $G \succ F$ means that $G \succeq F$ but $G \not\preceq F$. Notice that undominated rules are improving since they are preferred to Never Switch Strategies.

To increase average payoffs as often as possible ignores individual behavior to some extent. The fact that some individuals are worse off (measured by expected payoff) in the next round is ignored as long as sufficiently many are better off to make average payoffs increase. However, a justification based on individual player's concern for average payoffs can be given through slightly adjusting our model to include entry and exit (Schlag, 1998b).

Instead of short-run criteria to order behavioral rules, one might be interested in long-run performance of a rule; in particular, in how good a rule is at learning a best action. A set of states is called *asymptotically stable* if dynamic trajectories starting sufficiently close stay arbitrarily close and eventually converge to this set. It is *globally asymptotically stable* if, in addition, all trajectories starting in an interior state converge to the set.[9]

The following result of Schlag (1998b) characterizes preferred rules for Single-Decision Tasks and summarizes their properties according to the criteria introduced above.

[9] For formal definitions, see Bhatia and Szegö (1970).

Theorem 1 *(i) Improving rules for Multi-Armed Bandits are imitating. However, Imitate If Better is not improving.*

(ii) A behavioral rule F is undominated if and only if F is imitating and, for each unordered pair of different actions $\{i,j\}$,

$$F(i,x,j,y)_j - F(j,y,i,x)_i = \frac{y-x}{\omega-\alpha}. \tag{2}$$

(iii) An undominated rule is strictly improving, induces the dynamic

$$p_i^{t+1} = p_i^t + \frac{1}{\omega-\alpha}\left[\pi_i - \pi\left(p^t\right)\right] p_i^t \ , \ i \in S, \ t = 0,1,2,.. \ . \tag{3}$$

and causes the set of best actions to be globally asymptotically stable for all Multi-Armed Bandits in $MAB(S,[\alpha,\omega])$.

Equation (3) is a discrete version of the *(standard) replicator dynamic* which is the basic example of both a monotone and sign preserving selection dynamic (see Remark 1). Of course, there are other monotone and sign preserving selection dynamics beyond the replicator dynamic but the point here is that none of these can result from an undominated rule. In particular, Theorem 1 (iii) implies that undominated rules are *dynamically equivalent* in the sense that they all induce the same change of the population state.

Theorem 1 also shows that our most preferred rules are imitative. Three particular imitative rules, given by the following definition and characterized in Theorem 2, play a prominent role for both Single- and Multi-Decision Tasks.

Definition 1 *The imitating behavioral rules PIR, PRR and POR are defined as follows.*

(a) The Proportional Imitation Rule (with rate $\frac{1}{\omega-\alpha}$) *(PIR, Helbing, 1992; Schlag, 1998b) only imitates the sampled action if it realizes a higher payoff and does this with probability $\frac{1}{\omega-\alpha}(y-x)$. That is, $F(i,x,j,y)_j = \frac{1}{\omega-\alpha}(y-x)$ if $y > x$ and is 0 otherwise.*

(b) The Proportional Reviewing Rule (with rate $\frac{1}{\omega-\alpha}$) *(PRR, Björnerstedt and Weibull, 1996; Schlag, 1998b) imitates the sampled action with probability $\frac{1}{\omega-\alpha}(\omega-x)$ if own last payoff was x.*

(c) The Proportional Observation Rule (with rate $\frac{1}{\omega-\alpha}$) *(POR, Schlag, 1998a) imitates the sampled action that has payoff y with probability $\frac{1}{\omega-\alpha}(y-\alpha)$.*

Theorem 2 *(Schlag, 1998b) (i) PIR is the unique undominated rule that minimizes the probability of switching strategies.*[10]

[10] It is also the unique undominated rule that never imitates an action that achieved a lower payoff.

(ii) PRR is the unique undominated rule that does not depend on the sampled individual's payoff.

(iii) POR is the unique undominated rule that does not depend on the individual's own last payoff.[11]

3 Tasks with Unrelated Consequences

The concern in this paper is to find good learning rules for tasks that consist of multiple decisions. Throughout we maintain assumptions made on memory, information and preferences made in the previous section.

In this section, we consider only those situations where the decisions in a task have unrelated consequences. This is modelled as follows. Suppose there are m different decisions where $m \geq 2$. The task each individual faces is a game against nature where the only move nature makes is the initial one (see Figure 1 below for the extensive form of this game in a special case when $m = 2$). Specifically, at the beginning of each task, nature selects decision k with probability $q_k \geq 0$ (note $\sum_{k=1}^{m} q_k = 1$). Then the individual faces the selected decision. Each decision k is a Multi-Armed Bandit as in Section 2 with a fixed finite number n_k of possible actions S_k where $n_k \geq 2$ for at least one k.[12] The realized payoff from choice i at decision k continues to be drawn independently and at random from a probability distribution $P_i^{(k)}$ that has finite support in $\left[\alpha^{(k)}, \omega^{(k)}\right]$. Thus, an individual's plan of action in each of the m Multi-Armed Bandits (i.e., his *strategy*) can be described by $\eta = (\eta_1, ..., \eta_m) \in \times S_k$ where η_k denotes the action chosen at decision k for all $1 \leq k \leq m$. In particular, strategy and action no longer coincide as in Single-Decision Tasks. The tuple $\left\langle \left(q_k, \left(i, P_i^{(k)} \right)_{i \in S_k} \right)_{k=1,..,m} \right\rangle$ is called a *Parallel Bandit*. The set of Parallel Bandits that satisfy our restrictions on the support of the payoff distributions will be denoted by $PB\left(\left(S_k, \left[\alpha^{(k)}, \omega^{(k)} \right] \right)_{k=1,..,m} \right)$. Parallel Bandits can be represented in an extensive form as in Figure 1 where $\mathcal{L}(k, i)$ denotes the lottery with distribution $P_i^{(k)}$ induced by choosing action i in decision k. This extensive form is closely related to the usual extensive form in game theory as we will see in Section 4. The major difference is that, instead of specifying payoffs at the endnodes, the probability distribution of realized payoffs $P_i^{(k)}$ are indicated there.

[11] The proof of part (iii) is an easy exercise given the results contained in (Schlag, 1998b).

[12] That is, Section 2 concerns the case $m = 1$. In this section, we usually assume $n_k \geq 2$ for all k. However, in Section 3.1, we consider the case $m = 2$ with $n_1 = 1$ and $n_2 = 2$ (see Figure 2).

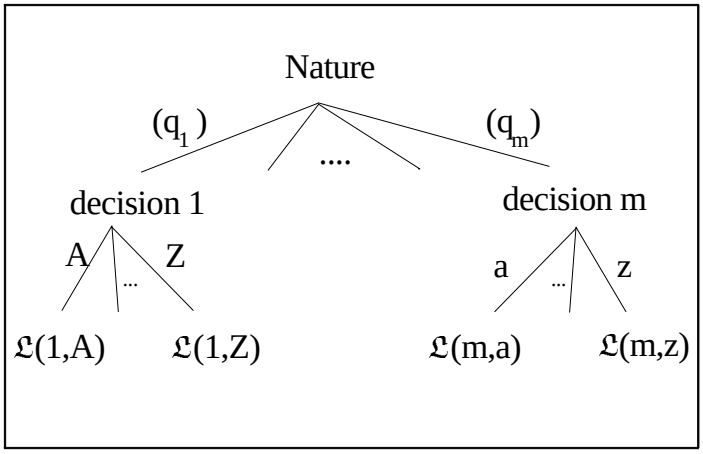

Figure 1: The extensive-form of a Parallel Bandit.

In a given Parallel Bandit, let $\pi_i^{(k)} = \sum_x x P_i^{(k)}(x)$ denote the expected payoff of choosing action i in decision k and let $\pi_\eta = \sum_{k=1}^m q_k \cdot \pi_{\eta_k}^{(k)}$ denote the expected payoff of strategy η. $\Delta\{\arg\max\{\pi_\eta : \eta \in \times S_k\}\}$ is the set of best strategies. A *state* p is an element of $\Delta(\times S_k)$. *Play* in a given task consists of the decision k faced and the action i chosen and will be denoted by i^k $(i \in S_k)$. Throughout the paper we assume that play, not strategy is sampled. Thus, play and strategy no longer necessarily coincide as in the Single-Decision Tasks of Section 2, when at least two decisions have at least two actions (i.e., $|\{k : n_k \geq 2\}| \geq 2$) they never coincide. To simplify notation we will write η^k instead of $(\eta_k)^k$ for play induced when an individual using strategy η faces decision k. Sampling remains random, the probability of sampling play i^k that yielded a payoff y is equal to $q_k \sum_{\eta:\eta_k=i} p_\eta P_i^{(k)}(y)$.

A *behavioral rule* for a Parallel Bandit is a map $F\left(\eta, k, x, j^l, y\right)_{\eta'}$ that specifies the probability an individual adopts strategy η' in the next round if own last strategy is η with payoff x obtained at decision k (from choosing action η_k) and the sampled action is j at decision l with payoff y. Consequently, $1 - F\left(\eta, k, x, j^l, y\right)_\eta$ is the probability of changing to a different strategy under these circumstances and will be denoted by $F_\circ\left(\eta, k, x, j^l, y\right)$.

Each rule induces a discrete-time deterministic dynamic on $\Delta(\times S_k)$ through our assumptions of random sampling and that all individuals use the same rule. For a given initial state $p^0 \in \Delta(\times S_k)$, the dynamic $(p^t)_{t=0,1,2,..}$ is given by the recursion

$$p_{\eta'}^{t+1} = \sum_{k,l=1}^m q_k q_l \sum_{\eta,\nu \in \times_k S} p_\eta^t p_\nu^t F\left(\eta, k, \nu^l\right)_{\eta'} \tag{4}$$

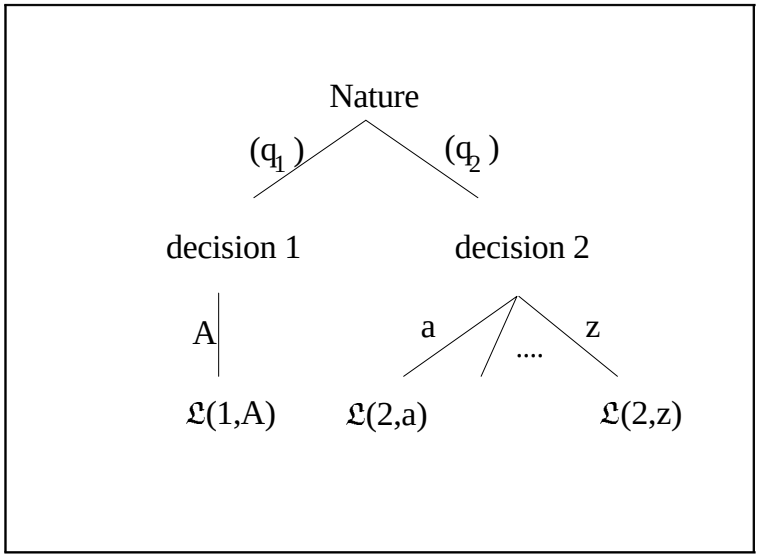

Figure 2: Simple Parallel Bandits

where $F\left(\eta,k,j^{l}\right)_{\eta'}=\sum_{x\in[\alpha^{(k)},\omega^{(k)}]}\sum_{y\in[\alpha^{(l)},\omega^{(l)}]}F(\eta,k,x,j^{l},y)_{\eta'}P_{\eta_{k}}^{(k)}(x)P_{j}^{(l)}(y)$ is the expected probability of changing from strategy η to η' if own last play was at decision k and the sampled individual is observed using action j at decision l. The notation $F_{\circ}\left(\eta,k,j^{l}\right)=1-F\left(\eta,k,j^{l}\right)_{\eta}$ also proves useful below.

3.1 Simple Parallel Bandits

We first consider a very simple Parallel Bandit with $m=2$, $n_1=1$ and $n_2=2$ as shown in Figure 2 that we call a *Simple Parallel Bandit*. Decision 1 consists of a degenerate decision with only a single action to choose from. Decision 2 is a Multi-Armed Bandit. Notice that rounds in which decision 1 is played need not be ignored since after such a play an individual might learn about decision 2 through sampling performance of someone else. Here, $(A,a)=Aa$ and $(A,z)=Az$ are the possible strategies, $\Delta\left(Aa,Az\right)$ is the set of states.

Since play sampled at decision 2 reveals the strategy used by the sampled individual, a behavioral rule for a Simple Parallel Bandit is imitative if the individual never switches when he samples play in decision 1 and only switches to the strategy used by the sampled individual when sampling play in decision 2.

A natural question at this point is whether good rules of thumb exist for Simple Parallel Bandits. Perhaps the most intuitive rule is the following:

(a) Do not change strategies if own last play and sampled play did not both occur at decision 2, i.e., $F\left(Ai,k,x,j^l,y\right)_{Ai}=1$ if $\{k,l\}\neq\{2\}$.
(b) If own last play and sampled play both occurred at decision 2, then act as if facing a Single-Decision Task consisting of decision 2. Choose an undominated rule such as PIR from Section 2 to determine whether the strategy of the sampled individual should be imitated. Under PIR we obtain $F\left(Ai,2,x,j^2,y\right)_{Aj}=1-F\left(Ai,2,x,j^2,y\right)_{Ai}=\max\left\{\frac{y-x}{\omega^{(2)}-\alpha^{(2)}},0\right\}$, if $i\neq j$.
Such a rule does not attempt to use observed information across two different decisions (*cross-decision learning*). Rather, individuals wait until they receive information from the same decision (*within-decision learning*). We will see that there are better rules that make use of cross-decision learning.

The rule described above utilizes results from Single-Decision Tasks whenever the individual finds himself facing such a situation. In particular, a Simple Parallel Bandit turns into a Single-Decision Task when $q_1=0$. Moreover, besides the fact that the individual knows of the existence of decision 1, within-decision learning in decision 2 cannot be distinguished from learning in a Single-Decision Task. Rules that apply undominated rules from simpler settings whenever possible will play an important role in our analysis. These rules will be called *consistent*. In Simple Parallel Bandits, a rule is consistent if it satisfies condition (b) above.

The following result characterizing undominated rules is intuitive. Experience about payoffs in decision 1 are meaningless in learning the best strategy (i.e., the best action at decision 2). Hence, only observed payoffs in decision 2 matter for updating behavior. By using POR under cross-decision learning whenever sampled play occurred at decision 2, own experience does not influence switching.

Theorem 3 *Consider Simple Parallel Bandits with $S_1=\{A\}$ and $m=2$. The behavioral rule F is undominated if and only if*

(i) it is imitating,

(ii) it is consistent, i.e., it uses an undominated rule from Single-Decision Tasks to determine switching probabilities under within-decision learning and so

$$F_\circ\left(Ai,2,x,j^2,y\right)-F_\circ\left(Aj,2,y,i^2,x\right)=\frac{y-x}{\omega^{(2)}-\alpha^{(2)}}\text{ for }i\neq j,$$

and,

(iii) it uses the undominated POR (with rate $\frac{1}{\omega^{(2)}-\alpha^{(2)}}$) to determine switching proba-

bilities under cross-decision learning and so

$$F_{\circ}\left(Ai,1,x,j^{2},y\right)=\frac{y-\alpha^{(2)}}{\omega^{(2)}-\alpha^{(2)}}\ \textit{for}\ i\neq j.$$

Corollary 1 *Undominated rules are strictly improving, induce the dynamic*

$$p_{Ai}^{t+1}=p_{Ai}^{t}+\frac{1}{\omega^{(2)}-\alpha^{(2)}}\left[\pi_{Ai}-\pi\left(p^{t}\right)\right]p_{Ai}^{t},\ i\in S_{2}\ . \tag{5}$$

and cause the set of best strategies to be globally asymptotically stable in any Simple Parallel Bandit in $PB\left(\left(\{A\},\left[\alpha^{(1)},\omega^{(1)}\right]\right),\left(S_{2},\left[\alpha^{(2)},\omega^{(2)}\right]\right)\right)$.

It is interesting to note that perfect information cannot make up for ignorance of the extensive-form character of Simple Parallel Bandits by the following argument. Suppose an individual had perfect information about the strategy Ai used by the sampled individual even when decision 1 is sampled. The Simple Parallel Bandit could then be treated as a Single-Task Multi-Armed Bandit $\left\langle\left(Ai,q_{1}P_{A}^{(1)}+q_{2}P_{i}^{(2)}\right)_{i\in S_{2}}\right\rangle$ with strategy set $\{Ai,i\in S_{2}\}$ and payoffs contained in $\left[\min\{\alpha_{1},\alpha_{2}\},\max\{\omega_{1},\omega_{2}\}\right]$. Using an undominated rule for this multi-armed bandit induces (5) except that the factor $1/\left(\omega^{(2)}-\alpha^{(2)}\right)$ must be replaced by the generally smaller term $1/\left(\max\{\omega_{1},\omega_{2}\}-\min\{\alpha_{1},\alpha_{2}\}\right)$. Thus, undominated rules under imperfect information (as characterized in Theorem 3) are preferred by our second order preference relation to any rule under complete information that treats the Simple Parallel Bandit as a Single-Decision Task.

In analogy to Theorem 2 (i), there is a unique undominated rule that minimizes the probability of switching strategies; it uses PIR for within-decision learning. However, an analogue of Theorem 2 (ii) no longer holds since any undominated rule must utilize information from sampled payoff. On the other hand, in a setting where only own payoffs are observable, strictly improving rules still exist. For instance, the consistent rule that is imitating, uses PRR under-within decision learning and does not switch under cross-decision learning (a rule included in the description of (a) and (b) above) is strictly improving. This rule induces a dynamic as in (5) except that the right hand side receives an additional factor of q_2 since an individual only switches after experiencing decision 2. It is easily shown that this rule is undominated among the rules that do not depend on sampled payoff.

Undominated rules are consistent and ignore own payoffs under cross-decision learning. How do other consistent rules perform that utilize own payoffs under cross-decision learning? PIR is not an effective rule to determine cross-decision switching probabilities

since it can cause each individual to adopt the worst strategy in the long run as follows. Given $S_2 = \{a, z\}$, consistent rules that are imitating induce the dynamic

$$p_{Aa}^{t+1} = p_{Aa}^t + p_{Aa}^t p_{Az}^t q_2 \left[q_1 \left(F_\circ \left(Az, 1, a^2 \right) - F_\circ \left(Aa, 1, z^2 \right) \right) + q_2 \frac{1}{\omega^{(2)} - \alpha^{(2)}} \left(\pi_a^{(2)} - \pi_z^{(2)} \right) \right] .$$

Given $\alpha_1 = \alpha_2 = 0$, $\omega_1 = \omega_2 = 1$, $P_A^{(1)}(0.5) = P_a^{(2)}(0.6) = 1$, $P_z^{(2)}(0) = P_z^{(2)}(1) = 0.5$ we obtain $F_\circ \left(Az, 1, a^2 \right) - F_\circ \left(Aa, 1, z^2 \right) = -0.15$ under PIR for cross-decision learning and hence $p_{Aa}^t \to 0$ if q_1 is sufficiently large although $\pi_{Aa} > \pi_{Az}$. If instead, PRR is used under cross-decision learning, then $F_\circ \left(Az, 1, a^2 \right) - F_\circ \left(Aa, 1, z^2 \right) = 0$. This alternative rule leads to the same dynamic as a consistent rule that does not switch under cross-decision learning and so cannot be undominated.

3.2 General Parallel Bandits

Next we consider behavioral rules for more general Parallel Bandits where there are at least two decisions with non trivial choices.

Example 1 *A particular Parallel Bandit we will use repeatedly in the demonstration of proofs and results is given in Figure 3 where* $S_1 = \{A, Z\}$, $S_2 = \{a, z\}$ *and* $m = 2$. *Abbreviating notation again, the set of strategies is* $S_1 \times S_2 = \{Aa, Az, Za, Zz\}$, states *are denoted by* $p = (p_{Aa}, p_{Az}, p_{Za}, p_{Zz}) \in \Delta(S_1 \times S_2)$. *Here we see why, in contrast to Single-Decision Tasks, play and strategy never coincide when there are at least two decisions with at least two actions each. For example, if sampled play is* z^2, *then this individual could be using either* Az *or* Zz. *Imitating behavior no longer exists except for the degenerate rule Never Switch Strategies.*

Imitation has been an important property of good (i.e., improving) rules in avoiding a decrease in average payoffs (Theorem 1 (i), Corollary 1). The natural generalization of imitative behavior to the present setting with imperfect information occurs through imitating play. An individual who plans on changing his strategy adopts (or pastes) the sampled action at the sampled decision and keeps his previous action at all other decisions. Formally,

Definition 2 *A behavioral rule* F *is* play-wise imitating *if the strategy played in the next round differs at most from the previous strategy by play at sampled decisions. For Parallel Bandits, this means that* $F\left(\eta, k, x, j^l, y\right)_{\eta'} = 0$ *unless* $\eta = \eta'$ *or* $\eta' = \eta \backslash j^l$ *where*

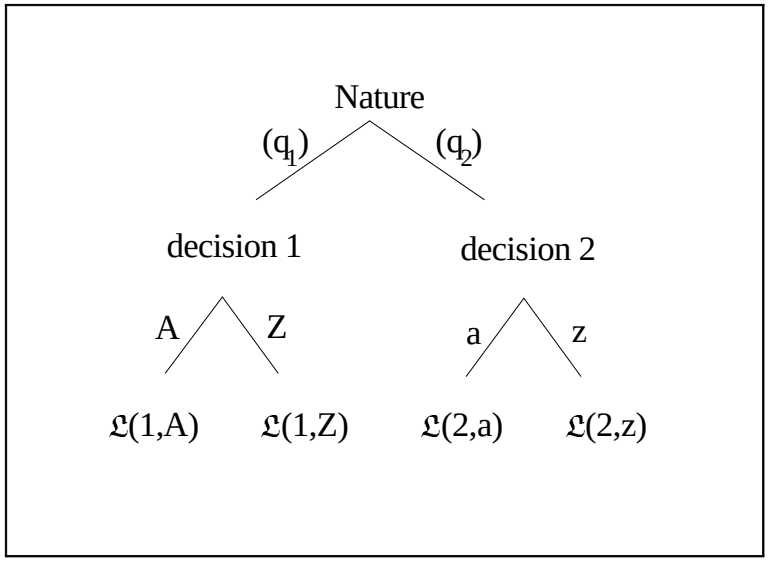

Figure 3: The extensive form of a Parallel Bandit with two decisions with two actions each.

$\eta\backslash j^l$ is the same strategy as η except at decision l where action j is chosen. In other words, $F_\circ\left(\eta,k,x,j^l,y\right)=F\left(\eta,k,x,j^l,y\right)_{\eta\backslash j^l}$ whenever $\eta_l\neq j$ and $F_\circ\left(\eta,k,x,j^l,y\right)=0$ otherwise.

The proof of Theorem 3 (i) in the Appendix shows that

Theorem 4 *Improving rules for Parallel Bandits are play-wise imitating.*

As the tasks become more complex with multiple decisions in parallel and/or in sequence, the set of feasible behavioral rules grows rapidly. We want to avoid specifying a separate rule for each possible task and search for characteristics of a rule that can be applied generally. Our method is to focus on behavioral rules that can be derived from (or are consistent with) undominated rules for simpler settings. Such consistency has already been pointed out for undominated rules of Simple Parallel Bandits. The individual behaves as he would in the simplest task that he could be facing given his present observations. Consistency in Parallel Bandits places the following restrictions on behavioral rules. Under within-decision learning in decision l, it is as if only decision l can occur and the task is a Multi-Armed Bandit. Under cross-decision learning, experiencing decision l and sampling play in decision k, it is as if only decision k and l can occur and the task is a Simple Parallel Bandit with own action at decision l the only feasible action.

These conditions on within- and cross-decision learning are summarized in (i) and (ii) respectively in the following definition.

Definition 3 *A behavioral rule F for Parallel Bandits in PB $\left(\left(S_k, \left[\alpha^{(k)}, \omega^{(k)}\right]\right)_{k=1,..,m}\right)$ is called* consistent *if:*

(i) Under within-decision learning in decision l, if $\eta\backslash i^l$ is identified with $\nu\backslash i^l$ for all $\eta, \nu \in \times S_k$ and $i \in S_l$ then the switching behavior of F is that of an undominated rule in the Multi-Armed Bandit $\left\langle \left(i, P_i^{(l)}\right)_{i\in S_l} \right\rangle$.

(ii) Under cross-decision learning, experiencing play i^l and sampling play in decision k, if $\nu\backslash j^k$ is identified with $\nu'\backslash j^k$ for all $j \in S_k$ and all ν, ν' with $\nu_l = \nu'_l = i$, then switching behavior of F is that of an undominated rule in the Simple Parallel Bandit $\left\langle \left(q_l, \left(i, P_i^{(l)}\right)\right), \left(1-q_l, \left(j, P_j^{(k)}\right)_{j\in S_k}\right)\right\rangle$.

The following result generalizes Theorem 3 to all Parallel Bandits by characterizing the undominated rules that are also consistent. However, unlike Theorem 3, there remains the possibility that undominated rules exist that are not consistent.

Theorem 5 *For a behavioral rule F, statement (a) is equivalent to statements (b1) - (b3):*
(a) F is a consistent rule that is undominated.
(b1) F is play-wise imitating,
(b2) for within-decision learning, F uses an undominated rule among the strictly improving rules in Single-Decision Tasks to determine switching probabilities, i.e.,

$$F_\circ\left(\eta\backslash i^k, k, x, j^k, y\right) - F_\circ\left(\eta\backslash j^k, k, y, i^k, x\right) = \frac{y-x}{\omega^{(k)} - \alpha^{(k)}} \text{ , and,}$$

(b3) for cross-decision learning, F uses the undominated POR to determine switching probabilities, i.e.,

$$F_\circ\left(\eta\backslash i^k, l, x, j^k, y\right) = \frac{y-\alpha^{(k)}}{\omega^{(k)} - \alpha^{(k)}} \text{ for all } i \neq j \text{ and } k \neq l.$$

The strategy dynamic (4) for a play-wise imitating rule F is

$$\begin{aligned} p_\eta^{t+1} &= p_\eta^t + \sum_k q_k \sum_{j\in S_k\backslash\{\eta_k\}} p^t_{\eta\backslash j^k} \cdot p^{(k)}_{\eta_k}(t) \cdot \sum_{l=1}^m q_l F_\circ\left(\eta\backslash j^k, l, \eta^k\right) \\ &\quad - p_\eta^t \sum_{s,l=1}^m q_s q_l \sum_{j\in S_l\backslash\{\eta_l\}} p_j^{(l)}(t) \cdot F_\circ\left(\eta, s, j^l\right) \end{aligned}$$

which, for an undominated consistent rule, simplifies somewhat to

$$p_\eta^{t+1} = p_\eta^t + \sum_k q_k \sum_{j \in S_k \setminus \{\eta_k\}} p_{\eta \setminus j^k}^t \cdot p_{\eta_k}^{(k)}(t) \frac{\pi_{\eta_k}^{(k)} - \alpha^{(k)}}{\omega^{(k)} - \alpha^{(k)}} - p_\eta^t \sum_{l=1}^m q_l \sum_{j \in S_l \setminus \{\eta_l\}} p_j^{(l)}(t) \frac{\pi_j^{(l)} - \alpha^{(l)}}{\omega^{(l)} - \alpha^{(l)}}$$

where $p_i^{(k)}(t)$ is the frequency of individuals who plan to use action $i \in S_k$ at decision k at time t; namely,

$$p_i^{(k)}(t) = \sum_{\{\nu | \nu_k = i\}} p_\nu^t.$$

Unlike Single-Decision Tasks or Simple Parallel Bandits, these strategy dynamics are not directly related to the replicator dynamic for the game against nature described by a Parallel Bandit. In particular, extinct strategies do not remain extinct (see also Remark 1). For instance, in Example 1, undominated consistent rules cause some individuals to use either Az or Za in round 1 even if Aa and Zz are the only strategies used in round 0.

On the other hand, the population adjustment process simplifies considerably if we restrict attention to evolution of $p_i^{(k)}(t)$ at each decision k. The following is an immediate consequence of Theorem 5 and the above equations.

Corollary 2 *Undominated consistent rules are strictly improving and induce the replicator dynamic at each decision since*

$$p_i^{(k)}(t+1) = p_i^{(k)}(t) + \frac{q_k}{\omega^{(k)} - \alpha^{(k)}} \left[\pi_i^{(k)} - \sum_{j \in S_k} p_j^{(k)}(t) \cdot \pi_j^{(k)} \right] p_i^{(k)}(t) \quad , \; i \in S_k \; . \tag{6}$$

The set of best strategies is globally asymptotically stable in any Parallel Bandit belonging to $PB\left(\left(S_k, \left[\alpha^{(k)}, \omega^{(k)}\right]\right)_{k=1,..,m}\right)$.

These fundamental properties of undominated consistent rules stand in stark contrast to the performance of alternative rules used in the literature. Whenever switching strategies, the behavioral rule used by Ruebeck (1997)[13] randomizes among the strategies the sampled individual could be using. For the task given in Example 1, this means that $F(Aa, 1, z^2)_{Az} = F(Aa, 1, z^2)_{Zz} = \frac{1}{2} F_\circ(Aa, 1, z^2)$. This rule is not play-wise imitating and hence fails to be strictly improving. In fact, when payoffs are given through unknown lotteries, the best strategy will typically not be asymptotically stable.[14]

[13] See also Cooper (1997).

[14] For this rule, when Az is the best strategy, $F_\circ(Az, 1, x, z^2, y) > 0$ for some x and y with $P_A^1(x) \cdot P_z^2(y) > 0$ and $0 < q_1 < 1$, $\{p_{Az} = 1\}$ is not even a *rest point* since $p^{t+1} \neq p^t$ when $p_{Az}^t = 1$.

3.3 Monotonicity

As mentioned in Section 2, monotonicity and related conditions (see Remark 1) are important intuitive properties of population adjustment processes in the Evolutionary Game Theory literature. Even though some individuals could switch to a worse strategy in any Single-Decision Task, the undominated rules always increase the frequency of the best strategy. We say such a rule induces a *best reply monotone* dynamic. Formally, a dynamic $p^{t+1} = f(p^t)$ is called best reply monotone if the frequency of the best strategies $B = \arg\max\{\pi_\nu : \nu \in \times S_k\}$ increases strictly in each interior state (i.e., $\sum_{\eta\in B} p_\eta^{t+1} > \sum_{\eta\in B} p_\eta^t$ whenever $C(p^t) = \times S_k$).

The reason best reply monotone dynamics based on individual behavioral rules can exist is that any switch to a worse strategy is offset by the sampled individual switching with higher probability. It is clear from (6) that any undominated rule for a Parallel Bandit increases the frequency of the best action in each decision. Nevertheless, the adjustment dynamic is not necessarily best reply monotone. In fact, there is no rule that guarantees best reply monotonicity (Theorem 6 below) due to the inability to sample strategies. That is, offsetting behavior found in Single-Decision Tasks cannot be used here to guarantee best reply monotonicity.

Theorem 6 *Assume that* $|S_1|, |S_2| \geq 2$. *There is no behavioral rule that induces a best reply monotone dynamic in all Parallel Bandits in* $PB\left(\left(S_k, \left[\alpha^{(k)}, \omega^{(k)}\right]\right)_{k=1,..,m}\right)$.

Proof. Assume that the behavioral rule F always induces best reply monotone trajectories. Following the arguments made in the first paragraph of the proof of Theorem 3 in the Appendix and using the fact that p^{t+1} depends continuously on p^t from (4), we obtain that F is play-wise imitating.

Assume that $m = 2$, $S_1 = \{A, Z\}$ and $S_2 = \{a, z\}$ as in Example 1. By best reply monotonicity, we can find a strategy μ, payoffs $x \in \left[\alpha^{(k)}, \omega^{(k)}\right]$ and $y \in \left[\alpha^{(l)}, \omega^{(l)}\right]$ with $k \neq l$ and an action $j \in S_l$ such that $F_\circ\left(\mu, k, x, j^l, y\right) > 0$. Then there are payoff distributions such that $F_\circ\left(\mu, k, j^l\right) > 0$ and $\pi_\mu > \pi_\eta$ for all strategies η not equal to μ. Without loss of generality, assume that $\mu = Aa$ is the unique best strategy. Consider p^t such that $C(p^t) = \{Aa, Zz\}$. Play-wise imitating implies $F(Aa, \cdot, \cdot)_{Zz} = 0$. This together with $F_\circ\left(Aa, i, j^k\right) > 0$ implies $p_{Aa}^{t+1} < p_{Aa}^t$. Continuity implies that $p_{Aa}^{t+1} < p_{Aa}^t$ will also hold when $p_{Za}^t > 0$ and $p_{Zz}^t > 0$ are sufficiently small. This contradicts best reply monotonicity. The proof for more general Parallel Bandits is analogous. ■

Since both monotone and sign preserving dynamics are best reply monotone, Theorem 6 raises the question as to whether such dynamics are realistic models of aggregate population behavior in multi-decision games against nature. Recall, from Remark 1, that monotone dynamics (e.g., the replicator dynamic) are commonly used in Evolutionary Game Theory to model a population that is myopically learning the best strategy by increasing the relative frequency of those strategies with higher payoffs. A more realistic expectation might be that the relative frequencies of actions with higher payoffs at a given decision should increase for good rules of thumb. That is, for any decision k for which $q_k > 0$, and any $i, j \in S_k$ for which $p_i^{(k)}(t)$ and $p_j^{(k)}(t)$ are both positive, we might expect

$$\frac{p_i^{(k)}(t+1)}{p_i^{(k)}(t)} > \frac{p_j^{(k)}(t+1)}{p_j^{(k)}(t)} \text{ if and only if } \pi_i^{(k)} > \pi_j^{(k)} \ . \tag{7}$$

Following (6), undominated consistent rules have this *decision-wise monotonicity* property.[15]

4 A Task involving Sequential Decisions

In this section, we consider tasks involving multiple decisions with related consequences. Choices at earlier decisions may influence decisions faced later in the same task. Here we restrict attention to the simple tasks illustrated in Figure 4 that we call *Centipede Bandits* (of length three). Nature moves down with probability q (and across otherwise, $0 \leq q \leq 1$) after a player choice of A at decision 1. We hasten to point out that the analysis of this specific task yields behavioral rules for general tasks involving sequential decisions (see Conclusion). Again, $\mathcal{L}(k,i)$ denotes the lottery with distribution $P_i^{(k)}$ that is encountered at decision k when action i is taken. The four probability distributions at the endnodes are assumed to be independent and to have finite support in $[\alpha, \omega]$. Their corresponding expected values are denoted $\pi_D^{(1)}$, $\pi_n^{(0)}$, $\pi_d^{(2)}$ and $\pi_a^{(2)}$ respectively. As before, the only information[16] an individual has is his last strategy, play and payoff as well as the play and payoff of one other individual sampled at random.

The set of strategies of an individual facing the Centipede Bandit is $\{Dd, Da, Ad, Aa\}$, where, for instance, Dd has the meaning (as in standard game theory) that player

[15] (7) can be considered a discrete version of the subgame monotonicity condition introduced by Cressman (1996) for general extensive form games.

[16] besides the knowledge of the actions and tree structure given in Figure 4, that the payoff distributions have support $[\alpha, \omega]$ and that the parameter q is in $[0, 1]$

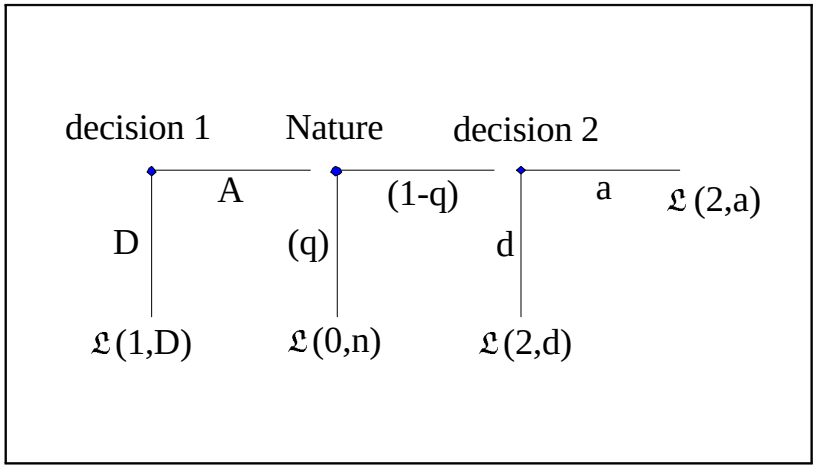

Figure 4: Centipede Bandit

one will choose D (d) if asked to move at decision 1 (2). In the following, let $\pi_D = \pi_{Di} = \pi_D^{(1)}$, $\pi_{Ai} = q\pi_n^{(0)} + (1-q)\pi_i^{(2)}$ for $i \in \{d, a\}$. The set of best strategies equals $\Delta\{\arg\max\{\pi_\eta : \eta \in \{Dd, Da, Ad, Aa\}\}\}$. Let $\pi(p) = (p_{Dd} + p_{Da})\pi_D + p_{Ad}\pi_{Ad} + p_{Aa}\pi_{Aa}$ denote the average payoff in state $p \in \Delta(\{Dd, Da, Ad, Aa\})$.

For a behavioral rule F (i.e., a specification of the switching probabilities among the individual's four strategies), experienced and sampled play will be described by the endnode in $\{D, n, d, a\}$ reached. We let $F(Ai, e, x, e', y)_\nu$ denote the probability of adopting strategy ν in the next round after playing strategy Ai, obtaining payoff x at the endnode $e \in \{n, i\}$ and sampling an individual who achieved payoff y at endnode $e' \in \{D, n, d, a\}$. When an individual is using strategy Di, we will abbreviate the notation somewhat by eliminating e in $F(Di, e, x, e', y)_\nu$ since the endnode D is trivially reached and let $F(Di, x, e', y)_\nu$ denote the corresponding probability.

In analogy to Parallel Bandits,

Definition 4 *A behavioral rule F is* play-wise imitating *if the strategy played in the next round can only differ from previous strategy by pasting sampled actions at some or all of the sampled decisions. For Centipede Bandits this means that, for $e, e' \in \{n, i\}$ and $i \in \{d, a\}$,*

$$F(Di, x, D, y)_{Di} = F(Ai, e, x, e', y)_{Ai} = 1$$
$$F(Di, x, n, y)_{Di} + F(Di, x, n, y)_{Ai} = 1$$
$$F(Ai, e, x, D, y)_{Ai} + F(Ai, e, x, D, y)_{Di} = 1$$

The following theorem clearly illustrates that sequential decision making is indeed

more complex than making choices in a Parallel Bandit. In particular, part (i) shows that the improving condition is too stringent since it implies that there can then be no learning in decision 1. Following part (ii), play-wise imitating remains necessary to avoid making individuals worse off in "ideal" population states where each is playing a best strategy and there are no plans at unreachable decisions to choose a bad action.

Theorem 7 *(i) There is no strictly improving rule. An improving rule (in particular, an undominated rule) for Centipede Bandits always maintains the same action at decision* 1*. There is no behavioral rule that, in each Centipede Bandit, makes the set of best strategies asymptotically stable or induces a best reply monotone dynamic.*

(ii) If average payoffs never decrease when currently all individuals are choosing a best strategy and plan to choose a best action at decision 2 *(even if this decision is unreachable), then the underlying rule is play-wise imitating.*

Proof. Assume that F is improving. Consider a Centipede Bandit with $\pi_D = \pi_{Ad} > \pi_{Aa}$ and a state p^t in which $p^t_{Aa} = 0$ and $p^t_{Di} > 0$. Then improving implies $p^{t+1}_{Aa} = 0$. This however implies that $F(Di, x, n, y)_{Aa} = F(Di, x, D, y)_{Aa} = 0$. Symmetric arguments, interchanging the roles of d and a, imply $F(Di, x, n, y)_{Ad} = F(Di, x, D, y)_{Ad} = 0$.

Consider now a Centipede Bandit with expected payoffs satisfying $\pi_D < \pi_{Ad} = \pi_{Aa}$. Following the above argument, when $q = 1$ and $p^t \in \Delta^4$ is an interior state, no one using Dd or Da will switch to A at decision 1 and hence the rule fails to be strictly improving.

The same arguments lead to a contradiction when F always induces a best reply monotone dynamic or always causes the set of best strategies to be asymptotically stable.

For part (ii), assume that $\pi_D = \pi_{Ad} > \pi_{Aa}$ and consider a state with $C(p^t) = \{Dd, Ad\}$. Then average payoffs can strictly decrease unless F is play-wise imitating. ■

As seen in the above proof, the added complexity arises from not knowing what to play at decision 2 after sampling a favorable outcome at endnode n and hence switching from D to A at decision 1. Here play-wise imitating implies the switch must be to Ai where i is the second component of the strategy Di. The condition of improving applies to any state, in particular, to any initial state where individuals have no previous experience of the task. In these initial situations, it is intuitively clear that it might take a rule a number of rounds to adjust the additional memory implicit in the relative play of Dd and Da before it can perform well. However, after a finite number of rounds, one expects a good rule to have adjusted memory in the population in such a way as to cause average

payoffs to increase strictly. In Section 4.4 below, we will see that this intuitive condition can be guaranteed. However we have to give up comparing such a rule to all other rules. Otherwise, such a rule must be preferred to Never Switch Strategies which means that it is improving. Following the proof of Theorem 7 (i), improving rules never increase average payoffs when nature always moves down.

4.1 Consistent Rules and Short-Run Population Dynamics

As in Parallel Bandits, we will restrict attention to consistent rules. Again, an individual behaves as he would in the simplest task that he could be facing given his present information. Whenever n is not the experienced or sampled action, the individual behaves as if nature never moves down and chooses an undominated rule for the three-armed bandit that is induced when $q = 0$. Similarly, whenever individuals using different actions at decision 1 sample each other with nature always moving down after A, behavior follows an undominated rule for the two-armed bandit induced when $q = 1$. These are our first two conditions (a) and (b) of consistency. Condition (c) concerns situations where sampled play and own intended play are as if action $j \in \{d, a\}$ at decision 2 does not exist whereas (d) applies when it is as if action D at decision 1 does not exist.

Definition 5 *A behavioral rule F for Centipede Bandits is* consistent *if:*

(a) Whenever the move down by nature is neither experienced nor sampled, if Dd and Da are identified with D, then switching behavior of F is that of an undominated rule in the three-armed bandit $\left\langle \left(D, P_D^{(1)}\right), \left(Ad, P_d^{(2)}\right), \left(Aa, P_a^{(2)}\right) \right\rangle$.

(b) Whenever the move across by nature is neither experienced nor sampled, if Dd and Da are identified with D, as well as Ad and Aa identified with A, then the switching behavior of F is that of an undominated rule in the two-armed bandit $\left\langle \left(D, P_D^{(1)}\right), \left(A, P_n^{(0)}\right) \right\rangle$.

(c) Whenever action j at decision 2 is neither planned nor sampled, then the switching behavior of F is that of an undominated rule in the single-decision two-armed bandit $\left\langle \left(Di, P_D^{(1)}\right), \left(Ai, qP_n^{(0)} + (1-q)P_i^{(2)}\right) \right\rangle$ *where* $\{i, j\} = \{d, a\}$.

(d) Whenever action D at decision 1 is neither experienced nor sampled, then the switching behavior of F is that of an undominated rule in the Simple Parallel Bandit $\left\langle \left(q, \left(A, P_n^{(0)}\right)\right), \left(1-q, \left(d, P_d^{(2)}\right), \left(a, P_a^{(2)}\right)\right) \right\rangle$.

Theorem 8 *A behavioral rule F for Centipede Bandits is consistent if and only if F is play-wise imitating, $F(Di, x, j, y)_{Ai} = 0$ for $\{i, j\} = \{d, a\}$, and there exist functions*

$f_1, f_2 : [\alpha, \omega] \to [0,1]$ *satisfying*

$$f_r(x,y) - f_r(y,x) = \frac{y-x}{\omega - \alpha} \text{ for } r \in \{1,2\} \tag{8}$$

such that,

$$F(Di,x,n,y)_{Ai} = F(Di,x,i,y)_{Ai} = F(Di,x,j,y)_{Aj} = f_1(x,y)$$
$$F(Di,x,n,y)_{Di} = F(Di,x,i,y)_{Di} = F(Di,x,j,y)_{Di} + F(Di,x,j,y)_{Dj} = 1 - f_1(x,y)$$
$$F(Ai,e,x,D,y)_{Di} = 1 - F(Ai,e,x,D,y)_{Di} = f_1(x,y) \text{ for } e \in \{n,i\}$$
$$F(Ai,n,x,j,y)_{Aj} = 1 - F(Ai,n,x,j,y)_{Ai} = \tfrac{y-\alpha}{\omega-\alpha}$$
$$F(Ai,i,x,j,y)_{Aj} = 1 - F(Ai,i,x,j,y)_{Ai} = f_2(x,y)$$

Consistent rules imitate play-wise in a particular fashion. An action observed at an earlier decision on a sampled path is never adopted without also imitating the action at the later decision (since $F(Di,x,j,y)_{Ai} = 0$ for $\{i,j\} = \{d,a\}$). It is in the sense that consistent rules have the flavor of *learning backwards in the tree.*

In the following we search among the consistent rules for a good rule of thumb. This means that we accept that average payoffs will sometimes decrease. Imposing consistency considerably reduces the set of rules under consideration. The only flexibility in specifying a consistent rule is the exact forms of the functions f_1 and f_2 satisfying (8) and the probabilities $F(Di,x,j,y)_{Di}$ and $F(Di,x,j,y)_{Dj}$ which must satisfy $F(Di,x,j,y)_{Dd} + F(Di,x,j,y)_{Da} = 1 - f_1(x,y)$ for $\{i,j\} = \{d,a\}$. The specific form of f_2 will turn out to have no effect on population dynamics (see Remark 2 below). f_1 affects how often implicit memory of i in Di is changed into play at a reachable node through switching from D to A at decision 1. Relying least often on such memory will have the most preferable properties (Theorem 9 below).

Finally, $F(Di,x,j,y)_{Dj}$ determines how implicit memory of i in Di is managed whenever planned play through choosing D is not changed. To point out the two extreme possibilities, we will speak of *Minimal Switch* if only entire sampled play is imitated, i.e., $F(Di,x,j,y)_{Dj} = 0$; we will speak of *Maximal Switch* if sampled action at decision 2 is always imitated, i.e., $F(Di,x,j,y)_{Dj} = 1 - F(Di,x,j,y)_{Aj}$. Another alternative, that will prove important for our analysis, is *Proportional Switch* where $F(Di,x,j,y)_{Dj} + F(Di,x,j,y)_{Aj} = (y-\alpha)/(\omega-\alpha)$. Using Proportional Switch, adopt the choice of sampled individual at decision 2 with probability given by POR, and, given change of own action at decision 2, adopt sampled choice at decision 1 with probability $F(Di,x,j,y)_{Aj} = f_1(x,y)$. It will turn out that Minimal Switch has the least desirable properties among

these three alternatives for adjusting play at unreachable nodes (see Theorem 10 below).[17]

We now develop the population dynamic (equations (9) to (13) below) generated when all individuals use the same consistent behavioral rule F. First, consider the change in frequency of strategy Dd from one round to the next. Change among individuals using Dd and sampling an individual with strategy Ad (even if action d is not sampled) or vice versa contribute the term $p^t_{Ad}p^t_{Dd}\frac{1}{\omega-\alpha}\left[\pi^{(1)}_D - q\pi^{(0)}_n - (1-q)\,\pi^{(2)}_d\right]$ to (9). Only an increase in play of Dd is possible when Da observes Ad or vice versa, the corresponding term is

$$\begin{aligned} & p^t_{Ad}p^t_{Da}\left[qF\left(Ad,n,D\right)_{Dd} + (1-q)\,F\left(Ad,d,D\right)_{Dd} + (1-q)\,F(Da,d)_{Dd}\right] \\ = \;& p^t_{Ad}p^t_{Da}\left[\frac{1}{\omega-\alpha}\left[\pi^{(1)}_D - q\pi^{(0)}_n - (1-q)\,\pi^{(2)}_d\right] + qF\left(Dd,n\right)_{Ad} + (1-q)\,F_\circ\left(Da,d\right)\right] \end{aligned}$$

where[18]

$$\begin{aligned} F(Di,e)_\nu &= \sum_{x\in[\alpha,\omega]}\sum_{y\in[\alpha,\omega]} P^{(1)}_D(x)P_e(y)F(Di,x,e,y)_\nu \text{ and} \\ F_\circ\left(Di,e\right) &= 1 - F\left(Di,e\right)_{Di} \text{ for } e\in\{n,d,a\} \\ F(Ai,e',D)_\nu &= \sum_{x\in[\alpha,\omega]}\sum_{y\in[\alpha,\omega]} P_{e'}(x)P^{(1)}_D(y)F(Ai,e',x,D,y)_\nu \text{ for } e'\in\{n,i\}\,. \end{aligned}$$

Only a decrease in play of Dd is possible when Dd players sample Aa players. The relevant term is $-p^t_{Dd}p^t_{Aa}\left[qF_\circ\left(Dd,n\right) + (1-q)\,F_\circ\left(Dd,a\right)\right]$. Since none of the other possible sampling configurations contribute to a change of Dd, we obtain

$$\begin{aligned} p^{t+1}_{Dd} - p^t_{Dd} &= p^t_{Ad}p^t_{Dd}\frac{1}{\omega-\alpha}\left[\pi^{(1)}_D - q\pi^{(0)}_n - (1-q)\,\pi^{(2)}_d\right] \\ &\quad + p^t_{Ad}p^t_{Da}\left[qF\left(Ad,n,D\right)_{Dd} + (1-q)\,F\left(Ad,d,D\right)_{Dd} + (1-q)\,F(Da,d)_{Dd}\right] \\ &\quad - p^t_{Dd}p^t_{Aa}\left[qF_\circ\left(Dd,n\right) + (1-q)\,F_\circ\left(Dd,a\right)\right] \\ &= p^t_{Ad}\left(p^t_{Dd}+p^t_{Da}\right)\frac{1}{\omega-\alpha}\left(\pi_D - \pi_{Ad}\right) - qw^tF\left(Dd,n\right)_{Ad} \\ &\quad + p^t_{Da}p^t_{Ad}\left(1-q\right)F_\circ\left(Da,d\right) - p^t_{Dd}p^t_{Aa}\left(1-q\right)F_\circ\left(Dd,a\right) \end{aligned} \tag{9}$$

where

$$w^t = p^t_{Dd}p^t_{Aa} - p^t_{Da}p^t_{Ad}. \tag{10}$$

By a similar calculation,

$$p^{t+1}_{Ad} - p^t_{Ad} = p^t_{Dd}p^t_{Ad}\frac{1}{\omega-\alpha}\left[q\pi^{(0)}_n + (1-q)\,\pi^{(2)}_d - \pi^{(1)}_D\right]$$

[17] Notice, however, that only consistent rules with Minimal Switch can be interpreted as learning "forwards in the tree" in the sense that $F\left(Di,x,j,y\right)_{Dj} = 0$ for $\{i,j\} = \{a,d\}$. See also Remark 3 below.

[18] The notation $P_e(y)$ has the obvious meaning (e.g. $P_e(y) = P^{(2)}_d(y)$ when $e = d$).

$$
\begin{aligned}
& +p_{Da}^{t}p_{Ad}^{t}\left(1-q\right)\left[\pi_{d}^{(2)}-\pi_{D}^{(1)}\right]+p_{Dd}^{t}p_{Aa}^{t}qF\left(Dd,n\right)_{Ad} \\
& -p_{Ad}^{t}p_{Da}^{t}qF\left(Ad,n,D\right)_{Dd} \\
& +p_{Aa}^{t}p_{Ad}^{t}\frac{1}{\omega-\alpha}\left(1-q\right)\left(\pi_{d}^{(2)}-\pi_{a}^{(2)}\right) \\
= \; & \left(p_{Dd}^{t}+p_{Da}^{t}\right)p_{Ad}^{t}\frac{1}{\omega-\alpha}\left(\pi_{Ad}-\pi_{D}\right)+qw^{t}F\left(Dd,n\right)_{Ad} \\
& +p_{Aa}^{t}p_{Ad}^{t}\frac{1}{\omega-\alpha}\left(\pi_{Ad}-\pi_{Aa}\right) \\
= \; & p_{Ad}^{t}\frac{1}{\omega-\alpha}\left[\pi_{Ad}-\pi\left(p^{t}\right)\right]+qw^{t}F\left(Dd,n\right)_{Ad} \qquad (11)
\end{aligned}
$$

Analogous expressions hold for $p_{Da}^{t+1}-p_{Da}^{t}$ and $p_{Aa}^{t+1}-p_{Aa}^{t}$. Since $F\left(Dd,n\right)_{Ad}=F\left(Da,n\right)_{Aa}$, we obtain

$$
\begin{aligned}
& p_{D}^{(1)}\left(t+1\right)-p_{D}^{(1)}\left(t\right) \\
= \; & \frac{1}{\omega-\alpha}\left[p_{Ad}^{t}\left(\pi_{D}-\pi_{Ad}\right)+p_{Aa}^{t}\left(\pi_{D}-\pi_{Aa}\right)\right]\cdot p_{D}^{(1)}\left(t\right) \\
= \; & \frac{1}{\omega-\alpha}\left[\pi_{D}-\pi\left(p^{t}\right)\right]\cdot p_{D}^{(1)}\left(t\right) \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
p_{d}^{(2)}\left(t+1\right)-p_{d}^{(2)}\left(t\right) \; = \; & p_{Ad}^{t}p_{Aa}^{t}\frac{1}{\omega-\alpha}\left(\pi_{Ad}-\pi_{Aa}\right) \qquad (13) \\
& +\left(1-q\right)\left[p_{Da}^{t}p_{Ad}^{t}F_{\circ}\left(Da,d\right)-p_{Dd}^{t}p_{Aa}^{t}F_{\circ}\left(Dd,a\right)\right]
\end{aligned}
$$

where $p_{D}^{(1)}=p_{Dd}+p_{Da}$, $p_{A}^{(1)}=1-p_{D}^{(1)}$ and $p_{d}^{(2)}=p_{Dd}+p_{Ad}$ are the analogues of the frequencies of different actions introduced in Section 3 for Parallel Bandits.

Remark 2 *Since the function f_2 underlying a consistent rule in Theorem 8 is associated to an undominated rule for within-decision learning at decision 2, the particular form of f_2 has no influence on the change of the population state (see part (iii) of Theorem 1 for Multi-Armed Bandits and part (b) of Theorem 5 for Parallel Bandits). On the other hand, the particular form of f_1 influences $F\left(Di,n\right)_{Ai}$ and, through the value of $F\left(Di,x,j,y\right)_{Dj}$, it also influences $F_{\circ}\left(Di,j\right)=F\left(Di,j\right)_{Dj}+F\left(Di,j\right)_{Aj}$.*

Imperfect information through sampling play instead of strategy affects learning at decision 2 only. Specifically, from (12), the frequency in which D is played at decision 1 evolves according to the same equation as under perfect information (see (3)). However, the perfect information analogue for the frequency evolution of Ad must be amended in (11) by the term $qw^{t}F\left(Dd,n\right)_{Ad}$ to account for switching from D to A at decision 1 when own planned action at decision 2 does not coincide with hidden action at decision

2 *when sampled play reaches n. Of course, we should expect no distortions from perfect information if the proportion of states in which d is planned does not depend on play in decision* 1, *i.e., if* $p_{Dd}/p_D^{(1)} = p_{Ad}/p_A^{(1)}$. *We see this intuition reflected in (11) since, from (10),* $p_{Dd}/p_D^{(1)} = p_{Ad}/p_A^{(1)}$ *if and only if* $w = 0$.

4.2 Selecting among Consistent Rules - The CPIRs

From the population dynamic developed in the previous section for consistent rules, we can calculate the change in average payoffs between rounds and through this apply the partial order introduced in Section 2. The following result is the analogue of Theorem 5 for Centipede Bandits. The only difference is that we are now only ranking rules belonging to the subclass of consistent rules.

Theorem 9 *A behavioral rule* F *is undominated among the consistent rules if and only if* $f_1(x,y)$ *in (8) is given by* $\frac{1}{\omega-\alpha}\max\{y-x,0\}$.

Proof. Let F be any consistent rule with $f_1(x,y) = \frac{1}{\omega-\alpha}\max\{y-x,0\}$ and G be any alternative consistent rule. In the following we will show that F is best at increasing average payoffs among the consistent rules in the sense that

$$\pi\left(p^{t+1}(F)\right) - \pi\left(p^t\right) < 0 \text{ implies } \pi\left(p^{t+1}(F)\right) - \pi\left(p^t\right) \geq \pi\left(p^{t+1}(G)\right) - \pi\left(p^t\right)$$

for any state p^t. Thus, $\pi\left(p^{t+1}(G)\right) \geq \pi\left(p^t\right)$ implies $\pi\left(p^{t+1}(F)\right) \geq \pi\left(p^t\right)$ which means that $I(G) \subseteq I(F)$.

The change in average payoff between rounds, $\pi\left(p^{t+1}(F)\right) - \pi\left(p^t\right)$, can be decomposed into a sum of separate terms involving $p_\eta p_\nu$ where η and ν range over all six possible strategy pairs. Three of these terms are

$$\frac{p_{Dd}^t p_{Ad}^t}{\omega-\alpha}\left(\pi_{Ad} - \pi_D\right)^2 \tag{14}$$

$$\frac{p_{Ad}^t p_{Aa}^t}{\omega-\alpha}\left(\pi_{Ad} - \pi_{Aa}\right)^2 \tag{15}$$

$$p_{Dd}^t p_{Aa}^t\left[\frac{1}{\omega-\alpha}\left(\pi_{Aa} - \pi_D\right)^2 + q(1-q)F_\circ(Dd,n)\left(\pi_d^{(2)} - \pi_a^{(2)}\right)\right] \tag{16}$$

We have omitted the other three analogous terms. From these three expressions, notice that only the second term in the right hand side of (16) depends on the exact specification of the consistent rule and this occurs only when Dd meets Aa. There is a similar expression when Da meets Ad.

Consider a state p^t in which $\pi\left(p^{t+1}\left(F\right)\right) < \pi\left(p^t\right)$. From (14) - (16),

$$\begin{aligned} 0 &> p^t_{Dd}p^t_{Aa}q\left(1-q\right)F_\circ\left(Dd,n\right)\left(\pi^{(2)}_d-\pi^{(2)}_a\right)+p^t_{Da}p^t_{Ad}q\left(1-q\right)F_\circ\left(Da,n\right)\left(\pi^{(2)}_a-\pi^{(2)}_d\right) \\ &= q\left(1-q\right)w^tF_\circ\left(Dd,n\right)\left(\pi^{(2)}_d-\pi^{(2)}_a\right) \end{aligned} \tag{17}$$

since $F_\circ\left(Dd,n\right)=F_\circ\left(Da,n\right)$. From (14) - (16) and the fact that $G_\circ\left(Dd,n\right)=G_\circ\left(Da,n\right)$ remains true for any consistent rule,

$$\begin{aligned} &\pi\left(p^{t+1}\left(F\right)\right)-\pi\left(p^{t+1}\left(G\right)\right) \\ =\ & p^t_{Dd}p^t_{Aa}q\left(1-q\right)\left[F_\circ\left(Dd,n\right)-G_\circ\left(Dd,n\right)\right]\left(\pi^{(2)}_d-\pi^{(2)}_a\right) \\ & +p^t_{Da}p^t_{Ad}q\left(1-q\right)\left[F_\circ\left(Da,n\right)-G_\circ\left(Da,n\right)\right]\left(\pi^{(2)}_a-\pi^{(2)}_d\right) \\ =\ & q\left(1-q\right)w^t\left[F_\circ\left(Dd,n\right)-G_\circ\left(Dd,n\right)\right]\left(\pi^{(2)}_d-\pi^{(2)}_a\right) \end{aligned} \tag{18}$$

Since PIR minimizes the probability of switching among the undominated rules for Multi-Armed Bandits (Theorem 2 (i)),

$$F_\circ\left(Dd,n\right)=\sum\frac{1}{\omega-\alpha}\max\left\{y-x,0\right\}P^{(1)}_D\left(x\right)P^{(0)}_n\left(y\right)\leq G_\circ\left(Dd,n\right)\ . \tag{19}$$

Since $w^t\left(\pi^{(2)}_d-\pi^{(2)}_a\right)<0$ from (17), expressions (18) and (19) combine to complete the proof that $\pi\left(p^{t+1}\left(F\right)\right)\geq\pi\left(p^{t+1}\left(G\right)\right)$ for all such p^t and hence, $I\left(G\right)\subseteq I\left(F\right)$.

In fact, we have shown above that the existence of p^t and payoff distributions for which $F_\circ\left(Dd,n\right)<G_\circ\left(Dd,n\right)$ and $\pi\left(p^{t+1}\left(F\right)\right)<\pi\left(p^t\right)$ implies $\pi\left(p^{t+1}\left(F\right)\right)>\pi\left(p^{t+1}\left(G\right)\right)$ which, together with a continuity argument, implies $I\left(G\right)\subsetneq I\left(F\right)$.

Assume that $I\left(G\right)=I\left(F\right)$. To complete the proof that $G\preceq F$ we will show $\pi\left(p^{t+1}\left(F\right)\right)\equiv\pi\left(p^{t+1}\left(G\right)\right)$. Following (14) - (16) it will suffice to prove $F_\circ\left(Dd,n\right)\equiv G_\circ\left(Dd,n\right)$. If this is not the case, then, by Theorem 2 (i), $F_\circ\left(Dd,x,n,y\right)<G_\circ\left(Dd,x,n,y\right)$ holds for some x,y. Choose payoff distributions of a Centipede Bandit such that $P^{(1)}_D\left(x\right)\cdot P^{(0)}_n\left(y\right)>0$ and $\pi_D=\pi_{Aa}>\pi_{Ad}$. Then $F_\circ\left(Dd,n\right)<G_\circ\left(Dd,n\right)$ and $\pi\left(p^{t+1}\left(F\right)\right)<\pi\left(p^t\right)$ when $C\left(p^t\right)=\left\{Dd,Aa\right\}$. Hence, $I\left(G\right)\subsetneq I\left(F\right)$ leads to a contradiction and the proof is complete. ■

Thus, only those consistent rules that use PIR for f_1 increase payoffs most often between rounds. Furthermore, by Remark 2, the choice of f_2 that satisfies (8) is immaterial in the population dynamic and so we can also take f_2 as PIR at decision 2 without loss of generality. For these reasons, we will call F a *Centipede Proportional Imitation Rule* (short, CPIR) if F is consistent and the functions f_1, f_2 in the statement in Theorem 8

are both given through proportional imitation; namely, $f_1 = f_2 = \frac{1}{\omega-\alpha}\max\{y-x,0\}$. From the discussion following Theorem 8, we will speak of *CPIR With Minimal Switch* ($F^{\min}$) if $F(Di,x,j,y)_{Dj} = 0\ \forall x,y \in [\alpha,\omega]$ when $\{i,j\} = \{d,a\}$. Analogously, *CPIR With Maximal Switch* ($F^{\max}$) satisfies $F(Di,x,j,y)_{Di} = 0$ and *CPIR With Proportional Switch* (F^{pro}) satisfies $F_{\circ}(Di,x,j,y) = (y-\alpha)/(\omega-\alpha)$.

Remark 3 *The following is easily verified.* $F^{\min}$ *is the unique CPIR where the plan at an unreachable decision equals the action last played at this decision (of course, provided this decision was faced in an earlier round).* $F^{\max}$ *is the unique CPIR that treats decisions* 1 *and* 2 *independently.*[19] $F^{\max}$ *is also the unique CPIR that must remember a plan for an unreachable decision the least number of rounds since memory is replaced by observed play whenever possible.*

4.3 Selecting among CPIRs

From Theorems 8 and 9, the only remaining flexibility in our program to select good rules of thumb is the choice of $F(Di,x,j,y)_{Dj}$ belonging to the interval $[0, 1-\max\{0,y-x\}/(\omega-\alpha)]$ in our CPIR. Clearly, this choice only changes planned play (i.e. memory) at unreachable nodes and so has no influence on change in average payoffs from one round to the next (in particular, all CPIRs are equivalent according to the partial order of Section 2). However, it will influence average payoffs two rounds later. This leads to the introduction of a third order preference in our partial order. As mentioned above, all consistent rules will sometimes lead to a decrease in average payoffs. When two rules decrease average payoffs between rounds under the same circumstances and at the same magnitude, we consider effects on average payoffs two rounds later. Specifically, a rule is preferred that is best at increasing average payoffs one round after such an unwanted decrease occurred. Formally, with $I(G)^c$ denoting the complement of $I(G)$, we extend our partial order $\preceq$ among behavioral rules for Centipede Bandits to be[20]

$$G \ \precsim \ F :\Leftrightarrow [I(G) \subsetneq I(F)]$$

[19] $\tilde{F}$ *always treats decision one and two independently* if, for all states p^t,

$$\tilde{F}(Da,d)_{Ad} = \left(\tilde{F}(Da,d)_{Ad} + \tilde{F}(Da,d)_{Aa}\right)\cdot\left(\tilde{F}(Da,d)_{Dd} + \tilde{F}(Da,d)_{Ad}\right)$$

[20] $\pi\left(p^{t+2}(G)\right) \leq \pi\left(p^{t+2}(F)\right)\big|_{I(G)^c}$ states that $\pi\left(p^{t+2}(G)\right) \leq \pi\left(p^{t+2}(F)\right)$ whenever $\pi\left(p^{t+1}(G)\right) < \pi\left(p^t\right)$.

$$\vee \left\{ [I(G) = I(F)] \wedge \left[\pi\left(p^{t+1}(G)\right) \leq \pi\left(p^{t+1}(F)\right) \right] \wedge \left[\pi\left(p^{t+1}(G)\right) \not\equiv \pi\left(p^{t+1}(F)\right) \right] \right\}$$
$$\vee \left\{ [I(G) = I(F)] \wedge \left[\pi\left(p^{t+1}(G)\right) \equiv \pi\left(p^{t+1}(F)\right) \right] \wedge \left[\pi\left(p^{t+2}(G)\right) \leq \pi\left(p^{t+2}(F)\right)\big|_{I(G)^c} \right] \right\}.$$

We will write *undominated'* for the domination concept with respect to the partial order $\precsim$. It turns out that there is enough cutting power through focusing on situations where payoffs are not increasing to select among the CPIRs. In particular, $F^{\min}$ is eliminated in this manner since it is strictly dominated by F^{pro} (see Corollary 3 below).

In our search for the best consistent rules (and hence best CPIRs rules) we will focus on CPIRs that satisfy an additional condition. Notice that, for certain CPIRs, when all strategies are equally good then the population state may not be a rest point of the dynamic. This can clearly occur in (11) when the proportion of planned action d at decision 2 differs among those using D and those using A at decision 1, i.e., $w \neq 0$. Less expected, it can also occur in (9) and (13) when $w = 0$ if $F_\circ(Da, d) \neq F_\circ(Dd, a)$. To avoid this last possibility, we will restrict our attention to *uncorrelated*[21] consistent rules which are defined as those that satisfy

$$p^{t+1} = p^t \text{ whenever } p^t_{Dd}/p^{(1)}_D(t) = p^t_{Ad}/p^{(1)}_A(t) \text{ and } \pi_{Ad} = \pi_{Aa} = \pi_D \ . \tag{20}$$

Uncorrelated CPIRs are characterized as follows.

Theorem 10 *A CPIR F is uncorrelated if and only if there exist functions $\tau_0(\cdot)$ and $\tau_1(\cdot)$ such that $\tau_1(x) \leq \frac{1}{\omega-\alpha}$ and*

$$F_\circ(Dd, x, a, y) = F_\circ(Da, x, d, y) = \tau_0(x) + \tau_1(x) \cdot y.$$

In particular, F^{pro} is uncorrelated with $\tau_0 \equiv \frac{-\alpha}{\omega-\alpha}$ and $\tau_1 \equiv \frac{1}{\omega-\alpha}$, and $F^{\max}$ is uncorrelated with $\tau_0 \equiv 1$ and $\tau_1 \equiv 0$. On the other hand, $F^{\min}$ is not uncorrelated which provides another reason to eliminate it as a good rule of thumb.

Through an analysis of the third order preference relationship in our partial order $\precsim$ we are able to select among the uncorrelated CPIRs. The following result shows that only convex combinations of F^{pro} and $F^{\max}$ are undominated' among the uncorrelated consistent rules.

[21] "$F_\circ(Da, d) = F_\circ(Dd, a)$ whenever $\pi^{(2)}_d = \pi^{(2)}_a$" (which is equivalent to (20)) may also be postulated directly as a plausible property of a good rule, since d and a are a priori indistinguishable with individual preferences over lotteries defined through expected payoffs.

Theorem 11 *Both F^{pro} and $F^{\max}$ are undominated among the uncorrelated consistent rules. If $\tilde{F}$ is undominated among the uncorrelated consistent rules, then there exists a function $\lambda : [\alpha, \omega] \to [0,1]$ such that*[22]

$$\tilde{F}_\circ(Dd, x, a, y) = \tilde{F}_\circ(Da, x, d, y) \tag{21}$$
$$= \lambda(x) \cdot F^{pro}_\circ(Dd, x, a, y) + (1 - \lambda(x)) \cdot F^{\max}_\circ(Dd, x, a, y) \ .$$

Checking expression (36) in the proof of Theorem 11 in all the relevant cases we obtain:

Corollary 3 $F^{\min} \precsim F^{pro}$

Following Remark 3, the bad performance of $F^{\min}$ shows that memory[23] implicit in Dd and Da must be managed in a more sophisticated manner than simply remembering own last play at decision 2.

Remark 4 *From expression (36) in the proof of Theorem 11, we see that there is a trade-off between increasing τ_1 and increasing $F_\circ(Da, d)$ in Theorem 10 when considering the third order effects of memory in states where average payoffs decrease. $F^{\max}$ maximizes $F_\circ(Da, d)$ at the cost of having $\tau_1(F^{\max}) \equiv 0$. F^{pro} maximizes τ_1 for all x at the cost of $F_\circ(Da, d) < 1$ (unless $\pi_d^{(2)} = \omega$). Common to any uncorrelated CPIR $\tilde{F}$ satisfying (21) is a high rate of imitation of successful play of others at decision 2 – regardless of own payoff (since $\tilde{F}_\circ(Da, x, d, y) \geq F^{pro}_\circ(Da, x, d, y) = (y - \alpha)/(\omega - \alpha)$).*

If asked to specify a unique best rule for Centipede Bandits, our choice falls on $F^{\max}$ primarily due to its simple implementation characterized by the two unique properties mentioned in Remark 3.[24] *As for our analysis of dynamic stability and long-run properties below, results will be generated for any specific convex combination of F^{pro} and $F^{\max}$.*

[22] Notice that $\tau_1(x) = \lambda(x) \geq 0$.

[23] Notice that, for any consistent rule, action i can be interpreted as memory of an action last experienced or sampled at decision 2.

[24] Another reason for selecting $F^{\max}$ is that it has unique properties concerning the eigenvalues at $p_{Ad} = 1$ when Ad is the optimal strategy. It can be shown that the eigenvalue in the direction of Da is more negative for the continuous version of (11) under $F^{\max}$ than for any other CPIR (and all other eigenvalues are the same) which suggests the basin of attraction of the best strategy is largest under $F^{\max}$.

4.4 Long-Run Population Dynamics, Asymptotic Stability and Long-Run Outcomes

We have already seen that global learning of best strategies is difficult in sequential decision tasks (cf., Theorem 7 (i)) compared to the long-run results for Parallel Bandits (cf., Corollary 2). The focus of this section lies in characterizing the long-run dynamic properties induced by the good rules selected in Theorem 11 (i.e., by convex combinations of F^{pro} and $F^{\max}$). The results show our good rules perform remarkably well. In fact, we derive results for more general CPIRs whenever possible. For instance, our first result (Theorem 12 below) shows that long-run outcomes are, in fact, identical for all consistent rules when D is either the best (i.e. $\pi_D > \max\{\pi_{Ad}, \pi_{Aa}\}$) or the worst (i.e. $\pi_D < \min\{\pi_{Ad}, \pi_{Aa}\}$) choice.

Theorem 12 *Suppose everyone in the population is using the same consistent rule. Consider a Centipede Bandit in which D is either the best or the worst choice. Then the set of best strategies is globally asymptotically stable.*

Proof. When $\pi_D > \max\{\pi_{Ad}, \pi_{Aa}\}$ then (12) implies $p_D^{(1)}(t+1) > p_D^{(1)}(t)$ for $0 < p_D^{(1)}(t) < 1$ and $\lim_{t\to\infty} p_D^{(1)}(t) = 1$ whenever $p_D^{(1)}(0) > 0$. Consequently, $\Delta(\{Dd, Da\})$ is asymptotically stable. Assume $\pi_D < \min\{\pi_{Ad}, \pi_{Aa}\}$. Then $p_D^{(1)}(t+1) < p_D^{(1)}(t)$ for $0 < p_D^{(1)}(t) < 1$ and $\lim_{t\to\infty} p_D^{(1)}(t) = 0$. If $\pi_{Ad} = \pi_{Aa}$ then $\Delta(\{Ad, Aa\})$ is asymptotically stable. Assume now that $\pi_{Ad} > \pi_{Aa}$. Since $\Delta(\{Ad, Aa\})$ is invariant and $\lim_{t\to\infty} p_{Aa}^t = 1$ on this subspace, interior trajectories get arbitrarily close to $\Delta(\{Ad, Aa\})$ and then $\lim_{t\to\infty} p_{Ad}^t = 1$. Furthermore, following (11), for every $\varepsilon > 0$ there exists $0 < \varepsilon' < \varepsilon$ such that i) $p_{Ad}^{t+1} > p_{Ad}^t$ whenever $p_{Ad}^t > 0$, $p_D^{(1)}(t) < \varepsilon'$ and $p_{Aa}^t > \varepsilon$, and, ii) $p_{Aa}^{t+1} < \varepsilon$ whenever $p_{Aa}^t < \varepsilon'$. Together with (12) this implies that $p_{Ad}^t > 1 - \varepsilon$ for all t if $p_{Ad}^0 > 1 - \varepsilon'$. Hence, p^* with $p_{Ad}^* = 1$ is asymptotically stable. ■

There is a simple heuristic explanation why good rules learn the best strategy in the long run in the two extreme cases of Theorem 12. Since imperfect information does not affect learning at decision 1 (see (12) and Remark 2) and since choice at decision 1 can be made independently of decision 2, populations using consistent rules learn what is best at decision 1. If D is the best choice then the argument is completed. If D is the worst choice then Dd and Da die out and consistency condition (d) in Definition 5 guarantees that the population finds the best strategy thereafter.

Thus, the difficulty with learning the best strategy for a CPIR can only occur when the choice at decision 1 cannot be made independently of decision 2, i.e., $\min\{\pi_{Ad}, \pi_{Aa}\} \leq$

$\pi_D \leq \max\{\pi_{Ad}, \pi_{Aa}\}$ and $\pi_{Ad} \neq \pi_{Aa}$. Moreover, q cannot be too close to either 0 or 1.[25] Theorem 13 and Corollary 4 below characterize the long-run outcomes in these cases for uncorrelated CPIRs with $\tau_1(\cdot) \geq 0$ (such as convex combinations of F^{pro} and $F^{\max}$). Here, one can only be sure that an interior trajectory does not end up in everyone choosing the worst strategy. The following lemma, needed in this analysis, is also important in its own right. It shows that in populations in which actions are initially chosen independently at each decision steadily learn which action is best at decision 2.[26]

Lemma 1 *Let F be an uncorrelated CPIR with $\tau_1(x) \geq 0$ for all $x \in [\alpha, \omega]$. Then $p_i^{(2)}(t+1) > p_i^{(2)}(t)$ for all t whenever p^0 is interior with $p_{Dd}^0/p_D^{(1)}(0) = p_{Ad}^0/p_A^{(1)}(0)$, $q < 1$ and $\pi_i^{(2)} > \pi_j^{(2)}$ for $\{i,j\} = \{d,a\}$.*

Theorem 13 *Suppose everyone in the population is using the same uncorrelated CPIR with $\tau_1(\cdot) \geq 0$. Assume that $\pi_{Aa} < \pi_D < \pi_{Ad}$ and that the initial state p^0 is interior. Then play at unreached decision 2 coincides in the long run with actual play at this decision, i.e.,*

$$\lim_{t\to\infty} p_d^{(2)}(t) = \lim_{t\to\infty} \frac{p_{Dd}^t}{p_D^{(1)}(t)} = \lim_{t\to\infty} \frac{p_{Ad}^t}{p_A^{(1)}(t)} ,$$

and either

i) $\lim_{t\to\infty} p_D^t = 1$ *and* $\pi_D \geq \lim_{t\to\infty} \left[p_d^{(2)}(t) \cdot \pi_{Ad} + p_a^{(2)}(t) \cdot \pi_{Aa}\right]$, *or*

ii) $\lim_{t\to\infty} p_{Ad}^t = 1$ $(= \lim_{t\to\infty} p_d^{(2)}(t))$.

Corollary 4 *With the same conditions as in Theorem 13, let $\gamma^* = (\pi_D - \pi_{Aa})/(\pi_{Ad} - \pi_{Aa})$. Then $\pi_D = \gamma^*\pi_{Ad} + (1-\gamma^*)\pi_{Aa}$ and $p_{Ad}^0 > \gamma^*$ implies $\lim_{t\to\infty} p_{Ad}^t = 1$. $\Delta(\{Ad\})$ is the unique minimal asymptotically stable set.*[27]

Remark 5 *We would like to present some intuition behind the above results.*

(Interior trajectories converge to outcome associated to Nash equilibrium of the agent normal form) *The agent normal form game G of the game against nature induced by a Centipede Bandit of Figure 4 is a 2×2 game with payoff matrix*

	d	a
D	π_D, π_D	π_D, π_D
A	π_{Ad}, π_{Ad}	π_{Aa}, π_{Aa}

.

[25] Following (14) - (16), average payoffs increase strictly in interior states when q is sufficiently close to 0 or 1 for given distinct $\pi_D^{(1)}, \pi_n^{(0)}, \pi_d^{(2)}$ and $\pi_a^{(2)}$.

[26] In fact, it can be shown that a CPIR being uncorrelated with $\tau_1(\cdot) \geq 0$ is actually a characterization of a CPIR satisfying Lemma 1 for all Centipede Bandits.

[27] Minimality refers to set-wise inclusion.

This game has unique equilibrium outcome whenever D is either the best or the worst choice. When $\pi_{Aa} < \pi_D < \pi_{Ad}$ then (D,a) and (A,d) are the only pure NE, their NE components are $\{(D, \gamma d + (1-\gamma)a \mid 0 \leq \gamma \leq \gamma^*\}$ and $\{(A,d)\}$ respectively. Given distinct expected payoffs[28]*, Theorems 12 and 13 combine with Corollary 4 to show that an interior trajectory $(p^t)_{t=0,1,2,..}$ induced by an uncorrelated CPIR with $\tau_1(\cdot) \geq 0$ converges to a limit point associated to a Nash equilibrium (q^*, r^*) of G where $(q^*, r^*) \in \Delta(\{D,A\}) \times \Delta(\{d,a\})$ is such that $q_D^* = \lim_{t\to\infty} p_D^{(1)}(t)$ and $r_d^* = \lim_{t\to\infty} p_d^{(2)}(t)$.*[29]

(Where does global learning fail?) *Trajectories for our dynamic are qualitatively similar to those of the replicator dynamic based on two populations and payoffs given by G (see Gale et al., 1995). This similarity gives some intuition why CPIRs do not necessarily lead to the best strategy when $\pi_{Aa} < \pi_D < \pi_{Ad}$ even when actions are initially chosen independently at each decision. $p_d^{(2)}$ is monotonically increasing (Lemma 1) but its rate of increase depends on how often decision 2 is reached. If $p_d^{(2)}$ is small, then D is currently the best choice at decision 1. The consequent increased play of D diminishes the number of times individuals can learn that Ad is better than Aa and eventually $p_D^{(1)}$ may converge to 1. Thus, play is locked in at the second best outcome. However, if $p_d^{(2)}(0) > \gamma^*$, then A is the best choice at decision 1 for all $t > 0$ and so we have both $p_A^{(1)}$ and $p_d^{(2)}$ converging monotonically to 1.*

(Long-run efficiency under rare mutations) *When $\pi_{Aa} < \pi_D < \pi_{Ad}$ the population can get locked in and individuals need not learn the best strategy in the long run. Nevertheless, we argue that individuals will eventually learn the best strategy if, as is common to the Evolutionary Game Theory literature, minimal asymptotically stable sets are taken to be the relevant solutions for making long-run predictions. A population state in any other set will eventually leave and never return after a rare mutation that causes a "small" discontinuity in the trajectory. When $\pi_{Aa} < \pi_D < \pi_{Ad}$, mutations can arise that make $p_d^{(2)}$ grow continuously until the trajectory enters the basin of attraction of the best strategy Ad out of which sufficiently rare mutations cannot cause the trajectory to leave.*

In order to make our understanding of optimal boundedly rational learning in Centipede Bandits more complete we are left to show that some trajectories actually converge to the inefficient outcome D when $\pi_{Aa} < \pi_D < \pi_{Ad}$. This was done through simulations of the dynamic process to get a feeling for the basins of attraction of D and Ad. For

[28] That is, $\pi_D, \pi_{Ad}, \pi_{Aa}$ are all different.

[29] Notice that this relationship to the agent normal form game also holds trivially for the Parallel Bandit setting. There, the Nash equilibria of the agent normal form game are precisely the best strategies.

parameters $\alpha = 0$, $\omega = 1$, $q = 0.5$ and payoff distributions $P_D^{(1)}(1) = 1 - P_D^{(1)}(0) = 0.6$, $P_n^{(0)}(0.8) = P_d^{(2)}(0.41) = P_a^{(2)}(0.1) = 1$, expected payoffs satisfy $\pi_{Aa} = 0.45 < \pi_D = 0.6 < \pi_{Ad} = 0.605$. From the simulations, we find that the maximal value of p_{Ad}^0 such that there are trajectories starting in the interior of $\left\{p : p_{Dd}/p_D^{(1)} = p_{Ad}/p_A^{(1)}\right\}$ that converge to $\Delta(\{Dd, Da\})$ is equal to 0.10 (to two decimals) for $F^{\max}$ whereas it is equal to 0.12 for F^{pro}. Notice that $\gamma^* = 0.9677$ in Corollary 4 (which guarantees convergence to the efficient outcome if $p_{Ad}^0 > \gamma^*$) for this Centipede Bandit which shows how coarse the statement made in Corollary 4 is when actions are initially chosen independently at each decision. When we take $P_d^{(2)}(0.5) = 1$ instead and hence $\pi_{Ad} = 0.65$ then the associated values are 0.02 and 0.03 respectively (whereas $\gamma^* = 0.75$).

Although strictly improving rules do not exist, the result below states that any CPIR with $\tau_1(\cdot) \geq 0$ is eventually strictly improving and decision-wise monotone (in the sense that it satisfies (12) and (22)).

Theorem 14 *Suppose everyone is using the same uncorrelated CPIR which satisfies $\tau_1(\cdot) \geq 0$ and suppose that π_D, π_{Aa} and π_{Ad} are distinct. Then, for any interior initial state, there exists a finite time t_0 such that $\pi(p^{t+1}) > \pi(p^t)$ and*

$$\frac{p_{Ad}^{t+1}}{p_A^{(1)}(t+1)} > \frac{p_{Ad}^t}{p_A^{(1)}(t)} \text{ when } \pi_{Ad} > \pi_{Aa} \tag{22}$$

for all $t \geq t_0$.

Remark 6 *At first glance, there seems to be a conflict between the existence of inefficient limit points of interior trajectories, the statement of Theorem 14 and known results on convergent trajectories. Specifically, the proof of Proposition 4.9 in Weibull (1995) shows that any limit point of an interior trajectory must be a Nash equilibrium if the trajectory is eventually best reply monotone. In our setting, this means that any such limit point must put all its weight on the best strategies. Thus, we obtain that best reply monotonicity cannot hold in the long run on all trajectories for our behavioral rules. Similarly, following Weibull (Proposition 4.11, 1995), any limit point of an interior trajectory must be a Nash equilibrium if the underlying dynamic is weakly payoff positive.* Weakly payoff positive *means that the dynamic is regular selection and the growth rate of at least one strategy present with an expected payoff strictly above average payoffs (if such a strategy exists) is strictly positive.*[30] *In particular, the second requirement holds after a finite time.*[31]

[30] That is, $p_i^{t+1}/p_i^t > 0$ for some $i \in B(p^t) = \{j : \pi_j > \pi(p^t)\}$ whenever $B(p^t) \neq \emptyset$.

[31] The only case not immediately obvious is when $\pi_{Aa} < \pi_D < \pi_{Ad}$ and $\lim p_D^t = 1$. However, here $\pi\left(p^{t+1}\right) > \pi(p^t)$ for all $t > t_0$ implies $\pi_D > \pi(p^t)$ and thus, following (12), $p_D^{t+1} > p_D^t$ for all $t > t_0$.

Hence, to assume a regular selection dynamic is a stronger condition than might appear at first sight, and, according to our approach, regularity is not a relevant concept when play instead of strategy is sampled.

4.5 Memory

Throughout Section 4, we have interpreted the specification of action i in strategy Di as a planned action at a decision node that cannot be reached. This also entered our conditions of consistency (see Definition 5). However, formally speaking, Dd and Da are two different states of memory. In the following we want to investigate the role of this memory by considering briefly the strategy space $\{D, Ad, Aa\}$.[32] Initially, this restricted strategy space would seem to simplify our analysis since a large part of Section 4 is devoted to the effects of $F(Di, x, j, y)_{Dj}$ and these are no longer relevant. However, behavioral rules based on $\{D, Ad, Aa\}$ have undesirable consequences that lead us to believe that individuals would develop good rules of thumb based on the larger strategy space $\{Dd, Da, Ad, Aa\}$.

Consider the strategy space $\{D, Ad, Aa\}$. As before, we search for a rule that is consistent with our good rules of thumb from tasks involving single or unrelated decisions. Hence the conditions in Definition 5 (a,b) must apply. This fixes the probability of switching from D to A at decision 1 when observing nature moving down (see (31)). Hence we need to determine the action at decision 2 in case of such a switch. Perhaps the most intuitive is to flip a coin between Ad and Aa; thus, $F(D, n)_{Ad} = F(D, n)_{Aa}$. However, under this specification, there exist parameter values[33] such that $\Delta(\{D\})$ is asymptotically stable although $\pi_{Ad} > \pi_D > \pi_{Aa}$. Thus, the best strategy cannot be reached with repeated rare mutations as shown in Remark 5 for $F^{\max}$ which is based on strategy space $\{Dd, Da, Ad, Aa\}$.

5 Conclusion

We investigate boundedly rational optimal individual behavior together with the resulting aggregate behavior when individuals update their strategies through sampling the

[32] In fact, this semi-reduced strategy space is often used when Game Theory is applied to the extensive form game of Figure 4.

[33] A sufficient condition is that $2 - \frac{2(\pi_{Ad} - \pi_D)}{(\omega - \alpha)q} > \frac{\pi_{Ad} - \pi_{Aa}}{\pi_D - \pi_{Aa}}$.

play rather than the strategy of others. Much of our analysis deals with the special consequences of imperfect information in these situations.

On the individual level, imperfect information calls for more memory than in Single-Decision Tasks. For this, we use the memory implicit in the specification of standard extensive-form game strategies. Imitative behavior selected for Single-Decision Tasks (Schlag, 1998a, 1998b) becomes a pasting procedure that proceeds backwards in the tree for Multi-Decision Tasks. We develop a lexico-graphic preference relationship based on how well expected average payoffs improve under different behavioral rules. We find the most preferred multi-decision behavior exhibits (i) a natural ignorance (through the use of POR) of own performance when sampling play in an independent decision, (ii) a high degree of imitating sampled play at decisions not reached in own last play (through the use of a convex combination of F^{pro} and $F^{\max}$), and (iii) the usefulness of the minimal switching property of PIR for minimizing errors that result when using memory to complete imperfect observations (through the use of CPIR).

On the aggregate level, we find that the improving condition can be recovered after a finite number of rounds. Nevertheless, long-run behavior contrasts with results under regular selection dynamics with complete information. When decisions are not independent it can happen that all individuals play an inefficient strategy in the long run even if each strategy is initially present. In particular, in the Centipede Bandit that we analyzed, long-run behavior can be predicted by considering all Nash equilibria (whether efficient or not) of a game where the individual plays against himself through the agent normal form.

Good rules for Centipede Bandits were derived from good rules for Parallel Bandits. Further analysis of good rules in more general Multi-Decision Tasks is not necessary in our approach. Consistent rules in general Multi-Decision Tasks remain play-wise imitating where this imitative behavior is determined through optimal behavior in Parallel and Centipede Bandits embedded in the Multi-Decision Task. We refrain here from explicitly stating the general definition of consistency. Instead we demonstrate, using CPIR With Maximal Switch, how rules need to be applied to general Multi-Decision Tasks in order to remain consistent. The key is to find the "deviation" node at which sampled play differs from own play for the first time. If the deviation node is associated to a decision of the individual, then imitate the sampled action (by pasting it into own last strategy) at this node with probability given by PIR and imitate sampled actions at all other nodes with probability one. If the deviation node is associated to a move of nature then

imitate simultaneously all sampled actions (again through pasting) with probability given by POR. This sets the stage for predicting long-run outcomes in general Multi-Decision Tasks. One may hope that solutions remain associated to Nash equilibria of the agent normal form (Remark 5).

We set off to analyze optimal boundedly rational behavior in extensive-form games and produced a detailed paper on optimal behavior in Multi-Decision Tasks instead. Due to its length, we have decided not to include our results on learning in games where each individual behaves non-strategically in the sense that he behaves as if he is facing a Multi-Decision Task. We intend to address these issues in a future paper.

Regarding related literature on optimal boundedly rational learning, Börgers et al. (1997) consider individual learning rules that rely on more memory than available in our setting but do not rely on observations of others' behavior. In their setting, there is no rule that is both improving (in terms of expected adjustment) and always leads to the best action in the long run.

6 References

D. Berry, and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London, 1985.

N.P. Bhatia, and, G.P. Szegö, *Stability Theory of Dynamical Systems,* Springer-Verlag, Berlin, 1970.

J. Björnerstedt and J. Weibull, Nash Equilibrium and Evolution by Imitation, In: *The Rational Foundations of Economic Behaviour*, *Proc. IEA Conference*, K. Arrow et al. (Eds.), MacMillan, London, 1996, 155-171.

T. Börgers, A. Morales, and, R. Sarin, "Simple Behaviour Rules Which Lead to Expected Payoff Maximizing Choices," mimeo, University College London, 1997.

B. Cooper, "Copying Fidelity in the Evolution of Finite Automata that Play the Repeated Prisoners' Dilemma," mimeo, Wolfson College, Oxford, 1997.

R. Cressman, Dynamic Stability in Symmetric Extensive Form Games, *Int. J. Game Theory* **26** (1997), 525-547.

R. Cressman, "Subgame Monotonicity in Two-Person Extensive Form Evolutionary Games," mimeo, Wilfrid Laurier University, 1996.

R. Cressman, and K. H. Schlag, "The Dynamic (In)Stability of Backwards Induction," University of Bonn, Disc. Paper **B-347**, Bonn, 1996.

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

D. Friedman, Evolutionary Games in Economics, *Econometrica* **59** (1991), 637-66.

D. Fudenberg, and D. Levine, Self-Confirming Equilibrium, *Econometrica* (1993), 523-545.

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

J. Harrington, "Rigidity and Flexibility in Social Systems," mimeo, The Johns Hopkins University, 1997.

D. Helbing, Interrelations between Stochastic Equations for Systems with Pair Interactions, *Physica A* **181** (1992), 29-52.

J. Hofbauer, and K. Sigmund, *The Theory of Evolution and Dynamical Systems,* Cambridge University Press, 1988.

J. H. Nachbar, "Evolutionary" Selection Dynamics in Games: Convergence and Limit Properties, *Int. J. Game Theory* **19** (1990), 59-89.

G. Nöldeke and L. Samuelson, An Evolutionary Analysis of Backwards and Forwards Induction, *Games Econ. Beh.* **5** (1993), 425-454.

M. A. Nowak, and, R. M. May, Evolutionary Games and Spatial Chaos, *Nature* **359** (1992), 826-829.

K. Ritzberger, and J. W. Weibull, Evolutionary Selection in Normal-Form Games, *Econometrica* **63** (1995), 1371-1400.

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

C. Ruebeck, "Imitation Dynamics in the Repeated Prisoners' Dilemma: an Exploratory Example," mimeo, The Johns Hopkins University, 1997.

L. Samuelson, *Evolutionary Games and Equilibrium Selection*, MIT Press, Cambridge, Mass., 1997.

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

K. H. Schlag, Which One Should I imitate?, *J. Math. Econ.* (1998a, forthcoming).

K. H. Schlag, Why Imitate and if so, How? A Bounded Rational Approach to Multi-Armed Bandits, *J. Econ. Theory* (1998b, forthcoming).

J. M. Swinkels, Adjustment Dynamics and Rational Play in Games, *Games Econ. Beh.* **5** (1993), 455-484.

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

J. W. Weibull, *Evolutionary Game Theory*, The MIT Press, Cambridge, Mass., 1995.

A Proofs and Notes

Remark 7 *The discrete-time dynamic* $(p^t)_{t=0,1,2,..}$ *is a* regular selection dynamic *if the growth rate* p_i^{t+1}/p_i^t *as a function of* p^t *can be extended to a Lipschitz continuous function on* $\Delta(S)$ *(Samuelson and Zhang, 1992). In particular, extinct strategies remain extinct since* $p_i^0 = 0$ *implies* $p_i^t = 0$ *for all* t*. A regular selection dynamic is* monotone *(Nachbar, 1990; Friedman, 1991*[34]*; Samuelson and Zhang, 1992) if* $p_i^{t+1}/p_i^t > (=)\, p_j^{t+1}/p_j^t$ *if and only if* $\pi_i > (=)\, \pi_j$ *holds whenever* $p_i^t \cdot p_j^t > (=)\, 0$*. A regular selection dynamic is* sign preserving *(Nachbar, 1990) if* $p_i^{t+1}/p_i^t > (=)\, 0$ *if and only if* $\pi_i > (=)\, \pi(p^t)$*. Both monotonicity and sign preserving imply that* $\pi(p^{t+1}) \geq \pi(p^t)$ *with strict inequality whenever* $\pi_i \neq \pi_j$ *and* $p_i^t \cdot p_j^t > 0$ *for some* i, j*. The proof of this statement for monotone dynamics is as follows. Given* $p^t \in \Delta(S)$*, monotonicity implies that there exists* $j \in S$ *such that* $p_i^{t+1} \geq p_i^t$ *if and only if* $\pi_i \geq \pi_j$*. Hence,*

$$\pi\left(p^{t+1}\right) - \pi\left(p^t\right) = \sum_i \left(p_i^{t+1} - p_i^t\right) \pi_i \geq \sum_i \left(p_i^{t+1} - p_i^t\right) \pi_j = 0.$$

Proof of Theorem 3. Assume that F is undominated. Then F is improving. If F is not imitating, then there exists some simple Parallel Bandit with $a \in S_2$ and $x, y \in \mathbf{R}$ such that $F(Aa, 1, x, A^1, y)_{Aa} + F(Aa, 1, x, a^2, y)_{Aa} < 2$. Consider payoff distributions such that $P_A^{(1)}(x) \cdot P_A^{(1)}(y) \cdot P_a^{(2)}(y) > 0$ and π_{Aa} is the unique best strategy. Then $F(Aa, 1, A^1)_{Aa} + F(Aa, 1, a^2)_{Aa} < 2$ and $p_{Aa}^t = 1$ implies $p_{Aa}^{t+1} < 1$ which contradicts the fact that F is improving. Thus, F is imitating.

Since the change in average payoffs between consecutive rounds can be decomposed into a sum of terms as in (23) below for any imitating rule F, we may restrict attention to the case where $S_2 = \{a, z\}$ when proving parts (ii) and (iii) of the theorem. When $C(p^t) = \{Aa, Az\}$, the change in average payoffs is

$$\begin{aligned}\frac{\pi\left(p^{t+1}\right) - \pi\left(p^t\right)}{p_{Aa}^t p_{Az}^t} &= q_1 q_2 \left[F_\circ\left(Aa, 1, z^2\right) - F_\circ\left(Az, 1, a^2\right)\right](\pi_{Az} - \pi_{Aa}) \\ &+ (q_2)^2 \left[F_\circ\left(Aa, 2, z^2\right) - F_\circ\left(Az, 2, a^2\right)\right](\pi_{Az} - \pi_{Aa}). \end{aligned} \tag{23}$$

[34]Friedman (1991) uses the term *order compatible pre-dynamic.*

Since F is improving, (23) must be nonnegative for any Parallel Bandit in which $\pi_{Aa} \neq \pi_{Az}$. By considering all allowable values of q_1, q_2, we find that

$$\left[F_\circ\left(Aa,1,z^2\right) - F_\circ\left(Az,1,a^2\right)\right]\left[\pi_z^{(2)} - \pi_a^{(2)}\right] \geq 0 \tag{24}$$

and

$$\left[F_\circ\left(Aa,2,z^2\right) - F_\circ\left(Az,2,a^2\right)\right]\left[\pi_z^{(2)} - \pi_a^{(2)}\right] \geq 0. \tag{25}$$

We will now show that (24) implies that there exist functions $c_{12} : \left[\alpha^{(1)}, \omega^{(1)}\right] \to \mathbb{R}$ and $d_{12} : \left[\alpha^{(1)}, \omega^{(1)}\right] \to \left[0, \frac{1}{\omega^{(2)} - \alpha^{(2)}}\right]$ such that

$$F_\circ\left(Aa,1,x,z^2,y\right) = F_\circ\left(Az,1,x,a^2,y\right) = c_{12}(x) + d_{12}(x) \cdot y\ . \tag{26}$$

Consider the following Parallel Bandit: given $\alpha^{(2)} \leq v < y < u \leq \omega^{(2)}$ and arbitrary $x \in \left[\alpha^{(1)}, \omega^{(1)}\right]$ and $\lambda \in [0,1]$, let $P_A^1(x) = 1 = P_a^2(y)$[35], $P_z^2(v) = \lambda$ and $P_z^2(u) = 1 - \lambda$. Let $\theta(y) = F_\circ(Az,1,x,a^2,y)$, $\beta(v) = F_\circ(Aa,1,x,z^2,v)$ and $\gamma(u) = F_\circ(Aa,1,x,z^2,u)$. Then (24) implies

$$\lambda\beta(v) + (1-\lambda)\gamma(u) \geq (\leq)\,\theta(y) \text{ if } \lambda < (>)\,\frac{u-y}{u-v}\ .$$

Especially it follows that

$$(u-v)\,\theta(y) = (u-y)\,\beta(v) + (y-v)\,\gamma(u)\,. \tag{27}$$

We will now use the fact that, for fixed x, θ only depends on y. Fixing v and u and varying y it follows that there exist $c_{12}(x)$ and $d_{12}(x)$ such that

$$\theta = c_{12} + y \cdot d_{12} \text{ for all } y \in (v,u)\,.$$

In fact, the above must hold for all $y \in \left(\alpha^{(2)}, \omega^{(2)}\right)$. Similarly, $\beta(v)$ is linear in v with constants c_β and d_β and $\gamma(u)$ is linear in u with constants c_γ and d_γ. Entering these expressions into (27) and comparing terms it follows that

$$c_{12}(x) = c_\beta(x) = c_\gamma(x) \text{ and } d_{12}(x) = d_\beta(x) = d_\gamma(x)\,.$$

From (24), it follows that $d_{12}(x) \geq 0$. Since $c_{12}(x) + d_{12}(x) \cdot \omega^{(2)} \leq 1$ and $c_{12}(x) + d_{12}(x) \cdot \alpha^{(2)} \geq 0$ we then obtain $d_{12}(x) \leq \frac{1}{\omega^{(2)} - \alpha^{(2)}}$ and this completes the proof of (26).

[35] That is, these two lotteries are δ-distributions and so realized payoff equals expected payoff.

Notice that (25) is precisely the improving condition for Single-Decision Tasks (i.e. for within-decision learning) at decision 2. Following (Schlag, 1998b), there exists $0 \leq \sigma_2 \leq \frac{1}{\omega^{(2)}-\alpha^{(2)}}$ such that

$$F_\circ\left(Aa,2,x,z^2,y\right) - F_\circ\left(Az,2,y,a^2,x\right) = \sigma_2 \cdot (y-x) \tag{28}$$

Thus

$$\begin{aligned}\frac{\pi\left(p^{t+1}\right)-\pi\left(p^t\right)}{p^t_{Aa}p^t_{Az}} &= \left(q_2\right)^2\left[q_1\sum d_{12}\left(x\right)P_A\left(x\right)+q_2\sigma_2\right]\left(\pi_z^{(2)}-\pi_a^{(2)}\right)^2 \\ &\leq \frac{1}{\omega^{(2)}-\alpha^{(2)}}\left(q_2\right)^2\left(\pi_z^{(2)}-\pi_a^{(2)}\right)^2\end{aligned}$$

with equality if and only if $d_{12}\left(x\right)=\sigma_2=\frac{1}{\omega^{(2)}-\alpha^{(2)}}$ and $c_{12}\left(x\right)=-\frac{\alpha^{(2)}}{\omega^{(2)}-\alpha^{(2)}}$. Our second order preference relationship now implies that undominated rules require equality which in turn implies parts (ii) and (iii) of the theorem. In particular, any undominated rule must have, for cross-decision learning, $F_\circ\left(Aa,1,x,z^2,y\right)=F_\circ\left(Az,1,x,a^2,y\right)=\frac{y-\alpha}{\omega^{(2)}-\alpha^{(2)}}$ which is the POR rule.

The proof that (i), (ii) and (iii) imply F is undominated is now straightforward. In fact, the above proof has characterized all improving rules in these simple Parallel Bandits as those that satisfy equations (26) and (28). ■

Proof of Theorem 5: Assume that F satisfies (b1) - (b3). Then F is consistent by Definition 3 and improving by Corollary 2. Assume that $F \prec G$. Then G is improving and by Theorem 4, G is play-wise imitating. Hence, behavior of G is uniquely determined through within- and cross-decision learning. However, F is consistent and performs best in these situations. This contradicts $F \prec G$ and hence F is undominated.

The fact that (a) implies (b1) - (b3) follows from the above argument together with the fact that consistent and play-wise imitating rules satisfy (b1) - (b3). ■

Proof of Theorem 8: Each of the four properties in Definition 5 put certain restrictions (equations (30) to (33) below respectively) on the switching probabilities for the behavioral rule F that are given as follows. The notation $i,j \in \{d,a\}$ is used throughout.

First, Definition 5 (a) restricts $F(\eta,x,e,y)_\nu$ where $e \in \{D,d,a\}$. By Theorem 1 (ii), there exist functions $f_{\{D,d\}}, f_{\{D,a\}}, f_{\{d,a\}} : [\alpha,\omega]^2 \to [0,1]$ that each satisfy

$$f_{\{D,i\}}\left(x,y\right) - f_{\{D,i\}}\left(y,x\right) = f_{\{d,a\}}\left(x,y\right) - f_{\{d,a\}}\left(y,x\right) = \frac{y-x}{\omega-\alpha} \text{ for } i \in \{d,a\} \tag{29}$$

such that

$$F\left(Di,x,D,y\right)_{Dd} + F\left(Di,x,D,y\right)_{Da} = F\left(Aj,x,j,y\right)_{Aj} = 1$$

$$\begin{aligned} F(Aj,x,D,y)_{Dd} + F(Aj,x,D,y)_{Da} &= F(Di,x,j,y)_{Aj} = f_{\{D,j\}}(x,y) \\ F(Aj,x,D,y)_{Aj} = F(Di,x,j,y)_{Dd} + F(Di,x,j,y)_{Da} &= 1 - f_{\{D,j\}}(x,y) \\ F(Ad,x,a,y)_{Aa} = F(Aa,x,d,y)_{Ad} &= f_{\{d,a\}}(x,y) \\ F(Ad,x,a,y)_{Ad} = F(Aa,x,d,y)_{Aa} &= 1 - f_{\{d,a\}}(x,y). \end{aligned} \tag{30}$$

Definition 5 (b) implies that the probability of switching actions from D to A at decision 1 is independent of the sampled action (i.e. the probability only depends on the payoffs x and y) if the sampled endnode is $e \in \{n,d,a\}$. When combined with (a) this means that

$$\begin{aligned} & F(Di,x,n,y)_{Ad} + F(Di,x,n,y)_{Aa} \\ = \; & F(Aj,n,x,D,y)_{Dd} + F(Aj,n,x,D,y)_{Da} = f_{\{D,d\}}(x,y) = f_{\{D,a\}}(x,y) \end{aligned} \tag{31}$$

Definition 5 (c) refers to behavior in states where $p_{Dj} + p_{Aj} = 0$ and implies, when combined with (b), that

$$\begin{aligned} F(Di,x,D,y)_{Di} &= F(Ai,x,n,y)_{Ai} = F(Ai,n,x,i,y)_{Ai} = F(Ai,n,x,n,y)_{Ai} = 1 \\ F(Di,x,n,y)_{Ai} &= F(Ai,x,D,y)_{Di} = F(Ai,n,x,D,y)_{Di} = f_{\{D,d\}}(x,y) \\ F(Di,x,n,y)_{Di} &= F(Di,x,i,y)_{Di} = F(Ai,n,x,D,y)_{Ai} = 1 - f_{\{D,d\}}(x,y) \end{aligned} \tag{32}$$

Finally, Definition 5 (d) refers to behavior in states where $p_{Dd} + p_{Da} = 0$ and implies, when combined with (b), that

$$\begin{aligned} F(Ad,n,x,a,y)_{Aa} &= F(Aa,n,x,d,y)_{Ad} = \frac{y-\alpha}{\omega-\alpha} \\ F(Ad,n,x,a,y)_{Ad} &= F(Aa,n,x,d,y)_{Aa} = \frac{\omega-y}{\omega-\alpha} \end{aligned} \tag{33}$$

In particular, the Proportional Observation Rule must be used to switch between Ai and Aj if own last strategy Ai lead to n and observed play is Aj.

Conversely, it is straightforward to show that any F satisfying these restrictions is consistent. The statement of Theorem 8 summarizes the above analysis. ■

Proof of Theorem 10. Following (9) and (11), a consistent rule F is uncorrelated if and only if

$$F_\circ(Da,d) = F_\circ(Dd,a) \text{ whenever } \pi_d^{(2)} = \pi_a^{(2)}. \tag{34}$$

Next we show that there exist functions $\tau_0(\cdot)$ and $\tau_1(\cdot)$ such that

$$F_\circ(Da,x,d,y) = F_\circ(Dd,x,a,y) = \tau_0(x) + \tau_1(x) \cdot y \tag{35}$$

with the help of the proof of Theorem 3. Consider a Simple Parallel Bandit with $S_1 = \left\{\tilde{A}\right\}$ and $S_2 = \{\tilde{a}, \tilde{z}\}$. Identify D with $\tilde{A}$, a with $\tilde{a}$ and d with $\tilde{z}$. Then (20) corresponds to (24) and (35) is identical to (26) if τ_0 is replaced by c_{12} and τ_1 is replaced by d_{12}. Following the arguments in the proof of Theorem 3, (34) implies (27) which is then used to prove (35).

It follows from $1 \geq F_\circ(Dd, x, a, \omega) - F_\circ(Dd, x, a, \alpha) = \tau_1(x) \cdot (\omega - \alpha)$ that $\tau_1(x) \leq \frac{1}{\omega - \alpha}$. Clearly, $F^{\max}$ and F^{pro} are both uncorrelated with $\tau_0 \equiv 1$ and $\tau_1 \equiv 0$ for $F^{\max}$ and $\tau_0 \equiv \frac{-\alpha}{\omega - \alpha}$ and $\tau_1 \equiv \frac{1}{\omega - \alpha}$ for F^{pro}. On the other hand, $F_\circ^{\min}(Dd, x, a, y) = \frac{1}{\omega - \alpha} \max\{y - x, 0\}$ does not have a representation as in (35). ■

Proof of Theorem 11. Suppose F is a CPIR. Let us compare the effects of $F(Dd, a)_{Dd}$ and $F(Dd, a)_{Da}$ on the increase in average payoffs, $\pi(p^{t+2}(F)) - \pi(p^t(F))$, between rounds t and $t+2$. Remember that $F(Dd, a)_{Dd} + F(Dd, a)_{Da} = 1 - F(Dd, a)_{Aa}$ is constant for any CPIR. Thus, the only difference among CPIRs in the change in average payoffs occurs from round $t+1$ to $t+2$ when there are differences in the number of switches from Dd to Ad and from Da to Aa resulting from sampling play of n in round $t+1$. However, since both these switches occur with the same probability for each individual in the appropriate circumstances, the net effect on average payoffs is

$$\left[F(Dd, a)_{Dd} \pi_{Ad} + F(Dd, a)_{Da} \pi_{Aa}\right] p_{Dd}^t p_{Aa}^t (1 - q) \cdot p_A^{(1)}(t+1) \cdot q F_\circ(Dd, n).$$

Combining this with the analogous effect of $F(Da, d)_{Dd}$ and $F(Da, d)_{Da}$ yields (up to the common positive factor $(1 - q) \cdot p_A^{(1)}(t+1) \cdot q F_\circ(Dd, n)$)

$$p_{Dd}^t p_{Aa}^t \left[F(Dd, a)_{Dd} \pi_{Ad} + F(Dd, a)_{Da} \pi_{Aa}\right] + p_{Da}^t p_{Ad}^t \left[F(Da, d)_{Dd} \pi_{Ad} + F(Da, d)_{Da} \pi_{Aa}\right]$$

which simplifies to

$$k + (\pi_{Ad} - \pi_{Aa}) \left[p_{Dd}^t p_{Aa}^t (F_\circ(Da, d) - F_\circ(Dd, a)) - F_\circ(Da, d) w^t\right] \tag{36}$$

where k does not depend on the particular form of the CPIR.

Now suppose F is an uncorrelated CPIR with functions $\tau_0(\cdot)$ and $\tau_1(\cdot)$. Assume that $\pi_{Ad} \geq \pi_{Aa}$. Let $\lambda(x) := \max\{\tau_1(x) \cdot (\omega - \alpha), 0\}$ and let $\tilde{F}$ be the uncorrelated CPIR given by (21). In particular, $\tau_1^{\tilde{F}}(x) = \max\{\tau_1(x), 0\}$. Since $\tau_0(x) + \tau_1(x) \cdot \omega \leq 1$ for all x, we obtain

$$F_\circ(Da, x, d, y) \leq \min\{1 - \tau_1(x) \cdot (\omega - y), 1\} = \tilde{F}_\circ(Da, x, d, y).$$

Consider now circumstances in which $\pi\left(p^{t+1}\left(F\right)\right) < \pi\left(p^{t}\left(F\right)\right)$. Following the proof of Theorem 9, $w^t < 0$. By continuity, we can also assume that $p^t_{Dd}p^t_{Aa} > 0$. From (36), we see that

$$\begin{aligned} & p^t_{Dd}p^t_{Aa}\left(F_\circ\left(Da,d\right) - F_\circ\left(Dd,a\right)\right) - F_\circ\left(Da,d\right)w^t \\ = \; & p^t_{Dd}p^t_{Aa}\left[\sum \tau_1\left(x\right)P_D\left(x\right)\right]\left(\pi^{(2)}_d - \pi^{(2)}_a\right) - F_\circ\left(Da,d\right)w^t \\ \leq \; & p^t_{Dd}p^t_{Aa}\left[\sum \tau_1^{\tilde{F}}\left(x\right)P_D\left(x\right)\right]\left(\pi^{(2)}_d - \pi^{(2)}_a\right) - \tilde{F}_\circ\left(Da,d\right)w^t \end{aligned} \tag{37}$$

where strict inequality holds in (37) whenever F does not have a representation as in (21). Hence, only convex combinations of F^{pro} and $F^{\max}$ can be undominated among the uncorrelated consistent rules.

In the following we will show that both F^{pro} and $F^{\max}$ are in fact undominated among the uncorrelated consistent rules. Since we only have to compare each one to convex combinations of the two, given linearity of (36), all we must show is that F^{pro} is not preferred to $F^{\max}$ (according to $\succsim$) nor vice versa.

For a given CPIR F, consider first a Centipede Bandit, and a state p^t in which $p^t_{Dd}p^t_{Aa} > 0$, $-w^t > 0$, $\pi^{(2)}_d > \pi^{(2)}_a = \pi^{(0)}_n = \pi^{(1)}_D$ and $F_\circ\left(Dd,n\right) > 0$. Then

$$\pi\left(p^{t+1}\left(F\right)\right) - \pi\left(p^t\right) = \left(\pi_{Ad} - \pi_{Aa}\right)\left[p_{Ad}\left(1-p_{Ad}\right)\frac{\pi_{Ad}-\pi_{Aa}}{\omega-\alpha} + qw^tF_\circ\left(Dd,n\right)\right] < 0$$

if $q < 1$ is sufficiently large and p_{Ad} is sufficiently small. Hence, under such circumstances, the third order preference of $\succsim$ applies when comparing CPIRs. When comparing the change in payoffs for such states between rounds t and $t+2$ for F^{pro} and $F^{\max}$, the relevant expressions from (36) are

$$\begin{aligned} & p^t_{Dd}p^t_{Aa}\left(F^{pro}_\circ\left(Da,d\right) - F^{pro}_\circ\left(Dd,a\right)\right) - F^{pro}_\circ\left(Da,d\right)w^t \\ = \; & p^t_{Dd}p^t_{Aa}\frac{\pi^{(2)}_d - \pi^{(2)}_a}{\omega-\alpha} - w^t\frac{\pi^{(2)}_d - \alpha}{\omega-\alpha} \end{aligned}$$

and

$$p^t_{Dd}p^t_{Aa}\left(F^{\max}_\circ\left(Da,d\right) - F^{\max}_\circ\left(Dd,a\right)\right) - F^{\max}_\circ\left(Da,d\right)w^t = -w^t$$

respectively. When $-w^t$ is chosen sufficiently large relative to $p^t_{Dd}p^t_{Aa}$, we see that F^{pro} is not preferred to $F^{\max}$. Similarly, by choosing $-w^t$ small relative to $p^t_{Dd}p^t_{Aa}$, it follows that F^{pro} is not preferred to $F^{\max}$. ■

Proof of Lemma 1. From equations (9), (11) and (12), using the fact that $w = p_{Dd} - p_D^{(1)} \cdot p_d^{(2)}$,

$$\begin{aligned} w^{t+1} - w^t &= p_{Dd}^{t+1} - p_{Dd}^t - \left[p_D^{(1)}(t+1) - p_D^{(1)}(t)\right] p_d^{(2)}(t) \\ &\quad -p_D^{(1)}(t+1) \cdot \left[p_d^{(2)}(t+1) - p_d^{(2)}(t)\right] \\ &= p_D^{(1)}(t) \frac{1}{\omega - \alpha} \left[p_{Da}^t p_{Ad}^t (\pi_D - \pi_{Ad}) - p_{Dd}^t p_{Aa}^t (\pi_D - \pi_{Aa})\right] \\ &\quad + \left[p_D^{(1)}(t) + p_D^{(1)}(t+1)\right] p_{Ad}^t p_{Aa}^t \frac{1}{\omega - \alpha} (\pi_{Aa} - \pi_{Ad}) \\ &\quad -w^t q F(Dd, n)_{Ad} \\ &\quad +p_A^{(1)}(t+1) \cdot (1-q) \left[p_{Da}^t p_{Ad}^t F_\circ(Da, d) - p_{Dd}^t p_{Aa}^t F_\circ(Dd, a)\right] \end{aligned} \tag{38}$$

Let F be an uncorrelated CPIR. Then (38) simplifies to

$$\begin{aligned} & w^{t+1} - w^t \\ = \; & -w^t \left[\frac{1}{\omega - \alpha} p_D^{(1)}(t) \cdot (\pi_D - \pi_{Aa}) + qF(Dd, n)_{Ad} + p_A^{(1)}(t+1) \cdot (1-q) F_\circ(Dd, a)\right] \\ & +p_{Ad}^t (\pi_{Aa} - \pi_{Ad}) \left[\begin{array}{c} \frac{1}{\omega - \alpha} \left(p_D^{(1)}(t) \cdot p_a^{(2)}(t) + p_D^{(1)}(t+1) \cdot p_{Aa}^t\right) \\ -\tau_{1,D} \cdot p_A^{(1)}(t+1) \cdot p_{Da}^t \end{array}\right] \end{aligned} \tag{39}$$

where $\tau_{1,D} = \sum \tau_1(x) P_D(x)$.

Next we will show that

$$\left\{w^t \leq 0\right\} \text{ is invariant whenever } \pi_d^{(2)} \geq \pi_a^{(2)} \; . \tag{40}$$

If $\tau_{1,D} = 0$ the claim follows from (38), using the fact that $F_\circ(Da, d) = F_\circ(Dd, a)$. Now assume that $\tau_{1,D} = \frac{1}{\omega - \alpha}$. Since $w^t = p_D^{(1)}(t) \cdot p_a^{(2)}(t) - p_{Da}^t$, the last line in (39) is bounded above by $-w^t \frac{1}{\omega - \alpha} p_{Ad}^t (\pi_{Ad} - \pi_{Aa})$. Given $\tau_{\mathbf{0},D} = \sum \tau_{\mathbf{0}}(x) P_D(x)$ we obtain from $\tau_{\mathbf{0},D} + \tau_{1,D} \cdot \alpha \geq 0$ and $\tau_{\mathbf{0},D} + \tau_{1,D} \cdot \omega \leq 1$ that $\tau_{\mathbf{0},D} = -\frac{\alpha}{\omega - \alpha}$. Consequently, $F_\circ(Dd, a) = \frac{\pi_a - \alpha}{\omega - \alpha}$. Since F is a CPIR, $\frac{1}{\omega - \alpha} \left(\pi_D - \pi_n^{(0)}\right) + F(Dd, n)_{Ad} = F(Ad, n, D)_{Dd} \leq 1$. Hence, $w^t \leq 0$ implies

$$\begin{aligned} w^{t+1} - w^t &\leq -w^t q \left[p_D^{(1)}(t) \frac{1}{\omega - \alpha} \left(\pi_D^{(1)} - \pi_n^{(0)}\right) + F(Dd, n)_{Ad}\right] \\ &\quad -w^t \frac{1}{\omega - \alpha} (1-q) \left[\begin{array}{c} p_D^{(1)}(t) \cdot \left(\pi_D^{(1)} - \pi_a^{(2)}\right) \\ +p_A^{(1)}(t+1) \cdot \left(\pi_a^{(2)} - \alpha\right) + p_{Ad}^t \left(\pi_d^{(2)} - \pi_a^{(2)}\right) \end{array}\right] \\ &\leq -w^t \end{aligned}$$

Finally, assume that $\tau_{1,D} \in \left(0, \frac{1}{\omega-\alpha}\right)$. $F_\circ(Dd,a) - \sum F_\circ(Dd,x,a,\omega) P_D(x) = \tau_{1,D} \cdot (\pi_a - \omega)$ implies $F_\circ(Dd,a) \leq 1 - \tau_{1,D} \cdot (\omega - \pi_a)$. $F' \equiv [1 - \tau_{1,D} \cdot (\omega - \alpha)] \cdot F^{\max} + \tau_{1,D} \cdot (\omega - \alpha) \cdot F^{pro}$ defines an uncorrelated Centipede Proportional Imitation Rule with associated $\tau_1' = \tau_{1,D}$ and $F_\circ'(Dd,a) = 1 - \tau_{1,D} \cdot (\omega - \pi_a)$ (in this definition $\tau_{1,D}$ is a constant).

$$w^{t+1}(F') - w^t = [1 - \tau_{1,D} \cdot (\omega - \alpha)] \left[w^{t+1}(F^{\max}) - w^t\right] + \tau_{1,D} \cdot (\omega - \alpha) \left[w^{t+1}(\tilde{F}) - w^t\right]$$

implies $w^{t+1}(F') - w^t \leq -w^t$ when $w^t \leq 0$. Since $F_\circ(Dd,a) \leq 1 - \tau_{1,D} \cdot (\omega - \pi_a) = F'(Dd,a)$, (39) implies $w^{t+1}(F) - w^t \leq w^{t+1}(F') - w^t$ when $w^t \leq 0$. This completes the proof of (40).

The proof of the lemma now follows from (40) together with

$$\begin{aligned} & p_d^{(2)}(t+1) - p_d^{(2)}(t) = p_{Ad}^t p_{Aa}^t \frac{1}{\omega - \alpha}(1-q)\left(\pi_d^{(2)} - \pi_a^{(2)}\right) - (1-q)\, w^t F_\circ(Dd,a) \\ & + p_{Da}^t p_{Ad}^t (1-q)\left[F_\circ(Da,d) - F_\circ(Dd,a)\right] \\ = \;& (1-q)\left(\pi_d^{(2)} - \pi_a^{(2)}\right) p_{Ad}^t \left[\frac{1}{\omega - \alpha} p_{Aa}^t + \tau_{1,D} \cdot p_{Da}^t\right] - (1-q)\, w^t F_\circ(Dd,a) \;. \end{aligned} \tag{41}$$

■

Proof of Theorem 13. Recall that $\pi_{Ad} > \pi_D > \pi_{Aa}$. Assume that $w^t > 0$ for all t. Following (11), p_{Ad}^t is monotonically increasing with $\lim_{t\to\infty} p_{Ad}^t = 1$. The conclusion of the theorem is then obvious.

Now assume there exists $t_0 \geq 0$ such that $w^{t_0} \leq 0$. $p_d^{(2)}(t)$ is monotonically increasing once $t > t_0$ by Lemma 1. Following (41), $w^t \to 0$ and $p_{Ad}^t p_{Aa}^t \to 0$ as $t \to \infty$. Consequently, $p_{Ad}^t \cdot w^t = \left(p_{Ad}^t\right)^2 p_{Da}^t - p_{Ad}^t p_{Aa}^t p_{Dd}^t \to 0$ and hence $p_{Ad}^t p_{Da}^t \to 0$ and $p_{Ad}^t \cdot p_a^{(2)}(t) \to 0$ as $t \to \infty$.

Let $\gamma = \lim_{t\to\infty} p_d^{(2)}(t)$. Then $\gamma \geq p_d^{(2)}(t_0) > 0$. Assume that $\gamma < 1$. Then $p_a^{(2)}(t) \geq 1 - \gamma > 0$ for all t and hence $p_{Ad}^t \to 0$ and $p_{Dd}^t \to \gamma$ as $t \to \infty$. $\lim_{t\to\infty} w^t \to 0$ implies $\lim_{t\to\infty} p_{Aa}^t = 0$. Consequently, $\lim_{t\to\infty} p_D^{(1)}(t) = 1$ and $\lim_{t\to\infty}\left[p_{Dd}^t / p_D^{(1)}(t)\right] = \gamma$. Since $\lim_{t\to\infty} w^t = 0$ we obtain
$\lim_{t\to\infty}\left[p_{Ad}^t / p_A^{(1)}(t) - p_{Dd}^t / p_D^{(1)}(t)\right] = 0$. Hence, $\lim_{t\to\infty}\left[p_{Ad}^t / p_A^{(1)}(t)\right] = \gamma$. Assume that $\pi_D < \gamma\pi_{Ad} + (1-\gamma)\pi_{Aa}$. Then (12) implies that $p_D^{(1)}(t)$ is decreasing for sufficiently large t. This contradicts $\lim_{t\to\infty} p_D^{(1)}(t) = 1$ and hence $\pi_D \geq \gamma\pi_{Ad} + (1-\gamma)\pi_{Aa}$.

Finally, consider the case when $\gamma = 1$. Then trajectories get arbitrarily close to $\Delta(\{Dd, Ad\})$, an invariant subspace on which movement is in direction of Ad. Similar arguments as in the first paragraph of this proof show that $\lim_{t\to\infty} p_{Ad}^t = 1$. ■

Proof of Corollary 4. Given p^0 with $p^0_{Ad} > \gamma^*$, let $t_0 \in N \cup \{0, \infty\}$ be such that $w^t > 0$ when $t < t_0$ and $w^t \leq 0$ when $t \geq t_0$. Following the proof of Theorem 13, $p^{t_0}_{Ad} \geq p^0_{Ad} > \gamma^*$ and $p^{(2)}_d(t) \geq p^{(2)}_d(t_0) \geq p^{t_0}_{Ad} > \gamma^*$ for all $t > t_0$ implies $\lim_{t\to\infty} p^t_{Ad} = 1$.

Consider any $\varepsilon > 0$ such that $p_{Ad} > (1-\varepsilon)^2$ implies $\pi_D < \pi(p)$. Consider p^0 with $p^0_{Ad} > 1-\varepsilon$. We will show that $p^t_{Ad} > (1-\varepsilon)^2$ for all t whenever $p^0_{Ad} > 1-\varepsilon$ and this proves the asymptotic stability of $\Delta(\{Ad\})$. For all $t \leq t_0$, $p^t_{Ad} > (1-\varepsilon)^2$ holds by the monotonicity of p^t_{Ad}. Assume that t_0 is finite. Assume that $p^{(1)}_A(\tilde{t}) > 1-\varepsilon$ and $p^{(2)}_d(\tilde{t}) > 1-\varepsilon$ for some $\tilde{t} \geq t_0$. Notice that $\tilde{t} = t_0$ has this property. Then $p^{(2)}_d(\tilde{t}+1) \geq p^{(2)}_d(\tilde{t})$ and $-w^t = p^t_{Ad} - p^{(1)}_A(t) \cdot p^{(2)}_d(t) \geq 0$ for $t \geq t_0$ implies $p^{\tilde{t}}_{Ad} \geq p^{(1)}_A(\tilde{t}) \cdot p^{(2)}_d(\tilde{t}) \geq (1-\varepsilon)^2$ and hence, $p^{(1)}_A(\tilde{t}+1) \geq p^{(1)}_A(\tilde{t})$ by assumption on ε. Thus, $p^{\tilde{t}+1}_{Ad} \geq p^{(1)}_A(\tilde{t}+1) \cdot p^{(2)}_d(\tilde{t}+1) \geq p^{(1)}_A(\tilde{t}) \cdot p^{(2)}_d(\tilde{t}) \geq (1-\varepsilon)^2$.

The monotonic increase of $p^{(2)}_d(t)$ whenever $w^0 \leq 0$, together with the fact that $p^{(2)}_d(0) > \gamma^*$ and $w^0 \leq 0$ implies $\lim_{t\to\infty} p^t_{Ad} = 1$, shows that $\Delta(\{Ad\})$ is contained in every asymptotically stable set. Thus, minimality of $\Delta(\{Ad\})$ is immediate. ■

Proof of Theorem 14. Fix p^0 in the interior of $\Delta(\{Dd, Da, Ad, Aa\})$. Without loss of generality, assume that $\pi_{Ad} > \pi_{Aa}$. If $w^t > 0$ for all t then, from (17), $\pi(p^{t+1}) > \pi(p^t)$ holds for all t. Assume therefore that there exists $t_0 \geq 0$ such that $w^{t_0} \leq 0$. Following (14) - (16), $t \geq t_0$ implies

$$\begin{aligned} \pi(p^{t+1}) - \pi(p^t) \quad &\geq \quad p^t_{Dd} p^t_{Aa} \frac{1}{\omega - \alpha} \left[(\pi_{Aa} - \pi_D)^2 + (\pi_{Ad} - \pi_D)^2 \right] \\ &\quad + w^t q(1-q) F_\circ(Dd, n) \left(\pi^{(2)}_d - \pi^{(2)}_a \right) . \end{aligned} \tag{42}$$

Following the proofs of Theorems 12 and 13, $\lim_{t\to\infty} w^t = 0$. Hence,

$$\lim_{t\to\infty} \frac{p^t_{Bi}}{p^{(1)}_B(t) \cdot p^{(2)}_i(t)} = 1 \tag{43}$$

for $B \in \{D, A\}$ and $i \in \{d, a\}$ and

$$\lim_{t\to\infty} \frac{w^t}{p^t_{Dd} p^t_{Aa}} = 1 - \lim_{t\to\infty} \frac{p^t_{Da} p^t_{Ad}}{p^t_{Dd} p^t_{Aa}} = 0. \tag{44}$$

From (42) and (44), $\pi(p^{t+1}) > \pi(p^t)$ holds once t is sufficiently large.

Inequality (22) holds if and only if

$$\begin{aligned} 0 \quad &< \quad p^{t+1}_{Ad} \cdot p^{(1)}_A(t) - p^t_{Ad} \cdot p^{(1)}_A(t+1) \\ &= \quad p^{(1)}_A(t) \cdot \left[\begin{array}{c} p^t_{Ad} \frac{1}{\omega - \alpha} (\pi_{Ad} - \pi(p^t)) + q w^t F(Dd, n)_{Ad} \\ -p^t_{Ad} \frac{1}{\omega - \alpha} \left(\frac{p^t_{Ad} \pi_{Ad} + p^t_{Aa} \pi_{Aa}}{p^{(1)}_A(t)} - \pi(p^t) \right) \end{array} \right] \end{aligned}$$

$$= p_A^{(1)}(t) \cdot \left[\frac{p_{Ad}^t p_{Aa}^t}{p_A^{(1)}(t)} \frac{1}{\omega - \alpha} (\pi_{Ad} - \pi_{Aa}) + w^t q F(Dd, n)_{Ad} \right]$$

Hence, the claim is complete once we show

$$\lim_{t\to\infty} \frac{w^t p_A^{(1)}(t)}{p_{Ad}^t p_{Aa}^t} = \lim_{t\to\infty} \left[p_A^{(1)}(t) \cdot \left(\frac{p_{Dd}^t}{p_{Ad}^t} - \frac{p_{Da}^t}{p_{Aa}^t} \right) \right] = \lim_{t\to\infty} \left[\frac{p_A^{(1)}(t)}{p_{Ad}^t} p_{Dd}^t \left(1 - \frac{p_{Da}^t p_{Ad}^t}{p_{Aa}^t p_{Dd}^t} \right) \right] = 0 \tag{45}$$

If (ii) holds in Theorem 13 or $\pi_D < \pi_{Aa} < \pi_{Ad}$ then (45) follows from (43) since $\lim_{t\to\infty} p_{Ad}^t = 1$. Alternatively, if (i) holds in Theorem 13 or $\pi_{Aa} < \pi_{Ad} < \pi_D$, then the second term in (45) equals 0, since for $i \in \{d, a\}$,

$$\lim_{t\to\infty} \left[p_A^{(1)}(t) \cdot \frac{p_{Di}^t}{p_{Ai}^t} \right] = \lim_{t\to\infty} \left[\frac{p_A^{(1)}(t)}{p_{Ai}^t} \cdot \frac{p_{Di}^t}{p_D^{(1)}(t)} \cdot p_D^{(1)}(t) \right] = \lim_{t\to\infty} p_D^{(1)}(t) = 1 \ .$$

■