Journal of Economic Behavior & Organization xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

journal homepage: www.elsevier.com/locate/jebo

Learning by (limited) forward looking players☆

Friederike Mengel[a,b,*]

[a] Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom
[b] Department of Economics (AE1), Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 22 January 2014
Received in revised form 5 August 2014
Accepted 10 August 2014
Available online xxx

JEL classification:
C73
C90
D03

Keywords:
Game theory
Learning
Forward-looking agents
Prisoner's Dilemma experiments

ABSTRACT

We present a model of adaptive economic agents who are k periods forward looking. Agents in our model are randomly matched to interact in finitely repeated games. They form beliefs by learning from past behavior of others and then best respond to these beliefs looking k periods ahead. We establish almost sure convergence of our stochastic process and characterize absorbing sets. These can be very different from the predictions in both the fully rational model and the adaptive, but myopic case. In particular we find that also Non-Nash outcomes can be sustained whenever they satisfy a "local" efficiency condition. We then characterize stochastically stable states in a class of 2×2 games and show that under certain conditions the efficient action in Prisoner's Dilemma games and coordination games can be singled out as uniquely stochastically stable. We show that our results are consistent with typical patterns observed in experiments on finitely repeated Prisoner's Dilemma games and in particular can explain what is commonly called the "endgame effect" and the "restart effect". Finally, if populations are composed of some myopic and some forward looking agents, parameter constellations exist such that either might obtain higher average payoffs.

1. Introduction

When trying to understand how economic agents involved in strategic interactions form beliefs and make choices, traditional game theory has ascribed a large degree of rationality to players. Agents in repeated games are, for example, assumed to be able (and willing) to analyze all possible future contingencies of play, and find equilibria via a process of backward induction, or to at least act as if they were doing so. In recent decades this model has been criticized by experimental work demonstrating that agents often do not seem to engage in backward induction when making choices in finitely repeated games.[1] In a different line of research some efforts have been made to develop models of learning, in which agents are assumed to adapt their beliefs (and thus choices) to experience rather than reasoning strategically. In these models agents usually display a substantial degree of myopia, learning e.g. through reinforcement or imitation or choosing

☆ I wish to thank Sam Bowles, Jayant Ganguli, Paul Heidhues, Jaromir Kovarik, Christoph Kuzmics, John Miller, RanSpiegler, Elias Tsakas, two anonymous Reviewers as well as seminar participants in Alicante, Bielefeld, Bilbao, Bonn, Curacao (FCGTC 2012), Faro (SAET 2011), Malaga (ESEM 2012), Muenchen, Santa Fe and Stony Brook forhelpful comments. Financial support by the European Union (Grant PIEF-2009-235973) and the NWO (Veni Grant 451-11-020) is gratefully acknowledged.

* Correspondence to: Department of Economics, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, United Kingdom. Tel.: +44 1206873417.
E-mail address: fr.mengel@gmail.com

[1] See Gueth et al. (1982) or Binmore et al. (2001) among others.

http://dx.doi.org/10.1016/j.jebo.2014.08.001
0167-2681/

myopic best responses.[2] Typically, though, one would expect that economic agents rely on both: adaptation and some degree of forward looking.[3]

In this paper, we present a learning model aiming to bring these two features together. While we recognize that agents are adaptive, we also allow them to be forward looking to some degree. Agents in our model are randomly matched to interact in finitely repeated two-player games. Such interactions are characteristic of many real-life situations. Work relationships often are finitely repeated games, where after completing one project, people start working with someone else.[4] Friends often interact repeatedly, but few stay friends for a life-time. And companies will bargain a deal with one client and afterwards start bargaining with another client.

Agents in our model form beliefs by relying on past experience in the same situation (after the same recent history) and then best respond to these beliefs looking k periods ahead. A researcher, for example, wondering how a co-author might react to a certain choice of action is likely to base her beliefs on this and previous co-authors' reactions to the same or a similar history of play. Standard models of adaptive play (see e.g. Young, 1993) implicitly or explicitly make two assumptions that rule out such reasoning. They assume (i) that agents are myopic and (ii) that agents believe that the distribution of opponent's choices is independent of the history of play. Both assumptions go well together since, if adaptive agents believe that the opponent's behavior is independent of the history, then it does not matter whether they are forward looking or not. In our model we relax both assumptions. We allow agents to be forward looking and we allow them to condition their beliefs about the opponent's choices on the recent history of play. Our model nests the model of adaptive play by Young (1993).

The stochastic process implied by our learning model can be described by a finite Markov chain of which we characterize absorbing and stochastically stable states. We find that absorbing sets are such that either a Nash equilibrium of the one shot game satisfying very mild conditions or an outcome that is "locally efficient", but not necessarily Nash, will be induced almost all the time as the length of the interaction grows larger. Outcomes can thus be very different from the predictions in both the fully rational and the myopic cases. We also establish almost sure convergence to such absorbing sets. We then characterize stochastically stable states in a class of 2×2 games and show that under certain conditions the efficient action in Prisoner's Dilemma games and coordination games can be singled out as uniquely stochastically stable. Again this contrasts with the results obtained for adaptive, but myopic agents analyzed by Young (1993).

We show that our results are consistent with typical patterns observed in experiments on repeated Prisoner's Dilemma games, such as e.g. by Andreoni and Miller (1993). In particular our model can explain why people cooperate in finitely repeated Prisoner's Dilemma games. It can explain what experimental economists often refer to as "endgame effect", namely the fact that after many periods of cooperation participants start to defect in the last periods in experiments with finitely repeated prisoner dilemma games. It can also explain the so-called "restart effect", i.e. the fact that if – after the endgame effect has been observed – participants are rematched and the finitely repeated game is "restarted", participants start to cooperate again.[5]

Finally, we also show that if populations are composed of some myopic and some forward looking agents there are some parameter constellations under which myopic agents obtain higher average payoff and others where forward-looking agents obtain higher average payoffs in absorbing states. Hence it is not clear ex ante whether myopic or forward-looking agents will have higher evolutionary fitness and there may be conditions where both coexist.

Some other authors have studied models with (limited) forward-looking agents. Jehiel (1995) has proposed an equilibrium concept for agents making limited horizon forecasts in two-player infinite horizon games, in which players move alternately. Under his concept agents form forecasts about their own and their opponent's behavior and act to maximize the average payoff over the length of their forecast. In equilibrium forecasts have to be correct. Jehiel (2001) shows that this equilibrium concept can sometimes single out cooperation in the *infinitely* repeated Prisoner's Dilemma as a unique prediction if players' payoff assessments are non-deterministic according to a specific rule. Apart from being strategic another difference between his and our work is that his concept is only defined for infinite horizon alternate move games whereas our model deals with finitely repeated (simultaneous move) games. In Jehiel (1998) he proposes a learning justification for limited horizon equilibrium.

Blume (2004) has studied an evolutionary model of unlimited forward looking behavior. In his model agents are randomly matched to play a one shot game. They revise their strategies sporadically taking into account how their action choice will affect the dynamics of play of the population in the future. He shows that myopic play arises whenever the future is discounted heavily or whenever revision opportunities arise sufficiently rarely. He also shows that the risk-dominant action evolves in the unique equilibrium in Coordination games. Unlike our agents, his agents anticipate how their behavior affects other players' beliefs in the future. In a recent paper Heller (2014) studies a repeated prisoner's dilemma where agents can choose their foresight ability ex ante and shows that agents will look at most three periods ahead. In his model foresight refers to anticipating the end of the interaction correctly. Hence a player with less foresight can consider more future periods if

[2] See e.g. Young (1993), Kandori et al. (1993) or the textbook by Fudenberg and Levine (1998).

[3] There is also some empirical evidence supporting this view. See e.g. Ehrblatt et al. (2010).

[4] Researchers' co-authorship relations or the work relations of flight crew on commercial airlines might be described in this manner. In some large organizations there are, in fact, explicit policies for staff rotation (see e.g. Bac, 1996).

[5] See e.g. Andreoni (1988), Burlando and Hey (1997) or Selten and Stoecker (1986). Selten and Stoecker (1986) also provide a (different) explanation of the endgame effects they observe based on learning.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

the game is "unusually" short in his model. As a consequence his notion of foresight is quite different from our notion of forward-looking behavior, where forward looking agents are defined by considering more future periods. A second major difference is that foresight is an endogenous choice in his model.[6] Fudenberg and Kreps (1995) have studied learning of individuals who repeatedly play a fixed extensive-form game. As in our model their players learn from past experience with the population to forecast future actions and as in our model they may not learn full behavioral strategies. Two key differences are (i) that their agents are not forward looking, i.e. they maximize only their immediate expected payoff ($k = 1$) and (ii) their players always learn correct beliefs on the path of play. These two key differences lead to very different results. Their players will learn self-confirming equilibria (see Fudenberg and Levine, 1993). As a consequence outcomes can be quite different from our model. Cooperation in the finitely repeated prisoner's dilemma, which can be an outcome of our learning process, is e.g. not a self-confirming equilibrium.[7]

The paper is organized as follows. In Section 2 we present the model. In Section 3 we collect our main results. Section 4 discusses extensions and Section 5 concludes. The proofs are relegated to an Appendix.

2. The model

2.1. Basic definitions

There is a finite number of individuals partitioned into two non-empty classes $i = 1, 2$. Every T periods 2 players are randomly drawn from the population, one from each class, to interact repeatedly in a symmetric normal form two-player game. We will index the player drawn from class i with the same index i as the class and will be explicit whether we are referring to the player or the class whenever doing otherwise could give rise to confusion. Each interaction consists of T repetitions of the stage game. In the stage game, each player in class i has a finite set of actions A_i to choose from. The payoff that player i obtains in a given period if she chooses action a_i and her opponent action a_{-i} is given by $\pi_i(a_i, a_{-i})$. We denote by $\vec{a}^t = (a_1^t, a_2^t)$ an action profile showing the action choices of both players at time t.

2.2. Histories

A history of play H^t lists the last (at most) h action profiles realized in the *current* T-period interaction. Hence

$$H^t = \begin{cases} (\vec{a}^{t-h}, \ldots, \vec{a}^{t-1}) & \text{if} \quad \forall \tau = t-h, \ldots, t-1 : \tau \neq 0 \mod T \\ (\vec{a}^{\max\{\tau<t:\tau=0 \mod T\}+1}, \ldots, \vec{a}^{t-1}) & \text{if} \quad \exists \tau \in t-h, \ldots, t-2 : \tau = 0 \mod T \\ H^0 & \text{if} \quad t-1 = 0 \mod T \end{cases}, \tag{1}$$

where H^0 is defined as the 0-tuple or empty sequence. Denote by $\mathcal{H}(h) = (A_i \times A_{-i} right)^h$ the set of all possible histories of length h and by $\mathcal{H} = H^0 \cup \mathcal{H}(1) \cup \ldots \mathcal{H}(h)$ the set of all possible histories of length smaller or equal than h.

2.3. Learning, memory, beliefs

Agents in our model are adaptive. They form beliefs about their opponent's action choices based on past play of the population and they condition these beliefs on the history of play. They also have limited foresight of k periods, meaning that – given their beliefs – they choose actions in order to maximize their expected utility across the following (at most) k periods. We now explain how beliefs are formed and show how choices are made in Section 2.4.

Memory: Agents have limited memory. For each history $H \in \mathcal{H}$ all agents i remember only the last m instances where the history was H and memorize the action choice of players in class $(-i)$ immediately following such a history. Denote by $M_i^t(H)$ the m-tuple of action choices of players in class $(-i)$ in the last m interactions (as seen from t) in which the history was H. Let $M_i^t = (M_i^t(H))_{H \in \mathcal{H}}$ be the collection of $M_i^t(H)$ for all possible histories and denote by $M^t = (M_i^t)_{i=1,2}$ the collection of memories across the two classes of players. Note that m is *not* history-dependent. This implies that agents can remember reactions to "rare" events even if they lie far back in time whereas they might not remember more "common" or "frequent" events even if they are closer in time. For example a consultant may remember clearly her superior's reaction to an event ("history") 10 years back in time where she badly mishandled a project and was almost fired as a consequence. But she may not remember the reaction to an event 5 years back where everything went "normal". Note also that we assumed that all agents in the same class share the same memory, though this assumption can be relaxed.

[6] Other studies include Fujiwara-Greve and Krabbe-Nielsen (1999) who study coordination problems, Selten (1991) or Ule (2005) who models forward looking players in a network.

[7] There is also some conceptual relation to the literature on long-run and short-run players. See also Fudenberg and Levine (1989) or Watson (1993) among others.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

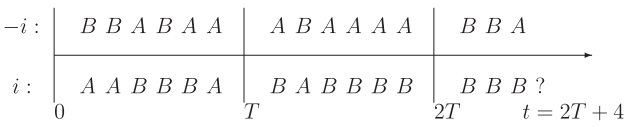

Fig. 1. Example time-line: At time $t = 2T+4$ player i wants to decide on an action plan. Assume that $h = 1$ and $m = 5$. The history at time t is $H^t = (B, A)$. The memory agent i has conditional on history (B, A) is denoted by $M_i^t(B, A) = (A, B, A, A, A)$. These are the last five action choices of agents in class $-i$ following the history (B, A). Assume now that both agents choose B. Then the new history is $H^{t+1} = (B, B)$ and the memory $M_i^{t+1}(B, A)$ is updated to (B, A, A, A, B).

Beliefs: After observing a given history H, agents then randomly sample (independently from others and without replacement) $\rho \leq m$ out of the last m periods where the history was H.[8] Given the realization of this random draw, the probability $\mu_i^t(a_{-i}|H)$ that agent i attaches to her opponent choosing action a_{-i} conditional on the current history being H then corresponds to the frequency with which a_{-i} was chosen after history H in the sample drawn. If a history occurred less than ρ times in the past, agents sample all periods in which the history occurred. If a history never occurred in the past, agents use a default belief $\mu_i^{t-DF}(a_{-i}^{t-1}) = 1$, i.e. they assume that the opponent keeps playing the same action as in the previous period.[9] Denote by $\mu_i^t(H)$ the (realized) belief of agent i given history H at time t. Fig. 1 illustrates an example of how memories are formed.[10]

2.4. Choices

Forward looking agents have beliefs not only about the opponent's choice in the current period, but also over the paths of play in the following k periods (conditional on their own choices). However, as we noted above, if there are less than k periods left to play, agents realize this and correspondingly form beliefs about the path of play only in the remaining periods. In the notation we reflect this by defining $t + k^* = t + k - 1$ if $\forall \tau = t+1, \ldots, t+k-1 : \tau \neq 0 \text{mod} T$ and $t + k^* = \min\{\tau > t : \tau + 1 = 0 \text{mod} T\}$ otherwise. For each action plan $(a_i^\tau)_{\tau=t,\ldots t+k^*}$ an agent entertains at t, conditional beliefs about the opponent's choice induce beliefs over "terminal nodes", where "terminal" is determined by the degree of forward looking k. Beliefs over terminal nodes are denoted by bold letters $\boldsymbol{\mu_i^t}((\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}|(\tilde{a}_i^\tau)_{\tau=t}^{t+k^*})$. The term $(a_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}$ reflects the fact that beliefs over terminal nodes are beliefs over paths of play of length k (or less than that if less periods are left to play) and the term $(a_i^\tau)_{\tau=t}^{t+k^*}$ reflects the fact that those beliefs are formed *conditional* on an agent's own action plans (see also Fig. 2). Beliefs over terminal nodes are constructed as follows:

$$\boldsymbol{\mu_i^t}((\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}|(\tilde{a}_i^\tau)_{\tau=t}^{t+k^*}) = \mu_i^t(a_{-i}^t|H^t) * \mu_i^t(a_{-i}^{t+1}|H^{t+1}_{(\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}}) \ldots * \mu_i^t(a_{-i}^{t+k^*}|H^{t+k^*}_{(\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}}),$$

where $H^{t+1}_{(\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}}$ is the history at time $t+1$ under the path of play $(\tilde{a}_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}$.

Fig. 2 illustrates how beliefs over terminal nodes are formed. At $t = 1$ we assume that agents choose an action randomly from A_i. In all subsequent periods $t > 1$ – given beliefs over terminal nodes – agents choose an action plan that maximizes their expected payoff over the next k periods.

$$\max_{(a_i^\tau)_{\tau=t,\ldots t+k^*}} V(\mu_i^t(H), (a_i^\tau)) = \sum_{(a_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}} \boldsymbol{\mu_i^t}((a_i^\tau, a_{-i}^\tau)_{\tau=t}^{t+k^*}|(a_i^\tau)_{\tau=t}^{t+k^*}) \sum_{\tau=t}^{t+k^*} \pi^i(a_i^\tau, a_{-i}^\tau). \tag{2}$$

Hence, when making a choice agents think about future paths of play and how their current choices might affect those. This idea seems inherent in the notion of forward looking behavior. Define by $\mathcal{BR}_i^t(\cdot)$ the instantaneous best response of player i for the repeated game, in the sense that for any plan of choices $(a_i^t; (a_i^\tau)_{\tau=t+1}^{t+k^*-1}) \in \text{argmax}\, V(\mu_i^t(H), (a_i^\tau))$ we have $a_i^t \in \mathcal{BR}_i^t(\cdot)$. We are interested in $\mathcal{BR}_i^t(\cdot)$, since only a_i^t is realized with certainty. The rest of the action plan is simply used to compute continuation payoffs. Since players revise their choice at each t, this can potentially lead to time inconsistencies. In other words, it is possible that an agent plans to choose some action at a future date $\tau > t$, but ends up choosing something else when that time arrives. Such time inconsistencies are characteristic of many real life decisions and seem inherent to the notion of limited foresight. Finally, note that for $(h, k) = (0, 1)$ this model nests the model of adaptive play by Young (1993).

[8] Note that if $h = 0$ then players just sample ρ out of the last m periods. We introduce imperfect sampling in order to nest the model of Young (1993) for the myopic case and to be able to establish almost sure convergence.

[9] This will imply that only Nash equilibria can be sustained by default beliefs, all other profiles have to be sustained via learned beliefs in an absorbing state.

[10] One may wonder why the 5th coordinate in $M_i^t(B, A)$ in Fig. 1 is not B, since after all at $2T$ the action profile was (B, A) followed by the opponent's choice of B at $2T+1$. The reason is that players were rematched at $2T$ and hence see the choice of B at $2T+1$ as a "reaction" to the empty sequence H^0 rather than to history (B, A).

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

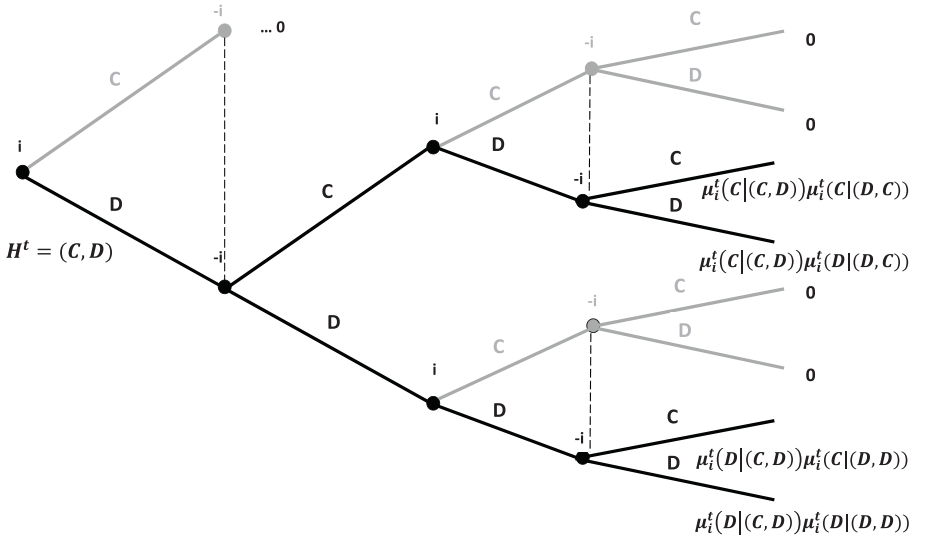

Fig. 2. Beliefs over "terminal nodes". The figure illustrates how beliefs over "terminal nodes" are formed, where "terminal" is determined by k. In the example agents play a 2×2 Prisoner's Dilemma game (with actions C and D – see also Sections 3.2 and 3.3), $k = 2$ and $h = 1$. At the beginning of the tree (at t) we have $H^t = (C, D)$. The figure shows beliefs over "terminal nodes" conditional on action plan $(a_i^\tau)_{\tau=t}^{t+1} = (D^t; D^{t+1})$. Consequently all "terminal nodes" that involve player i choosing C have probability zero under this plan. All other nodes may receive positive probability depending on player i's beliefs at t.

2.5. Discussion

As in many other learning models, our agents form beliefs by sampling from past interactions in the population and then best respond to these beliefs. A novelty in our model is that (i) agents are not myopic, i.e. they form beliefs also about future paths of play (nodes at distance k) and (ii) they condition their beliefs about their opponent's choice on the history of play ($h > 0$). In this subsection we discuss these two new assumptions. Standard models of myopic agents (e.g. Young, 1993) implicitly or explicitly assume that $h = 0$, i.e. that while agents learn from the history of play, they do not condition their beliefs on the (recent) history of play. It is important to note, though, that there is no conceptual discontinuity between the cases $h = 0$ and $h > 0$. In particular agents are *not* strategic under either model since they do *not* reason about the beliefs of their opponent but instead *learn* about the opponent's choices. One could think of the difference between the two models as a difference in the theory about the opponent. For example agents could view their opponent as a one-state automaton in the myopic case ($h = 0$) and as a multi-state automaton in the $h > 0$ case. An alternative interpretation could be that agents have the same "theory" about the opponent in both cases, but that h simply reflects their own reasoning constraints. Note that in the most sophisticated case $h = T$, agents would learn the "full strategies" of their opponents, i.e. they would learn a different belief for each decision node in the game. If $h < T$, this is not the case. Instead, in these cases, agents implicitly (and endogenously) categorize nodes according to the recent history of play, i.e. they form the same beliefs for every node that is preceded by the same history (of length h). In either case they treat all nodes equal – irrespective of whether they are at the beginning or end of the game – as long as they are preceded by the same history of play. (If $h = T$ then no two nodes will ever be preceded by the same history and hence all nodes will be distinguished.)

2.6. Techniques

State: The state at time t is given by the tuple

$$s^t := (M^t, H^t),$$

where H^t is the history at t and M^t the collective memory for both player classes. (See the definitions in Sections 2.2 and 2.3). Since memory m is finite and all decision rules are time-independent the process can be described by a stationary Markov chain on the state space $S = S_1 \times S_2$ where $S_i = (A_{-i}^m)^{\mathcal{H}} \times \mathcal{H}$ with transition matrix P. P has entries $P(s, s')$, that describe the probability to move from state $s \in S$ to state $s' \in S$. In Appendix A we provide more details about P.

Definition 1 *(Absorbing set).* A subset $X \subseteq S$ is called absorbing if $P(s, s') = 0, \forall s \in X, s' \notin X$.

In Section 3.1 we will characterize absorbing sets. Naturally, the question arises whether some absorbing sets are more likely to arise if the process is subjected to small perturbations. Let $P^\varepsilon(s, s')$ denote the transition matrix associated with the perturbed process in which players choose according to decision rule (2) with probability $1 - \varepsilon$ and with probability ε choose an action randomly (with uniform probability) from A_i.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

The perturbed Markov process $P^\varepsilon(s, s')$ is ergodic, i.e., it has a unique stationary distribution denoted by f^ε. This distribution summarizes both the long-run behavior of the process and the time-average of the sample path independently of the initial conditions.[11] The limit invariant distribution $f^* = \lim_{\varepsilon \to 0} f^\varepsilon$ exists and its support $\{s \in S | \lim_{\varepsilon \to 0} f^\varepsilon(s) > 0\}$ is a union of some absorbing sets of the unperturbed process. The limit invariant distribution singles out a stable prediction of the unperturbed dynamics ($\varepsilon = 0$) in the sense that for any $\varepsilon > 0$ small enough the play approximates that described by f^* in the long run. The states in the support of f^* are called stochastically stable states.

Definition 2. State s is stochastically stable $\Leftrightarrow f^*(s) > 0$.

We will characterize stochastically stable states in Section 3.2.

3. Results

3.1. Young's theorem (1993)

Before we move on to our results we would like to remind the reader of the result by Young (1993) corresponding to the case where $(h, k) = (0, 1)$, i.e. to the case where all agents are myopic (have foresight $k = 1$) and form beliefs without conditioning on the history ($h = 0$). Young considers a situation where $T = 1$, i.e. a case where actions and strategies coincide. Define the best reply graph of a game Γ as follows: each vertex is a tuple of action choices, and for every two vertices $\vec{a}$ and $\vec{a}'$ there is a directed edge $\vec{a} \to \vec{a}'$ if and only if $\vec{a} \neq \vec{a}'$ and there exists exactly one agent i such that a'_i is a best reply to a_{-i} and $a_{-i} = a'_{-i}$.

Definition 3. A game Γ is acyclic if its best reply graph contains no directed cycles. It is weakly acyclic if, from any initial vertex $\vec{a}$, there exists a directed path to some vertex $\vec{a}^*$ from which there is no exiting edge.

For each action-tuple, let $L(\vec{a})$ be the length of a shortest directed path in the best reply graph from $\vec{a}$ to a strict Nash equilibrium, and let $L_\Gamma = \max L(\vec{a})$.

Theorem 1. *(Young (1993))*
If Γ is weakly acyclic, $(h, k) = (0, 1)$, and $\rho \leq m/(L_\Gamma + 2)$ then the process converges almost surely to a point where a strict Nash equilibrium is played at all t.

The theorem by Young (1993) shows that in this special case of our model only strict Nash equilibria of the one shot game will be observed in the long run (in games with an acyclic best reply graph).

3.2. Absorbing sets

Now let us move to the case where $k > 1$. We will make the following assumption throughout.
Assumption A1 $h, k \leq (T/2)$.
This assumption will simplify the proofs considerably and some upper bound on h (or k) is crucial for some results as we will see. The bound assumed here is not tight. We will start by analyzing absorbing states. Recall that we defined a state to be a collection $s^t := (M^t, H^t)$. We are interested in characterizing behavior (action choices) that can be sustained in an absorbing state. In our discussion we will hence focus largely on what we call "pure absorbing states", which are states in which one particular action profile is induced "most of the time". More precisely we define a pure absorbing profile $\vec{a}^* = (a_1^*, a_2^*)$ as follows:

Definition 4. We say a profile $\vec{a}^*$ is *(pure) absorbing* if there exists an absorbing set $X \subset S$ and an integer $\lambda \in \{0, \ldots, k-1\}$ such that, in each state $s \in X$ and in each T-period interaction, $\vec{a}^*$ is played in $T - \lambda$ consecutive periods.

If a set $X \subset S$ induces a pure absorbing profile we will also refer to this set as pure absorbing. The intuitive reason why we want to allow pure absorbing states to be such that a different profile can be played in some periods is that forward-looking learning may be able to sustain some additional profiles (compared to myopic learning) as long as the time horizon is large enough, but not when the end of the interaction is near. We now proceed to characterizing such pure absorbing profiles.
It is intuitive (and non-surprising given what we know about the myopic case) that most Nash equilibria of the one-shot game can be absorbing.[12] To characterize absorbing profiles which involve outcomes that are not Nash, the following definition will be useful.

Definition 5. We call an action profile $\vec{a}^* = (a_i^*, a_{-i}^*)$ *locally efficient* if

[11] See for example the classic textbook by Karlin and Taylor (1975).
[12] Whenever we talk of (Non-)Nash actions, pareto efficient outcomes or curb sets (below), we always refer to the one shot game.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Table 1
Two games. Local efficiency of (A, a) is satisfied in Game 1, but not in Game 2.

Game 1	a	b	c
A	3.3	0.5	5.0
B	5.0	1.1	0.0
C	0.5	0.0	4.4

Game 2	a	b	c
A	3.3	0.5	5.6
B	5.0	1.1	0.2
C	6.5	2.0	4.4

(1) all unilateral deviations from $\vec{a}^*$ strictly hurt at least one player
(2) there exists a set $A' \subseteq (A_1 \times A_2)$ s.t. $\vec{a}^*$ is pareto efficient within A' and A' is closed under best replies to all beliefs $\mu \in \Delta A_{-i}'$ placing at least probability $1-\rho^{-1}\lceil m/T\rceil$ on a^*_{-i}, $\forall i = 1, 2$ and
(3) $\forall i : \exists\, a_{-i} \in A_{-i}'$ such that $\pi_i(a_{i'}, a_{-i}) < \pi_i(\vec{a}^*)$, $\forall a_{i'} \neq a^*_i$.

Part (1) of the definition of a "locally efficient profile" ensures locally efficiency is a "strict" criterion, in the sense that there exists a player i for which $\pi_i(a_i, a^*_{-i}) < \pi_i(\vec{a}^*)$, $\forall a_i \neq a^*_i$, i.e. for which a unilateral deviation leads to strictly lower payoffs or "strictly hurts the player". Part (2) is very close to the notion of a curb set (short for "closed under rational behavior") introduced by Basu and Weibull (1991). Essentially a subset of strategies in a normal form game is curb whenever the best replies to all the probability mixtures over this set are contained in the set itself. In more technical language a curb set is a non-empty product set $A' = \times_{i=1,2} A_i \subset A$ s.t. for each $i = 1, 2$ and each belief $\mu \in \Delta A_{-i}'$ of player i the set A_i contains *all* best responses of player i against this belief, i.e. $\forall i = 1, 2, \forall\, \mu \in \Delta A_{-i}' : BR_i(\mu) \subset A_i'$. Obviously any game $(A_1 \times A_2)$ is a curb-set itself, strict Nash equilibria are (minimal) curb-sets but also the set $A' = (A, B) \times (a, b)$ in Game 1 above is curb. Note that, since all $A_1 \times A_2$ are curb sets by definition, any profile that is pareto efficient in some game automatically satisfies Condition (2). The condition is weaker than pareto efficiency in a curb-set, since it requires closure only to beliefs placing at least probability $1-\rho^{-1}\lceil m/T\rceil$ on a^*_{-i}. (Remember that $\lceil m/T\rceil$ denotes the smallest integer bigger than (m/T)). The reason that Condition (2) does not require A' to be closed to all beliefs is as follows. Given the structure of pure absorbing profiles, a history of $\vec{a}^*$ is followed by a choice $a_{-i} \neq a^*_{-i}$ at most once in each T-period interaction and at most $\lceil m/T\rceil$ such instances will be remembered. As a consequence, given sample size ρ, agents will at an absorbing state always hold beliefs that – conditional on a history of $\vec{a}^*$ – place probability of at least $1-\rho^{-1}\lceil m/T\rceil$ on a^*_{-i} and it is under those beliefs that A' has to be closed.[13] Part (3) requires that for any deviation there should exist an action of the opponent that yields always worse payoffs to a player than $\pi_i(\vec{a}^*)$. Note that Conditions (1) and (3) together imply Condition (2) in a 2×2 game.

Table 1 shows two examples illustrating local efficiency. In Game 1 the action profile (A, a) can be sustained in a pure absorbing state despite the fact that it is not pareto efficient in the whole game. Such an absorbing state could be sustained by beliefs where $\mu_1(c|(C, .), .)$ is "small" and $\mu_1(b|(B, .), .)$ is "high enough". Condition (2) is satisfied in Game 1. In Game 2 (A, a) cannot be sustained, since $(A, B) \times (a, b)$ is not curb. In fact, the myopic best response to any belief with support on $(A, B) \times (a, b)$ is $C(c)$. But this means that "small" beliefs $\mu_1(c|(C, .), .)$ cannot be sustained. Condition (2) fails in this game. Local efficiency will matter for profiles which are not Nash. All Nash equilibrium profiles (a^*_i, a^*_{-i}) can be induced as long as the following Condition is satisfied:

Condition C1. $\forall i$ and $a_{i'} \neq a^*_i : \exists a_{-i} \in A_{-i}$ such that $\pi_i(a_{i'}, a_{-i}) < \pi_i(a^*_i, a^*_{-i})$

Obviously strict Nash equilibria satisfy C1, but even Nash equilibria in weakly dominated strategies will typically satisfy this requirement. With this observation we can state the following proposition.

Proposition 1. *Assume* $(h, k) \gg (0, 1)$. *There exists a real number* $\eta(h, k) > 0$ *such that a profile that can be reached with positive probability is pure absorbing if and only if it is either (i) a Nash equilibrium satisfying C1 or (ii) if it is locally efficient and* $\rho^{-1}\lceil m/T\rceil \leq \eta(h, k)$.

Proof. Appendix B. □

Proposition 1 shows that both Nash equilibria as well as profiles which are not Nash equilibria can be induced in pure absorbing states provided that they are efficient in a sense defined above. An example is cooperation in the Prisoner's Dilemma. If agents learn that their opponent takes actions with worse payoff consequences for them with higher probability after a history of Nash play than after a history of efficient (but possibly non Nash) play, then they will have incentives to refrain from choosing myopic best responses at least in early stages of a repeated game. More loosely speaking agents will

[13] Note that not all beliefs placing a higher probability than $1-\rho^{-1}\lceil m/T\rceil$ on a^*_{-i} can be drawn from the finite sample. However, if A' was not closed under some of these, it would also not be closed under some of those that can be drawn by continuity.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

anticipate that taking "aggressive" actions (like e.g. defection in the Prisoner's Dilemma) can deteriorate future relations, which is why they refrain from doing so in early stages of the repeated interaction. The forward looking part is crucial here. If myopic agents simply learned which strategies have yielded good payoffs in the past (e.g. via reinforcement learning), then efficient (but Non-Nash) profiles could *not* be absorbing. The only reason why players refrain from taking unilateral deviation that are profitable in the short run is that they take future payoffs into account and anticipate that the opponent's behavior is not stationary.

Some conditions are needed to obtain this result. The condition on $\rho^{-1}\lceil m/T\rceil$ ensures that samples remain informative enough. $\lceil m/T\rceil$ is a measure of the maximal number of "rare" or "untypical" events (read action choices other than a^*_{-i}) contained in an agent's memory at any time conditional on a history of $\bar{a}^*$. If ρ is too small compared to this expression, then it is possible that such "rare" events are over-represented in the sample on the basis of which agents form beliefs. This can destabilize the efficient absorbing profile. Note also that in Proposition 1 we have focused on pure absorbing profiles that can be reached with positive probability. The latter condition rules out states that are supported by off-path beliefs which are inconsistent with the learning process described in Section 2.

The threshold $\eta(h,k)>0$ is strictly increasing in k and not always monotone in h. The intuition for k is straightforward. The more forward looking agents are, the more do future payoffs matter for today's decisions. If future payoffs matter enough, then agents may refrain from choosing myopic best responses. The role of h is more subtle. If $h=0$, then agents do not condition their beliefs on the history of play and hence will hold the same belief at all decision nodes in the game. On the other hand if h were very large (in particular $h\geq T-1$), then histories would be of different length and hence necessarily different at all decision nodes. In this case agents will condition their beliefs on the decision node. But then only Nash equilibria (of the one shot game) are absorbing. The interesting cases are those with intermediate h, where agents implicitly (and endogenously) categorize nodes according to the recent history of play. In Section 3.4 we will see how these conditions play out in a numerical application to a Prisoner's Dilemma. This example will also illustrate that the conditions on ρ, m and T are "reasonable" in a typical game. Note that the exact value of $\eta(h,k)>0$ will also depend on payoff parameters of the game. We have omitted this dependency from the argument of η for notational clarity.

Note also that the result in Proposition 1 does not depend on there being a discrepancy between Nash and minmax outcomes in the game, nor per se on the time horizon being sufficiently long, nor on there being a multiplicity of Nash equilibria in the stage game. The result and the underlying intuition are thus fundamentally different from the standard repeated games literature. Proposition 1 implies for example that paths involving cooperation in the Prisoner's Dilemma can be absorbing under certain conditions. Such paths cannot be sustained, though, by standard folk theorems for finitely repeated games.

The following result shows that starting from a state which is *not* absorbing the process converges with probability one to one of the pure absorbing sets in acyclic games.

Proposition 2. *Assume the game is acyclic. Then, starting from any state which is not absorbing, the process converges almost surely to a pure absorbing set.*

Proof. Appendix B. □

The intuition is as follows. Since beliefs are formed by drawing imperfect samples from the past there is always positive probability to draw "favorable" beliefs which enable convergence after finitely many periods. This is only true for acyclic games. In games with best response cycles, such as e.g. the matching pennies game convergence to a pure absorbing state cannot be ensured and in fact pure absorbing states may even fail to exist in such games. In such cyclic games the process need not converge. Note also that the corresponding theorem in Young (1993) requires ρ to be "small enough" relative to m. This is needed in Young (1993) because agents sometimes need to be able to look back far enough to obtain a homogeneous sample. Because of the assumption in our model that memories are history dependent, i.e. that agents remember m instances for *each* history, the possibility of drawing homogeneous samples is guaranteed as long as $m\geq\rho$ which is satisfied by definition.

Proposition 2 establishes that the stochastic process converges with probability one to a pure absorbing set starting from a state which is not absorbing. Note that Propositions 1 and 2 do not imply that there may not be other absorbing sets which are not *pure* absorbing. In fact in many games of interest such states will exist.[14] However Proposition 2 shows that as soon as agents deviate slightly from such a state (to a non absorbing state) they will almost surely converge to a pure absorbing set. A natural question that arises is whether some absorbing sets are more likely to be observed in the long run than others. In the next subsection we will analyze which of the absorbing states are also stochastically stable.

[14] For example in a 2×2 pure coordination game states in which players alternate between the two equilibria are also absorbing. From any state "close" to those (where memory conditional on *either* of the pure histories contains *both* actions) the process will always surely converge to a pure absorbing state. The reverse is not true. From states "close" to a pure absorbing state, the process may not converge to such an alternating state, which is the case e.g. if the memory conditional on *each* pure action profile (history) contains that action only.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

3.3. Stochastically stable states

For our analysis of stochastically stable states we will focus on specific 2×2 games. Consider the following payoff matrix

	C	D
C	α, α	$0, \beta$
D	$\beta, 0$	γ, γ

. (3)

If $\beta > \alpha > \gamma > 0$ this matrix represents a Prisoner's Dilemma and if $\alpha > \beta$ and $\gamma > 0$ it represents a Coordination game. We will focus on the different cases in turn. Let us also assume that $\beta < 2\alpha$.[15] We adopt the notational convention that $\vec{C} = (C, C)(\vec{D} = (D, D))$ is the profile where action $C(D)$ is chosen by both agents.

3.3.1. Prisoner's Dilemma

Before we start our analysis of stochastically stable states, let us first describe the set of absorbing states for this game. States involving defection (D) in all periods can be absorbing by Proposition 1. (Since (D, D) is a strict NE of the one-shot game, it satisfies condition C1). The more interesting question is under which conditions states involving cooperation in some periods can be absorbing. Since cooperation is pareto efficient we know from Proposition 1 that such conditions will exist. Our first observation is the following.

Proposition 3. *The paths of play induced by absorbing sets involving cooperation satisfy non-increasing cooperation (NIC), i.e. they are such that* $\forall t$ *with* $t - l \neq 0 \bmod T$, $\forall l = 1, \ldots, h$ *the following is true: if* $a_i^t = C$ *then also* $a_i^{t-1} = C$.

Proof. Appendix B. □

Proposition 3 states that the probability to observe cooperation *within* a given T-period game is non-increasing in t (with the possible exception of early periods where histories are of length $<h$). This is intuitive, since cooperation (being efficient but dominated in the one shot game) can only be sustained if agents believe that defecting will lead to a higher probability of defection by their opponent in the future than cooperating. For any given degree of forward-looking k the perceived payoff loss from defection will be smaller the closer agents are to the end of their interaction T. Hence if agents find it in their interest to cooperate at t, they must also do so at $t-1$ (within the same T-period interaction).[16] Absorbing states in the Prisoner's Dilemma (PD) are denoted as follows. Let X_D be the absorbing set which induces defection in each period and denote by X_C the absorbing set which induces cooperation in some periods of every T-period interaction. All states $s \in X_C$ must satisfy the property NIC.

Proposition 4. *If* $(h, k) > (0, 1)$, $\rho^{-1}\lceil m/T \rceil < ((\alpha - \gamma)/\alpha)$ *and* $((\rho - 1)/\rho) \geq ((\alpha + 2\beta - 3\gamma)/(\alpha + 2\beta - 2\gamma))$, *then all stochastically stable states are contained in* X_C.

Proof. Appendix B. □

Two conditions are needed for this result. The condition $\rho^{-1}\lceil m/T \rceil < ((\alpha - \gamma)/\alpha)$ ensures that samples are "informative" enough such that agents' beliefs conditional on histories containing only $\vec{C}$ place high enough probability on the opponent choosing cooperation again. This is a necessary condition, which is needed for states in X_C to be absorbing at all. The condition $((\rho - 1)/\rho) \geq ((\alpha + 2\beta - 3\gamma)/(\alpha + 2\beta - 2\gamma))$ is sufficient to both prevent too "easy" transitions from any state in X_C to a state in X_D by ensuring that few trembles to defection are never enough to infect a pair of agents. On the other hand it is sufficient to enable "easy" transitions from any state characterized by defection to a state characterized by cooperation.

More loosely speaking the intuition is as follows. Transitions away from cooperative states are hard, since as long as it is in people's mind that the opponent responds to a history of joint cooperation by cooperating they will always have incentives to start new relations by cooperating. But this belief is very hard to destabilize since once a tremble to defection has occurred the history is not one of joint cooperation anymore. Transitions to cooperative states are easier, because once agents have experienced successful cooperation in one particular T-period interaction they will be willing to start new relationships by cooperating.

[15] This condition makes sure that cooperation is more efficient than players alternating between (C, D) and (D, C). For $k = 2$ such alternating states are not absorbing even without this condition. If an agent anticipates that – after a history ending with (D, C) – her opponent will defect, then she will not have an incentive to cooperate no matter what her beliefs about the opponent's choice after different histories are. For larger k one would need a progressively tighter condition to ensure that such states are not absorbing. If cooperation is efficient, though, i.e. if $\beta < 2\alpha$, then such alternating states are never absorbing.

[16] Ghosh and Ray (1996) have studied a setting where matching is not random but where agents can choose their interaction partners. Furthermore in their setting agents are (i) strategic and (ii) heterogeneous in the sense that some players have discount factor zero and some a strictly positive discount factor for payoffs obtained in the repeated game. Interestingly their characterization of equilibria comes closer to a property of non-decreasing cooperation rather than non-increasing cooperation as in our setting. In our setting non-increasing cooperation obtains because limited forward looking agents act as if they were "more myopic" towards the end of an interaction. In their setting non-decreasing cooperation obtains because agents test the willingness to cooperate of their match and continue to cooperate if their match has a high discount factor. Endogenous choice of who to play with guarantees that incentives are aligned in their setting.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Table 2
Average frequency of cooperation in last two 10-period interactions.

	1	2	3	4	5	6	7	8	9	10
Partner	**0.86**	**0.72**	**0.68**	**0.66**	**0.59**	**0.61**	0.34	0.29	0.07	0.04
Computer50	**0.64**	**0.72**	**0.68**	**0.71**	**0.71**	**0.65**	**0.61**	**0.65**	0.29	0.11

Periods are highlighted in bold, where cooperation would be expected according to the theoretical predictions outlined in this section.

Some remarks are at order. First notice that if the conditions are not satisfied then (depending on the parameters) stochastically stable states can be contained in either X_C or X_D. Note also that the conditions are not tight bounds, since we require in the proof that the maximal number of trembles needed for transitions from any state in X_C to a state in X_D requires less transition than the minimal number of transitions needed from any state in X_D to X_C. Since this kind of computation includes all the states, even those through which no minimal mutation passes, the bound is generally not tight, which is also the reason that it does not depend on h or k.

3.3.2. Coordination games

Since in the coordination game all locally efficient profiles are Nash equilibria which satisfy C1, pure absorbing sets induce either C or D at all periods. Denote these two absorbing sets by X_C and X_D, respectively. To make the problem more interesting, let us assume that additionally $\beta+\gamma>\alpha>\gamma$, implying that (C, C) is efficient and D is risk-dominant in the one-shot game. The question we then want to answer is: how does our adaptive learning process select among risk-dominance and efficiency if agents are forward-looking ? Young (1993) has analyzed this question for 2×2 games in the case where $(h, k)=(0, 1)$ and has found that risk-dominant equilibria are the only ones that are stochastically stable in this setting. In the presence of forward looking agents this is in general not the case as the following result shows.

Proposition 5. *There exists $\widehat{\rho}(\beta, \alpha, \gamma)$ such that whenever $\rho\geq\widehat{\rho}(\beta, \alpha, \gamma)$ and $(h, k)>(0, 1)$ all stochastically stable states are contained in X_C.*

Proof. Appendix B. □

The intuition is as follows. A unilateral tremble starting from a state in X_D is not as detrimental (yielding a payoff of $\beta>0$) as a tremble starting from the efficient equilibrium (yielding a payoff of zero) in the short run. If it is the case, though, that the opponent is likely to react to such a tremble by changing his action, then trembles starting from the efficient action can be less detrimental than those starting from the risk dominant action in the medium run. Forward looking agents will take this into account. There is also a second effect which favors the efficient convention, which is that agents will always be willing to start out new relationships by playing C even if in their previous relationship they converged to D as long as they are sufficiently convinced that a history of $\vec{C}, \ldots, \vec{C}$ will be followed by cooperation. Eliminating this belief requires many trembles. Hence, unlike in the myopic case, efficient outcomes can be part of an absorbing state in these two classes of games.

3.4. Application to experimental results

In this subsection we illustrate how the results from the previous subsection (in particular Section 3.3.1) can explain typical experimental results. An experiment that is relatively well suited to test our theory was conducted by Andreoni and Miller (1993). In their "Partner treatment" subjects were randomly paired to play a 10-period repeated prisoner's dilemma with their partner ($T=10$). They were then randomly rematched with another partner for another 10-period game. This continued for a total of 20 such 10-period games, i.e. for a total of 200 periods of the prisoner's dilemma. The payoffs in the Prisoner's Dilemma in their experiment were given by $\alpha=7$, $\beta=12$ and $\gamma=4$ (Table 3).

The second treatment we are interested in is the treatment they call "Computer50". This treatment coincides with "Partner," except that subjects had a 50% chance of meeting a computer partner programmed to play the "Tit-for-Tat" strategy. In the language of our model a "Tit-for-Tat" player is characterized by a level of sophistication $h=1$ and always mimics the action of the opponent in the previous period.

Table 2 shows the average cooperation rate in the last two 10-period interactions, where there are most chances that the learning process has converged. What is interesting about these results is (i) that the property of NIC seems satisfied on average, (ii) that there is a sharp drop after 6 periods in Partner treatment and that (iii) this sharp drop occurs two periods later in the Computer50 treatment. The results display two typical patterns of repeated Prisoner's Dilemma experiments. The sharp drop at the end is often referred to as "endgame effect" and the fact that cooperation rates are high again in initial periods of the next T-period interaction is often referred to as "restart effect".

We next ask whether we can explain their findings from both treatments with one common set of parameters of our model. Our sufficient condition to rule out defection as a stochastically stable state yields $\rho\in(2, 9]$ and $\rho^{-1}\lceil(m/10)\rceil<(3/7)$. This is satisfied e.g. if $\rho=5$ and $m=10$. But since we do not know ρ and m, we cannot rule out that both cooperative states and states characterized by defection might be stochastically stable. We start by analyzing the "Partner"-treatment. First note that the Condition from Proposition 4 boils down to $(\rho-1/\rho)\geq(19/23)$, which is the same as saying $\rho\geq 6$. We can state the following result.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Table 3
Frequencies with which cooperation was chosen in the experiment conditional on 1-period history and the (sufficient) restrictions on beliefs stemming from the theory. History of play in the table has the format (a_i, a_{-i}).

	Partner treatment			
	CC	CD	DC	DD
Pr(C)-Exp (Periods 1–180)	0.89	0.23	0.38	0.07
Pr(C)-Exp (Periods 81–180)	0.89	0.20	0.44	0.06
Pr(C)-Exp (Periods 1–100)	0.88	0.23	0.34	0.07
$\mu(C\|.)$-Theory	$\geq$0.83	–	$\in$[0, 0.47]	0
	Computer50 treatment			
	CC	CD	DC	DD
Pr(C)-Exp (Periods 1–180)	0.88	0.48	0.16	0.08
Pr(C)-Exp (Periods 81–180)	0.89	0.49	0.11	0.07
Pr(C)-Exp (Periods 1–100)	0.88	0.48	0.18	0.10
$\mu(C\|.)$-Theory	$\geq$0.76	–	$\in$[0.10, 0.54]	0

Proposition 6. *If $(h, k) = (1, 5)$, $\rho \geq 6$ and $\rho^{-1}\lceil (m/10)\rceil < (3/7)$ the path of play were agents cooperate in the first six periods of all T-period interactions and defect afterwards is induced in the unique stochastically stable state.*

Proof. Appendix B. □

Hence for a level of sophistication $h = 1$ and degree of forward looking $k = 5$ our model can rationalize this path of play.[17] What can we say about the beliefs required to sustain such a state? If m is not too large (in fact $m \leq 13$), this path of play induces beliefs $\mu(C|(C, C)) \geq 5/6$ and $\mu(C|(D, D)) = 0$. There are also some restrictions on off path beliefs. Table 3 shows the theoretically required beliefs and empirical frequencies in the first 100 periods of play. If participants do form beliefs by relying on empirical frequencies, as suggested by the theory, then our learning process can provide an explanation for their results.

Still our model has quite some free parameters. And of course we did choose parameters ($(h, k) = (1, 5)$ and $\rho \geq 6$) that – while appearing intuitively reasonable – can explain these data rather than choosing parameters at random. A better test of the theory is whether we can explain the data from a *different* treatment using the *same* parameters. In order to do this we consider the Computer50 treatment described above. Holding fixed the degree of forward looking for all agents, agents should have stronger incentives to cooperate in this case. The following proposition confirms this intuition.

Proposition 7. *If $(h, k) = (1, 5)$, $\rho \geq 6$ and $\rho^{-1}\lceil (m/10)\rceil < (3/7)$ and if there is a 50% chance of meeting a tit-for tat (computer) player the path of play were agents cooperate in the first eight periods of all T-period interactions and defect afterwards is induced in the unique stochastically stable state.*

Proof. Appendix B. □

If $m \leq 19$ this path induces beliefs $\mu(C|(C, C)) \geq 7/8$ and $\mu(C|(D, D)) = 0$, which is consistent with the empirical frequencies (see Table 3).[18]

Finally we ask whether individual decisions can be explained using our theory. We will consider three measures: (i) which percentage of participants satisfy the property of non-increasing cooperation (NIC) and hence are consistent with our theory for *some* k and h, (ii) which percentage of participants behave exactly in accordance with our theoretical prediction (for $h = 1, k = 5$) or cooperate one period longer or less long and (iii) whether the modal behavior coincides with our theoretical prediction ($h = 1$, $k = 5$).

Table 4 shows the results. In both treatments the modal behavior exactly coincides with our theoretical prediction. 86% of participants satisfy NIC in the Partner treatment and 77% in the Computer50 treatment. Not only aggregate behavior but also the distribution of individual behaviors responds to the treatment change in the direction predicted by the theory of limited forward looking players. Note also that, while just short of 50% of individual behavior coincides with the theoretical prediction (± 1) of our model, less than 20% of behavior is consistent with Nash equilibrium (+2) in the Partner treatment.

4. Heterogeneous agents

We ask whether agents with a higher degree of forward-looking (k) will always be able to exploit others with a lower degree of forward looking, i.e. whether there is an evolutionary sense in which agents should be more or less forward looking. We consider the following simple example. Assume that there are two types. k_1 is a myopic type with $(h, k) = (1, 1)$ and k_2

[17] One could also explain this path with higher values of h, but we find it most convincing to use the most simple decision rule (involving least sophistication).

[18] Note that cooperating until the opponent defects or until period 8 (whichever comes first) and defecting afterwards is also a sequential equilibrium of this game (Kreps et al., 1982). Cooperating in the Partner treatment, however, cannot be part of a sequential equilibrium.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Table 4
Percentage of 10-period behaviors that are in accordance with theory (for parameters $(h, k) = (1, 5)$, $\rho \geq 6$) in periods 181–200. LFP stands for "learning by limited forward looking players".

	Partner	Computer50
All C	0.04	0.04
(C,C,C,C,C,C,C,C,C,D)	0.04	0.18
(C,C,C,C,C,C,C,C,D,D)	0.11	0.25
(C,C,C,C,C,C,C,D,D,D)	0.18	0.04
(C,C,C,C,C,C,D,D,D,D)	0.25	–
(C,C,C,C,C,D,D,D,D,D)	–	–
(C,C,C,C,D,D,D,D,D,D)	0.04	–
(C,C,C,D,D,D,D,D,D,D)	0.04	–
(C,C,D,D,D,D,D,D,D,D)	0.04	–
(C,D,D,D,D,D,D,D,D,D)	0.14	0.07
All D	–	0.04
Other	0.14	0.35
Satisfy NIC ($h = 1$)	0.86	0.77
Theory prediction (LFP) ± 1	0.43	0.48
Modal behavior = theory (LFP)	Yes	Yes

is forward-looking characterized by $(h, k) = (1, 2)$. Denote the share of k_1 agents by σ. Irrespective of their type and class, agents are randomly matched to play a 4-period repeated Prisoner's Dilemma. The stage game payoffs are given by the payoff matrix (3). We want to consider two different scenarios. In the first agents know that the population is heterogeneous and are able to observe the type of their match at the end of an interaction, store this information in their memory and thus to form conditional beliefs. In the second scenario agents are not able to form conditional beliefs. The reason could be either that they (wrongly) assume that the population is homogeneous or that they are simply never able to observe (or infer) the type of their opponent.

4.1. Conditional beliefs

In this scenario all agents are aware that the population is composed of two different types and hence can react to this knowledge. In particular forward-looking types can update their priors on the type they are facing (and thus their conditional beliefs about behavior in future periods) depending on the behavior they observe in earlier periods. Remember that σ is the population share of myopic (k_1) types.

Proposition 8. *If $\sigma < ((3\alpha - \beta - 2\gamma)/(3\alpha - \beta - \gamma))$, then forward looking agents (k_2) obtain higher average payoffs in all absorbing states. If $\sigma \in [((3\alpha - \beta - 2\gamma)/(3\alpha - \beta - \gamma)), ((3\alpha - \beta - 3\gamma)/(3\alpha - \beta))]$ then myopic agents (k_1) obtain higher average payoffs in all absorbing states and if $\sigma > ((3\alpha - \beta - 3\gamma)/(3\alpha - \beta))$ all agents obtain the same average payoff in all states.*

Proof. Appendix B. □

The condition $\sigma < ((3\alpha - \beta - 3\gamma)/(3\alpha - \beta))$ is simply necessary for absorbing states with cooperation to exist at all. If the condition is not met, i.e. if there are too many myopic types who always defect, then all absorbing states will display full defection. Given that absorbing states with cooperation do exist, forward looking agents do only make higher profits in expectation if σ is not too high. Else myopic agents do make higher payoffs in these states. The reason is that when forward-looking agents decide on their action choice they expect to be able to exploit a cooperative opponent in later periods of of their horizon ($t+1, \ldots t+k$). But this is not true in an absorbing state, since other forward looking types do reason in the same way. Consequently they overestimate the relative benefit of cooperation and choose cooperation in a range of σ where they should be choosing defection.

These results have natural implications in terms of evolution. In particular they show that evolution need not eliminate myopic players, but that states where $\sigma \geq ((3\alpha - \beta - 2\gamma)/(3\alpha - \beta - \gamma))$ can be stable in an evolutionary model. Which states will be stable will depend of course on the precise evolutionary model considered. Finally note that if matching were assortative, i.e., if forward looking types were matched with increased probability with other forward-looking types and vice versa, forward-looking types will tend to have higher payoffs on average.[19]

4.2. Unconditional beliefs

In the case where agents are not able to infer the type of their opponents (or simply assume that the population is homogeneous) and thus form beliefs that are not conditional on the type of their opponent. In this case the only absorbing state involves full defection, as the following Claim illustrates.

[19] See e.g. Myerson et al. (1991) or Mengel (2007, 2008) for models of assortative matching in the prisoner's dilemma.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Proposition 9. *If beliefs are unconditional all absorbing states involve full defection and all agents obtain the same payoff in expectation.*

Proof. Appendix B. □

The intuition is simply that if forward-looking types are repeatedly matched with myopic types their beliefs will eventually decrease below the cooperation threshold. But given this, there is positive probability that even a small number of myopic types can induce the beliefs of all forward-looking types to decrease. In such states forward-looking types might still have high beliefs about the cooperation probability following a history of joint cooperation (since myopic types never cooperate). The problem is that their beliefs about initial cooperation (after the null history) and about cooperation after unilateral cooperation will be too low to induce cooperative outcomes. The lack of strategic reasoning is in this case responsible for them *not* being able to restore cooperative outcomes.

5. Conclusions

We studied agents interacting in finitely repeated games, who are adaptive, but also forward-looking to some degree. We have shown that in a pure absorbing set either Nash equilibria satisfying very weak conditions or locally efficient profiles can be induced. In 2×2 prisoner's dilemma and coordination games there are parameter conditions under which only the efficient outcomes are induced in stochastically stable states. We have also seen that these results can provide explanations for common findings in experiments, such as cooperation in finitely repeated games, the "endgame effect" and the "restart effect"

A number of other papers have shown that cooperation in the prisoner's dilemma can arise as the outcome of a learning process (see e.g. Karandikar et al., 1998 or Levine and Pesendorfer, 2007. A recurrent pattern in these papers seems to be that the rationality of agents has to be "bounded enough" in order to achieve cooperation. In particular agents are not allowed to choose best responses in these models. In the present paper, on the other hand, agents are allowed to be quite rational. In particular they are more sophisticated than myopic best response learners. Still they are able to achieve cooperation.

Further research could build in Section hyperlinkTDSEC:44 and study under which conditions forward looking behavior emerges as a result of evolutionary selection. It seems also worthwhile to test forward-looking behavior experimentally to distinguish this from other possible explanations of the "endgame" and "restart" effects in social dilemma games.

Appendix A. The transition matrix

Denote by $H(s)$ the history associated with state s and by $M_i(H(s))$ the memory of a player in class i associated with that history and let $M(H(s)) = (M_1(H(s)), M_2(H(s)))$. Call $\widehat{s}$ a successor of $s \in S$ if $\widehat{s}$ is obtained from s by (i) deleting the first coordinate from $M_i(H(s))$ (if $|M_i(H(s))| = m$) and by adding a new element $r_i(\widehat{s})$ to the right (i.e. as m-th coordinate) and (ii) by deleting the first coordinate of $H(s)$ (if $|H(s)| = h$) and by adding $\vec{r}(\widehat{s}) = (r_1(\widehat{s}), r_2(\widehat{s}))$ as h-th coordinate or (if $t = 0 \mathrm{mod} T$) by setting $H(s) = H^0$. The learning process can then be described by a transition matrix $P \in \mathcal{P}$ where $\mathcal{P}$ is defined as follows.

Definition (Transition matrices) Let $\mathcal{P}$ be the set of transition matrices P that satisfy $\forall s, s' \in S$:

$$P(s, s\prime) > 0 \quad \Leftrightarrow \begin{cases} s\prime \text{ is a successor of } s \text{ and} \\ r_i(s\prime) \in \mathcal{BR}_i^t(\mu_i(H(s))). \end{cases}$$

Appendix B. Proofs

Remember that we denoted by $BR_i(\cdot)$ player i's best response correspondence for the one shot game. We also denoted by $\mathcal{BR}_i^t(\cdot)$ the instantaneous best response of player i for the repeated game in the sense that for any plan of choices $(a_i^t, a_i^{t+1}, \ldots a_i^{t+k}) \in arg\max V(\mu_i^t(H), (a_i^\tau))$ the first element of the plan $(a_i^t \in \mathcal{BR}_i(\cdot))$.

The first property we establish is that all pure absorbing profiles are *individually rational* in the sense that they guarantee each player at least the (pure strategy) minmax payoff.

Lemma 1. *All pure absorbing profiles are individually rational.*

Proof. Consider a pure absorbing action profile $(a_i^*, a_{-i}^*)_{\tau=t,\ldots t+(T-\lambda).}$ where the same actions are chosen at all $t, \ldots t+(T-\lambda)$ by both players. If $a_i^* \in BR_i(a_{-i}^*)$, then a_i^* guarantees the minmax payoff $\widehat{\pi}$ to player i, $\forall t, \ldots t+(T-\lambda)$.

If $a_i^* \notin BR_i(a_{-i}^*) \wedge \pi_i(a_i^*, a_{-i}^*) < \widehat{\pi}$ then this must be because player i believes that a deviation at t (to say $a_i{}'$ with $\pi_i(a_i\prime, a_{-i}^*) > \pi_i(a_i^*, a_{-i}^*)$) yields a payoff lower than $\pi(a_i^*, a_{-i}^*) < \widehat{\pi}$ for some $\tau \in [t+1, t+k]$. (Since $(a_i^*, a_{-i}^*)_{\tau=t,\ldots t+(T-\lambda)}$ is a pure absorbing profile the payoffs without deviation are $\pi(a_i^*, a_{-i}^*) < \widehat{\pi}$ for all such τ. Hence, if this were not the case then i would have incentives to deviate to $a_i{}'$ at t and ensure herself (at least) the minmax payoff at t.) τ has to be within the same T-period interaction and within i's foresight (k). Denote her belief at time t about $-i$'s choices at τ by $\mu_i^t(a_{-i}^\tau | H^{\tau(t)})$. Now if she believes at t that at τ she will choose an action $a_i^\tau \in BR_i(\mu_i^t(a_{-i}^\tau | H^{\tau(t)}))$, then her (instantaneous) payoff at τ will not be below $\widehat{\pi}$. Hence the deviating profile $(a_i'^t, \ldots, a_i'^{t+k})$ must be such that she plans *not* to choose an (instantaneous) best response at τ. But she

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

will find it optimal at t not to choose a (myopic) best response (or any other action guaranteeing her $\widehat{\pi}$) at τ only if there is a $\tau' \in [\tau+1, t+k]$ where she expects to obtain a lower payoff than $\widehat{\pi}$ in case of a deviation etc. At t, though, she certainly *expects* to choose a (myopic) best response at $t+k$, because of limited foresight. Since she will *expect* to obtain at least $\widehat{\pi}$ at $t+k$, and hence at all τ'', τ', τ etc, it cannot be that $\pi_i(a_i^*, a_{-i}^*) < \widehat{\pi}$.

Let us now focus on periods t' where a pure absorbing state does not require (according to definition 4) that a_i^* is chosen. Assume first that $t' \in \{[T], \ldots, [T]+\lambda\}$. Then the exact same reasoning as above guarantees that payoffs must lie above $\widehat{\pi}$ for all such t'. Assume next that $t' \in \{[T]-\lambda+1, \ldots, [T]\}$. If $a_i^{t'} = a_i^*$, then player i can guarantee herself the minmax payoff by the previous arguments. Now assume that $a_i^{t'} \neq a_i^*$ for some t'at an absorbing state. Take the first such t'. At t' the history (of length h) coincides with that of $t'-1$ (because of A1 and since $\lambda < k+1$ by assumption) and hence in an absorbing state beliefs do as well. But then the only reason why at t' a different action may be chosen is that the horizon of play is shorter than before. But if this is the case it must also be the case that (i) $a_i^* \notin BR_i(a_{-i}^*) \wedge \pi(a_i^*, a_{-i}^*) > \widehat{\pi}$ and (ii) $a_i'^t \in BR_i(a_{-i}^*)$. Hence average payoffs above the minmax level can be guaranteed. □

Lemma 2. *Assume $(h,k) > (0,1)$. For any game there exists a real number $\eta(h,k) > 0$ such that action profiles which are not Nash are pure absorbing if and only if they are locally efficient and $\rho^{-1}\lceil m/T\rceil \leq \eta(h,k)$.*

Proof. First we show **sufficiency**. Denote by $\vec{a}^* = (a_i^*, a_{-i}^*)$ a locally efficient action profile and consider a state where T-period interactions have the following structure: $(\underbrace{\vec{a}^*, \ldots, \vec{a}^*}_{T-\lambda \text{ periods}}, \vec{a}\prime, \ldots)$ with $\lambda \in \{1, \ldots, k-1\}$. (If there is no such state that is absorbing, then there will also not be a state of the form $(\ldots, \underbrace{\vec{a}^*, \ldots, \vec{a}^*}_{T-\lambda \text{ periods}}, \ldots)$ that is absorbing since beliefs conditional on history H^0 can never be ruled out to coincide with beliefs after the 'pure' history $(\vec{a}^*, \ldots, \vec{a}^*)$.)

We have to find beliefs that sustain this profile and are consistent with choices made under decision rule (1). We know that $\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*)) \geq 1 - \rho^{-1}\lceil m/T\rceil$ and $\mu(a'_{-i}|(\vec{a}^*, \ldots, \vec{a}^*)) \leq \rho^{-1}\lceil m/T\rceil$, $\forall a'_{-i} \neq a_{-i}^*$ since memory of size m permits to draw a'_{-i} at most $\lceil m/T\rceil$ times in a sample of size ρ. (In states which induce pure absorbing profiles such as above there is only one instance in each T-period interaction where a history $(\vec{a}^*, \ldots, \vec{a}^*)$ of any length is followed by a profile $\vec{a}\prime \neq \vec{a}^*$. At most $\lceil m/T\rceil$ such instances are remembered.) Now a sufficient condition for the profile to be pure absorbing is that $\mathcal{BR}_i^t[\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*))] = a_i^*$, $\forall t \leq T-\lambda$, $\forall i$ whenever $\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*)) \geq 1 - \rho^{-1}\lceil m/T\rceil$. Whenever $\mu(a_{-i}|H)$ is s.t. $\mathcal{BR}_i^t[\mu(a_{-i}|H)] \in A\prime$, $\forall t$ and for every history H of the form $(\vec{a}^*, \ldots, \vec{a}\prime)$, $(\vec{a}^*, \ldots, \vec{a}\prime, \vec{a}'')$, …, $(\vec{a}^*, \ldots, \underbrace{\vec{a}\prime, \vec{a}'', \ldots}_{\lambda \text{ periods}})$, it is possible to find $\eta(h,k)$ small enough such that $\forall \rho^{-1}\lceil m/T\rceil \leq \eta(h,k) : \mathcal{BR}_i^t[\mu(a_{-i}|H)] = a_i^*$, $\forall t \leq T-\lambda$.

The reason is the following: because of condition (2) of the definition of local efficiency, play will remain within A' in all periods $t \in T-\lambda, \ldots T$, i.e. for all histories of the form $(\vec{a}^*, \ldots, \vec{a}\prime)$, $(\vec{a}^*, \ldots, \vec{a}\prime, \vec{a}'')$, …, $(\vec{a}^*, \ldots, \underbrace{\vec{a}\prime, \vec{a}'', \ldots}_{\lambda \text{ periods}})$: $\mathcal{BR}_i^t[\mu(a_{-i}|H)] \in A\prime$.

We have already seen that $\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*)) \geq 1 - \eta(h,k)$ is possible. Now (because of conditions (2) and (3)) there exists an action $\widehat{a}_{-i} \in A_{-i}\prime$ for both i such that $\pi_i(BR(\widehat{a}_{-i}), \widehat{a}_{-i}) < \pi_i(\vec{a}^*)$. Since $\widehat{a}_{-i} \in A_{-i}\prime$ and $\vec{a}^*$ is not a Nash equilibrium, this action $\widehat{a}_{-i} \in A_{-i}\prime$ will be reached via best responses and hence be observed after a deviation history. But this means that there exist beliefs sustaining profile $(\underbrace{\vec{a}^*, \ldots, \vec{a}^*}_{T-\lambda \text{periods}}, \vec{a}\prime, \ldots)$ with $\lambda \in \{1, \ldots, k-1\}$.

Next we show **necessity**. (i) First assume that $\vec{a}^*$ is locally efficient but that the condition $\rho^{-1}\lceil m/T\rceil \leq \eta(h,k)$ is not satisfied. Note that then (if $\rho^{-1}\lceil m/T\rceil > \eta(h,k)$) there is positive probability (for either i) that beliefs are drawn such that $\mathcal{BR}_i[\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*))] \neq a_i^*$. If this is the case then at some $\widehat{t}$ agent i will not choose a_i^* (or conversely $-i$ will not choose a_{-i}^*) and $\forall t > \widehat{t}$ the memory conditional on history $(\vec{a}^*, \ldots, \vec{a}^*)$ will contain at most as many elements a_{-i}^* at t than at $\widehat{t}$. But then it is possible to construct a path away from the candidate absorbing profile $\vec{a}^*$ by repeatedly drawing beliefs such that $\mathcal{BR}_i[\mu(a_{-i}^*|(\vec{a}^*, \ldots, \vec{a}^*))] \neq a_i^*$.

Now we show that Non-Nash profiles have to be locally efficient starting with part (2) of the definition of local efficiency (ii). Assume first that (2) is violated for A'. Note then that as $\vec{a}^*$ is not a Nash equilibrium, some player i must have a best response $BR_i(a_{-i}^*) = a_i\prime$, which will be chosen in a T-period interaction for some $t \in \{T-\lambda, \ldots, T\}$ after a history $(\vec{a}^*, \ldots \vec{a}^*)$. Note that any set A' with property (2) has to contain $a_i\prime$ by definition.

Now if $A\prime = \{(a_i^*, a_{-i}^*), (a_i\prime, a_{-i}^*), (a_i^*, a_{-i}\prime), (a_i\prime, a_{-i}\prime), \ldots\}$ does not satisfy (2), then there is a strictly positive probability that at some point t player i will hold a belief $\mu_i^{\cdot} \in \Delta A_i\prime$ such that $\mathcal{BR}_i^t(\mu_i^{\cdot}) = a''_i \notin \Delta A\prime$. (Note that this belief can be sampled even if $a_i\prime$ is played only in the last period of each T-period interaction, since it still counts as a reaction to the history at $[T]-1$: $H^{[T]-1} = (a_i^*, a_{-i}^*)$.)

Furthermore either the set $A'' = (A_i\prime \cup (a''_i)) \times A_{-i}\prime$ does not satisfy (2) or $\vec{a}^*$ is not efficient in A'' by assumption. We show why efficiency is necessary in step (iii). Assume hence the former and denote by $\Delta_\rho(M)$ the distributions on M which respect

the sampling procedure ρ.[20] Then since there exist $\mu\prime, \mu\prime\prime \in \Delta_\rho(M_i(\vec{a}^*, \ldots, \vec{a}^*))$ such that $a_i{}' \in BR_i(\mu')$ and $a''_i \in BR_i(\mu\prime\prime)$ it is possible that beliefs are repeatedly drawn from $M_i(\vec{a}^*, \ldots, \vec{a}^*)$ such that another action $a_i{}'''$ is played etc. Repeating this argument it can be seen that paths can be constructed which lead away from the absorbing profile $\vec{a}^*$.

(iii) The fact that $\vec{a}^*$ has to be pareto efficient in A' follows from the following observation. Assume $A' = \{(a_i^*, a^*_{-i}), (a_i^*, a'_{-i}), (a'_i, a^*_{-i}), (a'_i, a'_{-i})\}$, where $a'_i \in BR_i(a^*_{-i})$. We will show that $\vec{a}^*$ has to be pareto efficient in A'. If it fails to be pareto efficient in A', it will also fail to be pareto efficient in any $A'' \supset A'$. Now since the profile $\vec{a}^*$ is not a Nash equilibrium, there must exist at least one player i such that $a_i^* \notin BR_i(a^*_{-i})$. Thus (a_i^*, a^*_{-i}) can only be optimal for player i if she believes that deviating at t will reduce her payoff in some periods $\tau \in \{t+1, \ldots, t+k\}$. But if $\vec{a}^*$ is not pareto efficient then there must be $a'_i, a'_{-i} \in A'$ such that either (a'_i, a^*_{-i}) or (a'_i, a'_{-i}) must yield a higher payoff to both players for $(a'_i, a'_{-i}) \neq (a_i^*, a^*_{-i})$. (If this is not true for player i it must be true for player $-i$.) But since $a'_i \in BR_i(a^*_{-i})$, this means that (by Condition (1) of the definition of local efficiency) that $\pi_{-i}(a'_i, a^*_{-i}) < \pi_{-i}(a_i^*, a^*_{-i})$. Hence (since (a_i^*, a^*_{-i}) should fail to be pareto efficient) we will have $\pi_i(a'_i, a'_{-i}) \geq \pi_i(a_i^*, a^*_{-i}) \forall i$. But if this is the case the best response to any belief with support on $A_{-i}{}'$ will be $a_i{}'$ irrespective of k. Hence there are no beliefs supporting $\vec{a}^*$ as an absorbing profile.

(iv) Finally if part (1) of the definition of local efficiency is not satisfied, then there is positive probability to diverge from $\vec{a}^*$ simply because there is positive probability that players repeatedly choose a different element from $\mathcal{BR}_i^t[\mu(a_{-i}|(\vec{a}^*, \ldots, \vec{a}^*))]$. If part (3) is not satisfied then irrespective of the belief about $-i$'s choice after deviating from $\vec{a}^*$ player i has an instantaneous best response guaranteeing (weakly) higher payoffs irrespective of the future path and hence has incentives to deviate. □

Proof of Proposition 1

Proof. Part (ii) follows directly from Lemmas 1 and 2. For part (i) the proof is as follows. Consider any state where the NE $\vec{a}^*$ is played at each t. We will first show that if C1 is satisfied such a state is absorbing. It is sufficient that beliefs satisfy $\mu(a^*|(\vec{a}^*, \ldots, \vec{a}^*)) = 1$ and that $\mu(a_{-i}|(\vec{a}^*, \ldots, (a'_i, a^*_{-i}))$ is such that $\sum_{\tau=t}^{t+k-1} \sum_{a_{-i} \in A} \mu^{i\tau}(a_{-i}|H^{\tau-1(t)}) \pi^i(a_i, a_{-i}) - k\pi(\vec{a}^*) < 0$, holds whenever C1 is satisfied. Finally if C1 is not satisfied, i.e. if there exists a'_i such that $\pi^i(a'_i, a_{-i}) \geq \pi^i(\vec{a}^*), \forall\, a_{-i} \in A_{-i}$, then there is no belief for which player i would strictly prefer to choose a_i^* rather than a'_i. □

Proof of Proposition 2

Proof. We will show that there exists a number $K \in \mathbb{N}$ and a probability $p > 0$ such that from any $s \in S$ the probability is at least p to converge within K periods to a pure absorbing set. K and p are time independent and state independent. Hence the probability of not reaching a pure absorbing set after at least rK periods is at most $(1-p)^r$ which tends to zero as $r \to \infty$.

(i) Let $s^t = (M^t, H^t)$ be the state in period $t \geq m$. Denote by $\vec{a}^*$ the profile chosen at t. If $H^{t+1} = H^t = (\vec{a}^*, \ldots, \vec{a}^*)$ then we can go to step (ii) of the proof (setting $t = \tau''$, which will be defined in step (ii)). Assume $H^{t+1} \neq H^t$. Then, since the set of all possible histories $\mathcal{H}$ is finite, $\exists \tau' > t$ such that $H^{\tau'} = H^\tau$ for some $\tau \in [t, \tau' - 1]$. But then there is positive probability that $H^{\tau'+1} = H^{\tau+1}$ etc., i.e., there is positive probability to return to history H^τ any finite number of times. At history H^τ, there is positive probability, that each agent i samples the last ρ plays in her memory associated with that history $M_i(H^\tau)$. This is always possible, since each element $M_i(H)$ of an agent's memory contains m instances where this history occurred. Denote this sample by ξ. There is also positive probability that the next ρ times that the history is H^τ the agent samples ξ again and chooses the same best response.

(ii) Order the histories according to τ as follows: $H^\tau, H^{\tau+1}, \ldots, H^{\tau'-1}$. Now assume there exists $H^{\tau''} \in [H^\tau, H^{\tau'-1}]$ where $H^{\tau''} =: ((\vec{a}^*, \ldots, \vec{a}^*))$ is part of an absorbing set. Then there is positive probability to sample only the last ρ periods for the next $m - \rho$ periods thereby creating a homogeneous memory $M(H^{\tau''}) = (a^*_{-i}, \ldots, a^*_{-i})$. This is possible whenever $m \geq \rho$, which is true by definition. Since $a_i^* \in \mathcal{BR}(a^*_{-i})$ an absorbing set has been reached.

(iii) Assume now instead that there does not exist $H^{\tau''} \in [H^\tau, H^{\tau'-1}]$ with this property. Now for any $\tau'' \in [\tau, \tau' - 1]$ there is positive probability that each agent samples the last ρ periods where the history was $H^{\tau''}$, i.e., takes a homogeneous sample $(a, \ldots a)$. The best response to $(a, \ldots a)$ for each agent lies on a directed path leading to an absorbing set since the game is acyclic. Again now $\exists \tau''' > \tau''$ such that $H^{\tau'''} = H^{\tau^{iv}}$ for some $\tau^{iv} \in [\tau'', \tau''' - 1]$, since the set of all histories is finite. But then again there is positive probability that all agents take the same sample and choose the same best response to this sample in the next ρ periods $\forall H^{\tau^{iv}} \ldots H^{\tau''\prime - 1}$. If there is a history in $H^{\tau^{iv}}, \ldots, H^{\tau'''-1}$ that is part of an absorbing set, then jump to (ii). Else repeat step (iii). Note next that since the game is acyclic a directed path from any $(a, \ldots a)$ to a history $(\vec{a}^*, \ldots, \vec{a}^*)$ which is part of a pure absorbing set exists. Using the algorithm above, there is thus a positive probability p_s to reach any history on that path and eventually a history which is part of an absorbing set. This is possible whenever $m \geq \rho$, which is true by definition.

To sum up, we have shown that from any state s there is positive probability p_s to converge to a pure absorbing set. By setting $p = \min_{s \in S} p_s > 0$ it follows that from any initial state the process converges with at least probability p to an absorbing set in K periods. □

[20] For example if $M = (A, A, B)$ and $\rho = 2$, the degenerate distribution placing probability one on B does not respect the sampling procedure, while distributions placing probability (1/2) on both A and B or probability 1 on A do.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Proof of absorbing sets Prisoner's Dilemma:

Proof. That the set X_D is absorbing follows directly from Proposition 1. The proof that X_C induces pure absorbing profiles (under the conditions mentioned) follows from Lemma 2. It remains to show that the upper bound on $\rho^{-1}\lceil m/T\rceil$ is given by $((\alpha-\gamma)/\alpha)$. The most restrictive conditions (for the efficient profile to be absorbing) are encountered in the case $(k, h)=(2, 1)$ where $\mu(C|(D, C))=0$. In this case the condition is that both players have to find it advantageous to choose C after a history of $\vec{a}_1$, i.e. that

$$V(\mu(\vec{a}_1), C) > V(\mu(\vec{a}_1), D) \Leftrightarrow \mu(C|\vec{a}_1) > \frac{\gamma}{\alpha}.$$

But then since $M(s)$ contains at most $\lceil m/T\rceil$ choices of D and since ρ coordinates from $M(s)$ are randomly drawn to form this belief, the inequality $\rho^{-1}\lceil m/T\rceil < 1-(\gamma/\alpha)=((\alpha-\gamma)/\alpha)$ follows. Also note that there can be no other absorbing states not contained in either X_C or X_D, since every absorbing state involving some cooperation must be in X_C. Condition (ii) of the definition of X_C is implied by the property of non-increasing cooperation (see the proof of Proposition 3 below). If condition (i) fails, then beliefs may be drawn (placing "too high" probability on the opponent choosing D after a cooperative history) which lead to convergenve to X_D. □

Proof of Proposition 3

Proof. Assume that at period t (such that $t-l \neq 0 \bmod T, \forall\, l=1, \ldots, h$) beliefs of agent i are such that she finds it optimal to choose cooperation (C). If $\forall \tau = t+1, \ldots, t+k-1 : \tau \neq 0 \bmod T$, then the maximization problem at $t+1$ is identical to that at t. But then (since we are in a pure absorbing state) the same action has to be chosen at t and $t+1$. If not, then at $t+1$ the agent will have strictly "less foresight" than at t. But then defection (D) will seem relatively better to cooperation (D) at t compared to the situation at t where the agent looks k periods forward. The reason is that choosing defection must always reduce the probability with which the opponent is expected to cooperate in the future. (If this were not the case both agents would defect at all $t+1$.) Hence if the agent cooperates at $t+1$ she will cooperate as well at t (if $t-l \neq 0 \bmod T, \forall\, l=1, \ldots, h$). □

s-trees

For most of the following proofs we will rely on the graph-theoretic techniques developed by Freidlin and Wentzell (1984).[21] They can be summarized as follows. For any state s an s-tree is a directed network on the set of absorbing states Ω, whose root is s and such that there is a unique directed path joining any other $s' \in \Omega$ to s. For each arrow $s' \to s''$ in any given s-tree the "cost" of the arrow is defined as the minimum number of simultaneous trembles (ϵ – perturbations) necessary to reach s'' from s'. The cost of the tree is obtained by adding up the costs of all its arrows and the stochastic potential of a state s is defined as the minimum cost across all s-trees.

Proof of Proposition 4

Proof. (i) Consider first transitions from $X_D \to X_C$. Denote by $\kappa_{C(1)}$ the minimal number of mistakes necessary in order for one pair of players in a T-period interaction to start choosing cooperation in $T-\lambda$ consecutive periods for some $\lambda \in \{0, \ldots k-1\}$. Note that $\kappa_{C(1)} > 1$ will hold for any $s \in X_D$, since otherwise s could not have been absorbing in the first place. (The reason is that if one player can induce the opponent to cooperate by switching once unilaterally, she will have incentives to do so).

Next we will show that 2 trembles ($\kappa_{C(1)}=2$) are sufficient. Assume that in the first period of a T-period interaction characterized by joint cooperation (denote this period by t) player 1 trembles such that $\vec{a}^t=(C, D)$ and that then at $t+1$ player 2 trembles such that $\vec{a}^{t+1}=(D, C)$. Consider choices at $t+2$. Player 1 will choose C if $\mu_1(C|(C, D)) > (\gamma/(\alpha+2(\beta-\gamma))) =: \widehat{\mu}_1$ (where $\mu_1(C|(C, D))$ is player 1's belief that player 2 will cooperate after a history $H_{21}=(C, D)$ where player 2 defected and player 1 cooperated). The sufficient threshold $\widehat{\mu}_1$ is derived as follows. First note that the least favorable case for such a transition is the case with $(h, k)=(1, 2)$. Then we observe that

$$\begin{aligned} V(\mu(\cdot), (C, D)) &= \mu_1(C|(C, D))[\alpha + (\mu(C|\vec{C})\beta + (1-\mu(C|\vec{C}))\gamma \\ &\quad + (1-\mu(C|(C, D))[\mu_1(C|(C, D)\beta + (1-\mu(C|(C, D))\gamma] \quad \text{and}] \qquad (4) \\ V(\mu(\cdot), (D, D)) &= \mu_1(C|(C, D))[\beta + \mu(C|(D, C))\beta + (1-\mu_1(C|(D, C)))\gamma] \\ &\quad + (1-\mu_1(C|(C, D)))[\gamma + \mu(C|\vec{D})\beta + (1-\mu_1(C|\vec{D}))\gamma]. \end{aligned}$$

We want to find conditions on $\mu(C|(C, D))$ such that $V(\mu(\cdot), (C, D)) > V(\mu(\cdot), (D, D))$ *for all* candidate states $s \in X_D$. Clearly $\mu(C|\vec{D})=0$ is determined "on the outcome path". By setting $\mu(C|(D, C))=0$, $\mu(C|\vec{C})$ to either $\{0, 1\}$ and taking the maximum of the two critical values obtained this way we will get the threshold $\widehat{\mu}_1$ from above. ($\mu(C|(D, C))=0$ is the worst case for such a transition. (Remember that we are looking for a sufficient condition.) Now note that since player 2 cooperated at t+1 following the history $H_{21}=(C, D)$ we know that $\mu_1^{t+2}(C|(C, D)) \geq (1/\rho)$. The same is true for player 2 at t+3, i.e. $\mu_2^{t+3}(C|(C, D)) \geq (1/\rho)$. Hence if $(1/\rho) \geq (\gamma/(\alpha+2(\beta-\gamma)))$, then both players will start to cooperate in this T-period interaction.

[21] See also Young (1993, 1998).

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Finally note that after two agents have been "infected" (through $\kappa_{C(1)} = 2$ trembles as described above) the whole population can be infected. Note first that the "infected" players have beliefs $\mu(C|H^0) \geq (1/\rho)$. Furthermore their beliefs $\mu(C|\vec{C}) \geq \min\{((T-\lambda-1)/\rho), 1\}$, since they both cooperated for at least $T-\lambda$ consecutive periods in their previous interaction. Hence they will have incentives to cooperate after the null history. If the "non-infected" player trembles and chooses C after the null history (say at t') then at $t'+1$ we will either observe $\vec{a}^{t'+1} = (C, C)$, in which case the new agent will be infected or we will observe $\vec{a}^{t'+1} = (C, D)$, in which case the "non-infected" agent can be infected as described above. Hence at most one tremble *per player* is needed for this transition.

(ii) Let us then turn to the reverse transitions $X_C \to X_D$. Again we are interested first in the minimal number of mistakes $k_{D(1)}$ needed for a pair of players to start choosing defection at each t. But while above we were looking for a sufficient condition, we are now interested in a necessary condition for this transition to be possible. First assume that two players simultaneously make a mistake and choose (D, D) at some time t. Then it can be shown by comparing the analogous expressions to (4) that a necessary condition for either player to choose D (D) also at $t+1$ is that $2\gamma > \beta$. Secondly assume that player 1 makes two mistakes and chooses D at t and $t+1$.[22] Now we want to identify a sufficient condition for a transition *not* to be possible, so we consider the most favorable case for such a transition which is again $(h, k) = (1, 2)$.

Next we consider both player's decisions at $t+2$. We will show that a necessary condition for player 2 to choose D at $t+2$ is that $\mu(C|(D, C)) > \frac{\gamma}{\beta-\gamma}$. To see this compare

$$\begin{aligned} V(\mu, (C, D)) = {} & \mu(C|(D, C))[\alpha + \mu(C|\vec{C})\beta + (1 - \mu(C|\vec{C}))\gamma] \\ & + (1 - \mu(C|(D, C)))[\mu(C|(D, C))\beta + (1 - \mu(C|(D, C)))\gamma] \text{ and} \\ V(\mu, (D, D)) = {} & \mu(C|(D, C))[\beta + \mu(C|(C, D))\beta + (1 - \mu(C|(C, D)))\gamma] \\ & + (1 - \mu(C|(D, C))[\mu(C|\vec{D})\beta + (1 - \mu(C|\vec{D}))\gamma]. \end{aligned}$$

Then it can be seen that a necessary condition for a transition to be possible from *any* state in X_C is that $\mu_2^{t+2}(C|(D, C)) > (\gamma/(\beta-\gamma))$. Now there is some state in X_C where player 2 has only one observation C in the memory conditional on (D, C). But then since Since ρ periods are drawn from the memory to form this belief we need $((\rho-1)/\rho) > ((\beta-2\gamma)/(\beta-\gamma)$ for a transition *not* to be possible from *any* state in X_C. By analyzing the analogous expressions for player 1 it can be shown that player 1 has no incentives to start choosing D at $t+2$. Hence under condition $((\rho-1)/\rho) < (\gamma)/(\beta-\gamma)$ at least three trembles are needed to "infect" one pair of agents.

But note that for the two infected agents beliefs are still $\mu(C|H^0) \geq ((\rho-1)/\rho)$ and $\mu(C|\vec{C}) \geq ((\rho-1)/\rho)$. But this means that "infected" agents will choose C again after the null history. (If this were not true then s could not have been absorbing in the first place). Hence at least three trembles *per player* are needed to induce this transition (under the conditions above).

(iii) Combining the conditions found in (i) ad (ii) we first note that $((\rho-1)/\rho) > ((\beta-2\gamma)/(\beta-\gamma)) \Rightarrow 2\gamma < \beta$. Furthermore we have that $((\beta-2\gamma)/(\beta-\gamma)) < ((\alpha+2\beta-3\gamma)/(\alpha+2(\beta-\gamma))$. Hence a sufficient condition thus is $((\rho-1)/\rho \geq ((\alpha+2\beta-3\gamma)/(\alpha+2(\beta-\gamma))$, which is the condition from Proposition 3.

(iv) To finish the proof take any state $s \in X_D$ and consider a minimal s-tree. Assume first that there exists a state $s' \in X_C$ such that the transition from s' to s requiring the least amount of trembles is direct (i.e., does not pass through another absorbing state). Under our conditions the transition $s' \to s$ requires more trembles than $s \to s'$. But then we can simply redirect the arrow $s' \to s$ thereby creating an s' tree with smaller stochastic potential. If the shortest transition $s' \to s$ is indirect (passing through other states in X_C) do the following. Take the arrow $s'' \to s$ leading to s and reverse it. Since $s'' \to s$ has a cost of at least two under our conditions we have created an s''-tree with potential $\psi(s'') \leq \psi(s)$. If strict inequality holds the proof is complete. Assume thus $\psi(s'') = \psi(s)$. Then consider the arrow $s''' \to s''$ and reverse it etc Now at some point there must exist a state s^{iv} on the path $s' \to s''$ such that reversing this link saves one "tremble" per player. Else the s-tree could not have been minimal in the first place. Reversing this link will yield an s^{iv} tree with $\psi(s^{iv}) < \psi(s'') \leq \psi(s)$. □

Proof of Proposition 5:

Proof. The proof follows from the proof of Proposition 3. Since now the efficient outcome (C, C) is also a Nash equilibrium of the one-shot game, condition (2) is not needed for the result. □

Proof of Proposition 6

Proof. Assume that $\mu(C|(C, C)) = 5/6$ and $\mu(C|(D, D)) = 0$ (determined on the "outcome" path) and denote "off-path" beliefs $\mu(C|(D, C)) =: x$ and $\mu(C|(C, D)) =: y$. By Proposition 3, if an agent finds it optimal to cooperate in period 6, she will find it optimal to cooperate in period 2, . . ., 5. Also if an agent finds it optimal to defect in period 7, she will find it optimal to do so in periods 8, . . ., 10. We show next that under the conditions of the Proposition all agents will find it optimal to cooperate in period 6 and to defect in period 7. Denote the vectors $(C, D, D, D, D) =: \vec{a}(C)$ and $(D, D, D, D, D) =: \vec{a}(D)$. (Note that only

[22] No other constellation of two trembles can induce the transition. If first player 1 trembles and then player 2, the probability that both players attach to the event that the opponent defects after a history where they themselves defected and the opponent cooperated will increase, making it even more attractive for them to cooperate.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

the first choice is realized. The remaining choices determine the continuation payoff. Since we assume that defection will be optimal from period 7 on we know the continuation path must be all D in both cases.) To show the first claim, it is then sufficient to verify that $V(\mu^{it}(C|\vec{C}), \vec{a}(C)) \simeq 27.72$ (where we have set $y=0$ as worst case) exceeds $V(\mu^{it}(C|\vec{C}), \vec{a}(D)) = 13.3 + (5/6)12\sum_{j=1}^{4} x^j + 16(1-x)(5/6)$. To show the second claim (that defection is optimal in period 7) it is sufficient to establish that $V(\mu^{it}(C|\vec{C}), \vec{a}(C)\prime) \simeq 25.82$ is smaller than $V(\mu^{it}(C|\vec{C}), \vec{a}(D)\prime) = 13.3 + (5/6)12\sum_{j=1}^{3} x^j + 12(1-x)(5/6) \geq 26.13$ where $\vec{a}(C)\prime := (C, C, C, D)$ and $\vec{a}(D)\prime := (D, D, D, D)$. Both inequalities are satisfied whenever $x \in [0, 0.49]$. Whenever $m \leq 13$ beliefs will always lie in the relevant intervals. We still need to show that agents cooperate in period 1, since this case is not covered by Proposition 3. Note that in any state where agents cooperate in period 2, . . ., 6 the memory after history (*D*, *D*) must contain sufficiently many *D* entries to deter defection in periods 2, . . ., 6. But if this is true, then agents will have incentives to cooperate at $t=1$ as well. We have now shown that all absorbing states that involve any cooperation at all are characterized by 6 periods of mutual cooperation followed by 4 periods of mutual defection. But then if $\rho \geq 6$ and $m \leq 13$ we know from Proposition 2 that all stochastically stable states must involve some cooperation. Hence the stochastically stable states must be of the form above. □

Proof of Proposition 7:

Proof. Assume that $\mu(C|(C, C)) = 7/8$, $\mu(C|H^0) = 1$ and $\mu(C|(D, D)) = 0$ and denote off-equilibrium beliefs $\mu(C|(D, C)) =: x$ and $\mu(C|(C, D)) = y$. In analogy to the proof of Proposition 6, we will show that under the conditions of the Proposition all agents will find it optimal to cooperate in period 8 and to defect in period 9. For this we verify that $V(\mu^{it}(C|\vec{C}), (C, D, D)) \simeq 22.13 + 7x$ exceeds $V(\mu^{it}(C|\vec{C}), (D, D, D)) \simeq 19 + 7x + (21/2)x^2$ which requires $x < 0.54$ and that $V(\mu^{it}(C|\vec{C}), (C, D)) \simeq (65/4) + y$ is smaller than $V(\mu^{it}(C|\vec{C}), (D, D)) \simeq 16 + 7x$. Note that y will be at least $(1/2)$ since a tit-for-tat player will always respond with cooperation to (C, D). But then $\forall x > 0.1$ the latter inequality is satisfied. But then whenever $m \leq 19$ beliefs will always lie in the relevant intervals. □

Proof of Proposition 8:

Proof. First note that absorbing states with full defection exist for all σ. Obviously in these states all agents will have the same average payoffs. Note also that myopic types will always choose defection since it is a dominant strategy in the one-shot game. Hence whenever $\sigma > ((3\alpha - \beta - 3\gamma)/(3\alpha - \beta))$ or whenever $3\alpha - \beta < 0$, all absorbing states will be characterized by full defection. If $\sigma \leq ((3\alpha - \beta - 3\gamma)/(3\alpha - \beta))$ forward-looking types k_2 will find it always optimal to cooperate after the null history (given all beliefs $\mu(C|H^0, k_2) = 1$; $\mu(C|\vec{C}, k_2) \geq (2/3)$; $\mu(C|H^0, k_1) = \mu(C|\vec{C}, k_1) = 0$). But then given that k_2 types cooperate in the first threeand defect in the fourth period, k_1 types will make higher expected payoffs whenever

$$\Pi^e(k_1) \geq \Pi^e(k_2) \Leftrightarrow$$
$$\sigma\gamma + (1-\sigma)\beta + 3\gamma \geq (1-\sigma)[3\alpha + \gamma] + \sigma 3\gamma \Leftrightarrow$$
$$\sigma \geq \frac{3\alpha - \beta - 2\gamma}{3\alpha - \beta - \gamma}. \quad \square$$

Proof of Proposition 9:

Proof. Note that whenever $\sigma > 0$ there is always positive probability that some k_2 agents are matched with only k_1 agents for at least m periods. Consequently their (unconditional) beliefs will converge to $\mu(C|H^0) = 0$ (or at least will fall below the cooperation threshold) and they will start choosing defection at all initial s. There is then again positive probability that such "infected" agents will be matched amongst each other (thereby continuing to defect) and that the k_1 types will be matched with the remaining k_2 types. Hence from any state there is positive probability to reach a state where all agents defect. □

References

Andreoni, J., 1988. Why free ride? Strategies and learning in public goods experiments. J. Public Econ. 37 (3), 291–304.
Andreoni, J., Miller, J., 1993. Rational cooperation in the finitely repeated Prisoner's dilemma: experimental evidence. Econ. J. 103, 570–585.
Bac, M., 1996. Corruption, supervision and the structure of hierarchies. J. Law Econ. Org. 12, 277–298.
Basu, K., Weibull, J., 1991. Strategy subsets closed under rational behavior. Econ. Lett. 36, 141–146.
Binmore, K., Mc Carthy, J., Ponti, G., Samuelson, L., Shaked, A., 2001. A backward induction experiment. J. Econ. Theory 104 (1), 48–88.
Blume, L., 2004. Evolutionary Equilibrium with Forward-looking Players, Working Paper. Santa Fe Institute.
Burlando, R., Hey, J., 1997. Do Anglo-Saxons free-ride more? J. Public Econ. 64, 41–60.
Ehrblatt, W.Z., Hyndman, K., Oezbay, E., Schotter, A., 2010. Convergence: an experimental study of teaching and learning in repeated games. J. Eur. Econ. Assoc. 10 (3), 573–604.
Freidlin, M.I., Wentzell, A.D., 1984. Random Perturbations of Dynamical Systems. Springer-Verlag, New York.
Fudenberg, D., Levine, D., 1989. Reputation and equilibrium selection in games with a patient player. Econometrica 57, 759–778.
Fudenberg, D., Levine, D., 1993. Self-confirming equilibrium. Econometrica 61 (3), 523–545.
Fudenberg, D., Levine, D., 1998. The Theory of Learning in Games. MIT-Press, Cambridge.
Fudenberg, D., Kreps, D.M., 1995. Learning in extensive form games. I. Self confirming equilibria. Games Econ. Behav. 8, 20–55.
Fujiwara-Greve, T., Krabbe-Nielsen, C., 1999. Learning to Coordinate by Forward Looking Players. Riv. Int. Sci. Soc. CXIII (3), 413–437.
Ghosh, S., Ray, D., 1996. Cooperation in community interaction without information flows. Rev. Econ. Stud. 63, 491–519.
Gueth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Org. 3 (4), 367–388.
Heller, Y., 2014. Three steps ahead. Theor. Econ., forthcoming.
Jehiel, P., 1995. Limited horizon forecast in repeated alternate games. J. Econ. Theory 67, 497–519.
Jehiel, P., 1998. Learning to play limited forecast equilibria. Games Econ. Behav. 22, 274–298.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001

Jehiel, P., 2001. Limited foresight may force cooperation. Rev. Econ. Stud. 68, 369–391.
Karlin, S., Taylor, H.M., 1975. A First Course in Stochastic Processes. Academic Press, San Diego.
Kandori, M., Mailath, G., Rob, S., 1993. Learning, mutation, and long run equilibria in games. Econometrica 61, 29–56.
Karandikar, R., Mookherjee, D., Ray, D., Vega-Redondo, F., 1998. Evolving aspirations and cooperation. J. Econ. Theory 80, 292–331.
Kreps, D., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated Prisoner's dilemma. J. Econ. Theory 27 (2), 245–252.
Levine, D., Pesendorfer, W., 2007. The evolution of cooperation through imitation. Games Econ. Behav. 58, 293–315.
Mengel, F., 2007. The evolution of function-valued traits for conditional cooperation. J. Theor. Biol. 245, 564–575.
Mengel, F., 2008. Matching structure and the cultural transmission of social norms. J. Econ. Behav. Org. 67, 608–623.
Myerson, R.B., Pollock, G.B., Swinkels, J.M., 1991. Viscous population equilibria. Games Econ. Behav. 3, 101–109.
Selten, R., Stoecker, 1986. End behaviour in sequences of finite Prisoner's dilemma supergames: a learning theory approach. J. Econ. Behav. Org. 7, 47–70.
Selten, R., 1991. Anticipatory learning in two-person games. In: Selten, R. (Ed.), Game Equilibrium Models I. Springer-Verlag, Berlin, pp. 98–154.
Ule, A., 2005. Exclusion and Cooperation in Networks (Ph.D. thesis). Tinbergen Institute.
Watson, J., 1993. A reputation refinement without equilibrium. Econometrica 61, 199–205.
Young, P., 1993. The evolution of conventions. Econometrica 61 (1), 57–84.
Young, P., 1998. Individual Strategy and Social Structure. Princeton University Press, Princeton, New Jersey.

Please cite this article in press as: Mengel, F., Learning by (limited) forward looking players. J. Econ. Behav. Organ. (2014), http://dx.doi.org/10.1016/j.jebo.2014.08.001