# Learning by (limited) forward looking players[*]

Friederike Mengel[†]
Maastricht University

November 2008

**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 relying on their past experience in the same situation (after the same recent history) 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 are individually rational and satisfy an efficiency condition. We then characterize stochastically stable states in $2 \times 2$ 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. Finally, if populations are composed of some myopic and some forward looking agents parameter constellations exist such that either might obtain higher average payoffs.

Keywords: Game Theory, Learning, Forward-Looking Agents.
JEL-Classification: C70, C73.

[*] I wish to thank Jayant Ganguli, Jaromir Kovarik, Ran Spiegler and Elias Tsakas as well as seminar participants in Bielefeld and Muenchen for helpful comments.

[†] Department of Economics (AE1), Maastricht University, PO Box 616, 6200MD Maastricht, Netherlands. F.Mengel@algec.unimaas.nl

# 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 at least act as if they were doing so. In recent decades this model has been criticized for ascribing too much rationality to agents. Experimental work has for example demonstrated that agents often do not 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 are usually ascribed a substantial degree of myopia or "irrationality", assuming e.g. that players learn through reinforcement, imitation or at best choose 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 display a (limited) degree of forward-looking. In our model, agents in a large, but finite population are randomly matched to interact in finitely repeated games. They form beliefs by relying on their past experience in the same situation (after the same recent history) and then best respond to these beliefs looking $k$ periods ahead. Beliefs are conditional on the recent history of play in the finitely repeated game with their current match. This implies that when choosing an action plan agents take into account how their choice in the current period alters the history at and thus typically the choices of their match in future periods.

In general being forward looking implies that two kinds of effects can be anticipated. On the one hand agents will be aware that their action choices will affect the history of relations and hence the future path of play. On the other hand they might anticipate that their action choices affect future beliefs of their opponents. Our model allows for agents to learn the first kind of effect explicitly. The second effect is only present implicitly. The reasons are that a) forming beliefs about the opponent's beliefs implies a kind of strategic thinking that is absent in our model and b) since the degree of forward looking $k$ is assumed to be rather small such belief changes will be negligible over the horizon of the agent.

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. The model nests the model of adaptive play by Young (1993).

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

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

[3] There is also quite some empirical evidence for this. See e.g. Ehrblatt et al (2008) or Boyd and Richerson (2005) among others.

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 individually rational and 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 $2 \times 2$ 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). Finally we also show that if populations are composed of some myopic and some forward looking agents there are some parameter constellations such that myopic agents obtain higher average payoff and others such that forward-looking agents obtain higher average payoffs in absorbing states. These results suggest that in an evolutionary model polymorphic populations (composed of both myopic and forward-looking agents) or populations composed of only forward looking agents might evolve.

There are other models with limited forward looking agents. Most of them take a strategic perspective. 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. In Jehiel (2001) he 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. Also in Jehiel (1995) he shows that the length of memory does not matter for equilibrium outcomes, whereas in our model it can be crucial, as we will show below.[4]

Blume (2004) has proposed 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. Other works include Fujiwara-Greve and Krabbe-Nielsen (2006) who study coordination problems, Selten (1991) or Ule (2005) who models forward looking players in a network.[5] There is also some conceptual relation to the literature

[4] In Jehiel (1998) he proposes a learning justification for limited horizon equilibrium.

[5] The idea of sophistication is also present in e.g. Stahl (1993), who analyzes agents that

on long-run and short-run players (see e.g. Kreps and Wilson, 1982).[6] Results more closely related to this literature have been tested in experiments that investigate strategic sophistication and the existence of some "teachers" among adaptive players. Examples are Ehrblatt et al. (2008), Terracol and Vaksman (2008) or Camerer, Ho and Chong (2002).

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

**The Game:** There is a large, but finite population of (infinitely lived) individuals that is partitioned into two non-empty classes $C_1$ and $C_2$. At each $t = 0, T, 2T, 3T...$, $2n$ players are randomly drawn ($n$ from each class) and matched in pairs to interact repeatedly in a (normal form) two-player game. The members of $C_{1(2)}$ are candidates to play role 1 (2) in the game. Each interaction consists of $T$ repetitions of the stage game.

Each player has a finite set of actions $A_i$ to choose from. The payoff that player $i$ obtains in a given round if she chooses action $a$ and her opponent action $b$ is given by $\pi^i(a, b)$. A history of play for player $i$ of length $h < T$, denoted $H_{ij}(h)$, is a vector that summarizes the past action choices in the last $h$ rounds of the *current interaction* with player $j$. In the first round of each $T-$period interaction we set $H_{ij}(h) = \varnothing$. Denote by $\mathcal{H}(h) \subset (A_i \times A_{-i})^h$ the set of all possible histories of length $h$.

**Limited Foresight and Sophistication:** Players have limited foresight of $k$ periods, meaning that they choose actions in order to maximize their expected utility across the following $k$ rounds.[7] They also have limited "sophistication" of $h$ periods, meaning that they condition their beliefs on histories of length $h$. For most of the paper we will assume that $h, k < T/2$ and that all agents in the population display the same degree of forward-looking $k$ and the same $h$. We will investigate alternative assumptions in Section 4.

**Beliefs:** Agents in our model are adaptive forming beliefs by relying on their past experience. More precisely, at each period in time $t$ players form predictions about their opponent's action choices based on their experience with the population and conditional on the history of play in their current ($T-$period) interaction. Agents have limited memory, meaning that they remember only the last $m$ instances where history $H$ occurred and memorize the action choice of the opponent immediately following such a history.[8]

are $n$ smart according to the levels of rationalizability.

[6] See also Fudenberg and Levine (1989) or Watson (1993) among others.

[7] Whenever there are less than $k$ rounds left to play agents simply take into account all remaining rounds.

[8] Note that this implies that agents can remember "rare" events lying far back in time whereas they might not remember more "common" or "frequent" events even if they are closer in time.

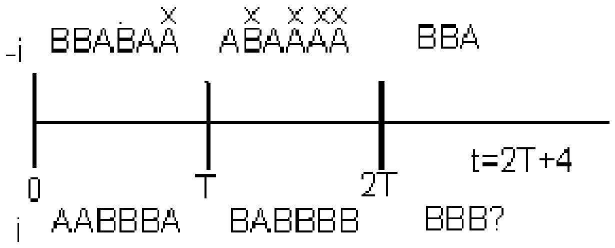

Figure 1: Example Timeline

After observing a history $H$, they then randomly sample $\rho \leq m$ out of the last $m$ periods where the history was $H$.[9] The probability $\mu^{it}(b|H)$ that agent $i$ attaches to her opponent choosing action $b$ conditional on history $H$ then corresponds to the frequency with which $b$ was chosen after history $H$ in the agents sample. Denote by $\mu^{it}(H)$ the beliefs of agent $i$ given history $H$ at time $t$. Note that if $h = 0$ then players just sample $\rho$ out of the last $m$ periods. We introduce such imperfect sampling in order to nest the model of Young (1993) for the myopic case and to be able to establish almost sure convergence.

**Action Choice:** They then choose an action vector $(a^\tau)_{\tau=t,..t+k-1}$in order to maximize their expected payoff over the next $k$ rounds, i.e. in order to maximize[10]

$$V(\mu^{it}(H), (a^\tau)) = \sum_{\tau=t}^{\max\{t+k-1,[T]\}} \sum_{b \in A} \mu^{i\tau}(b|H^{\tau-1}) \pi^i(a^\tau, b), \tag{1}$$

where $[T]$ denotes any multiple of $T$. Expression (1) illustrates how forward looking players take into account the impact their action choice at time $t$ has on their opponent's action choice in the following $k-1$ periods (by altering the history of play). If there are less than $k$ periods left to play agents realize this and sum only over the remaining periods. The following example illustrates the process.

**Example 1** *Consider the following hypothetical situation for player $i$ illustrated in Figure 1. At time $t = 0$ she was matched with some player $-i$ who chose actions $(B, B, A, B, A, A)$ in the $6-$period interaction. She herself chose $(A, A, B, B, B, A)$. At time $T = 6$ she was rematched with another player and at $2T = 12$ she was rematched again now to player $j$. The current round of interaction is $t = 2T + 4$ and player $i$ wants to decide on an action plan.*

[9] If a history occured only $m' \leq m$ times in the past agents sample $\min\{m', \rho\}$ periods from the last $m'$ periods. If a history never occurred in the past agents use a default belief having full support on $A_{-i}$.

[10] We chose not to include an explicit discount factor for notational simplicity. A discount factor could be easily included in the model, but wouldn't affect any of the results qualitatively.

*Assume that $h = 1$ and $m = 5$. Then it is easy to see that the history at time $t$ for player $i$ is $H_{ij}^t = (B, A)$. The memory agent $i$ has conditional on this history is $M_i(t, (B, A)) = \{A, A, A, B, A\}$. These are the last five action choices of her previous opponents following the history $(B, A)$. Note that whereas history refers always to the current interaction, agents' memories (possibly) contain experiences from many past interactions.*

**State:** Denote by $M_i(t, H)$ the collection of the opponent's action choices in the last $m$ interactions of player $i$ in which the history was $H$, let $M_i(t) = (M(t, H))_{H \in \mathcal{H}}$ and denote by $M(t) = (M_i(t))_{i \in C_1 \cup C_2}$ the collection of memories across all players. The state at time $t$ is then given by the tuple

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

i.e. by the collective memory of all agents together with the current history in all agents' interactions.[11] 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 $S = (A^m \times \mathcal{H}) \times \mathcal{H}$. Furthermore denote by $H_i^s$ the history of player $i$ associated with state $s$ and by $M(H_i^s)$ the memory associated with that history and let $H^s$ and $M(H^s)$ be the collection of these histories and memories across all players. Call $\widehat{s}$ a successor of $s \in S$ if $\widehat{s}$ is obtained from $s$ by deleting the leftmost element from some $M(H_i^s)$, adding a new element to the right of $M(H_i^s)$ and by updating $H_i^s$ consistently. Denote this last element added to $M(H_i^s)$ by $r_i(\widehat{s})$.

**Techniques:** The learning process can 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 $\forall s, s' \in S$ :

$$P(s, s') > 0 \Leftrightarrow \begin{cases} s' \text{ is a successor of } s \text{ and} \\ r_i(s') \in \arg\max \ V(\mu(H_{-i}^s), k), \forall i \end{cases}$$

**Definition (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 (1) with probability $1 - \varepsilon$ and with probability $\varepsilon$ choose an action randomly (with uniform probability) from $A_i$.

[11] If an agent is currently not interacting with another agents simply set $H(t) = \varnothing$. Note that in some sense $H(t)$ is redundant in the description of the state. We decided to include it explicitly for clarity.

The perturbed Markov process $P^\varepsilon(s, s')$ is ergodic, i.e. it has a unique stationary distribution denoted $\mu^\varepsilon$. This distribution summarizes both the long-run behavior of the process and the time-average of the sample path independently of the initial conditions.[12] The limit invariant distribution $\mu^* = \lim_{\varepsilon \to 0} \mu^\varepsilon$ exists and its support $\{s \in S | \lim_{\varepsilon \to 0} \mu^\varepsilon(s) > 0\}$ 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 $\mu^*$ in the long run. The states in the support of $\mu^*$ are called stochastically stable states.

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

We will characterize stochastically stable states in Section 3.2.

# 3 Results

First let us comment on the standard case where $(h, k) = (0, 1)$ and $T = 1$, i.e. where each agent plays one round of a normal form game with his opponent before being rematched and where all agents have foresight $k = 1$, i.e. are myopic and take into account only their payoffs in the current period when deciding on an action. In this case the process corresponds to the process of adaptive play described by Young (1993). Define the best reply graph of a game $\Gamma$ as follows: each vertex is a tuple of action choices, and for every two vertices $a$ and $b$ there is a directed edge $a \to b$ if and only if $a \neq b$ and there exists exactly one agent $i$ such that $b^i$ is a best reply to $a^{-i}$.

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

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

**Theorem (Young, 1993)** *If $\Gamma$ 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. But note that, unlike in Young's (1993) model, in our model generally $T > 1$. Of course Young's theorem still applies (even if $T > 1$), as long as agents are myopic and choose their actions via the process outlined above. One might ask, though, whether it makes sense to think of myopic learners if $T > 1$. We do not make such a claim. Young's (1993) result, though, can serve as a benchmark for the case where agents are not myopic.[13]

[12] See for example the classical textbook by Karlin and Taylor (1975).

[13] Note that if agents are myopic ($k = 1$), but sophisticated ($h > 0$) then results will not be very different from the $h = 0$ case, since eventually all history-dependent memories will converge.

## 3.1 Absorbing States

We will start by analyzing absorbing states. In our discussion we will focus exclusively on what we call "pure absorbing states", i.e. states in which one action profile is played almost all of the time if $T \to \infty$. This should not be read to imply that we will assume that $T$ is large. Rather a pure absorbing state can be one in which several different action profiles are chosen. What we require though is that the fraction of times in which the pure absorbing profile $\overrightarrow{a}^*$ is chosen is strictly increasing in $T$ while that of all other profiles is not. (For example if $\overrightarrow{a}^*$ is chosen in rounds $t = 1, ...T - 2$ and $a' \neq \overrightarrow{a}^*$ is chosen at $T - 1$ and $T$ we will still call $\overrightarrow{a}^*$ a *pure* absorbing profile, whenever this is the case irrespective of the particular value of $T$).

**Definition** We say a profile $\overrightarrow{a}^*$ is *(pure) absorbing* if there exists an absorbing set $X \subset S$ s.t. $\overrightarrow{a}^*$ is induced in $X$ almost all the time (as $T \to \infty$).

If a set $X \subset S$ induces a pure absorbing profile we will also refer to this set as being pure absorbing. We now proceed to characterizing such pure absorbing profiles. The first property we establish is that all absorbing sets are *individually rational* in the sense that they guarantee each player at least the (pure strategy) minmax payoff at each $t$.

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

**Proof.** Appendix. ■

This property is not very surprising given that agents in our model choose $k-$period best responses at each period. The underlying logic is then essentially the same as in the repeated games literature. Players will only be willing to choose an action which is not a best-response in the one shot game, because they believe that doing otherwise will be "punished" by the other player's reaction to such a history in future periods. The worst such threat is the mutual minmax profile. Since agents look only $k$ periods ahead, every sequence of $k$ periods has to satisfy this property.

The second property we would like to establish is an efficiency condition. In fact it is not hard to see that in $2 \times 2$ games absorbing profiles (unless they are Nash equilibria) have to be Pareto efficient (in the one-shot game).[14] The intuition is again quite simple. Players will only refrain from choosing a myopic best response if they believe that the induced history of play will induce the opponent to choose an action which will yield a lower payoff than the absorbing path. But this is only possible if the non-Nash action is Pareto efficient.

**Lemma 2** *In any $2 \times 2$ game: If an action profile which is not a Nash equilibrium is absorbing then it must be Pareto efficient.*

**Proof.** Appendix. ■

[14] Whenever we talk of (Non-) Nash actions, pareto efficient outcomes or curb sets (below), we always refer to the one shot game.

Note that the requirement of pareto-efficiency or Nash is not particularly strong in $2 \times 2$ games, ruling out only very unintuitive outcomes. In games with a unique Nash equilibrium for example (such as matching pennies) all outcomes will be Pareto efficient. The following example illustrates why in larger games this assertion need not be true.

|   | L | C | R |
|---|---|---|---|
| T | $2,2$ | $0,4$ | $4,0$ |
| M | $4,0$ | $1,1$ | $0,0$ |
| B | $0,4$ | $0,0$ | $3,3$ |

(2)

Intuitively in this game it seems that the action profile $(T, L)$ can be sustained in a pure absorbing state in spite of the fact that it is not pareto efficient. This could be the case for example if players believe that switching to $M$ $(C)$ will induce the opponent to respond with $C$ $(M)$ while switching to $B$ $(R)$ will not change the opponent's future action choice. Of course there is a sense in which the profile $(T, L)$ is "locally efficient" in a sense that we will make precise below. For this we need the following definition.

**Definition (Basu and Weibull 1991)** A subset of actions $A' \subseteq (A_1 \times A_2)$ is called curb whenever it is closed under best replies to all distributions $\mu \in \Delta A'_{-i}, \forall i = 1, 2$.

The definition of a curb set (short for "closed under rational behavior") was 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. Obviously any game is a curb-set itself, strict Nash equilibria are (minimal) curb-sets but also the set $A' = (T, M) \times (L, C)$ in the example above is curb.

What we will require for a Non-Nash action profile to be pure absorbing also in larger games is roughly that it is efficient in a curb set of the one shot game. This is quite intuitive. Assume that an efficient (non Nash) profile $\overrightarrow{a}^*$ was played in all rounds of an interaction $1, ..T - 1$. Since in round $T$ players will choose myopic best responses (since $T$ is the last round of the interaction and this is commonly known among the agents), they will choose a different action in this round. But this means that (given the "efficient history") beliefs are never guaranteed to lie on the boundary of the simplex. In fact if $\rho$ is small enough compared to $m$ (i.e. if samples are not very informative), then conditional on the pure history containing only $\overrightarrow{a}^*$ all beliefs putting positive probability on either $a^*_{-i}$ or the action chosen in the last period $T$ may be drawn (as long as they are multiples of $\frac{1}{\rho}$). But then if the set $A'$ , containing both $a^*$ and the best response chosen in $T$, is not curb divergence may occur with positive probability. The necessary condition is then that $\overrightarrow{a}^*$ is efficient in this set $A'$ for the same reasons underlying Lemma 2.

**Lemma 3** *If an action profile $\overrightarrow{a}^*$ which is not a Nash equilibrium is pure absorbing for any value of $\rho/m$, then there must exist a curb set in which $\overrightarrow{a}^*$ is pareto efficient.*

**Proof.** Appendix. ■

Lemma 3 shows that efficiency in a curb set is a necessary condition for an action profile to be induced in a pure absorbing state. The reason is that the smaller and smaller time horizon towards the end of a $T-$period interaction leads to different choices (and hence heterogenous memories). Imperfect sampling from such heterogenous memories then induces strong variation in beliefs, what makes efficiency in a curb set a necessary condition.

The discussion preceding Lemma 3 though suggests that the same condition may not be sufficient. The reason is that if $\rho/m$ is such that beliefs placing too little probability on the efficient action may be drawn, then choosing the efficient action almost all of the time will also not be absorbing. Since then, there is positive probability for "bad" beliefs to be drawn repeatedly, there is positive probability that agents will converge to a Nash equilibrium instead.

Consequently such efficient profiles cannot be absorbing for any $\rho/m$ (even though they are efficient in a curb-set). To derive a sufficient condition some restrictions on $\rho/m$ will be needed. But then again - given that we will impose such restrictions on $\rho/m$ we can relax our condition on curb sets. In fact to derive necessary and sufficient conditions we will use the following definition of "local efficiency".

**Definition** We call an action tuple $a^*$ *locally efficient* if

1) all unilateral deviations from $a^*$ strictly hurt at least one player and

2) $a^*$is pareto efficient within a $\underline{\mu}(a^*)-$curb-set $A' \subseteq (A_1 \times A_2)$, i.e. a set closed under best replies to all beliefs placing at least probability $\underline{\mu}$ on $a^*$.

The exact value of $\underline{\mu}(\cdot)$ will of course depend on $\rho, m$ and $T$ as well as the game payoffs. Note also that, since all games are curb sets any profile that is pareto efficient in some game automatically satisfies Condition 2). The first condition 1) ensures that efficient profiles are singletons.

The next Lemma now shows the sufficient condition guaranteeing that non-Nash profiles can be induced in a pure absorbing and the additional qualifications needed.

**Lemma 4** *Assume* $(h,k) \gg (0,1)$. *For any game there exists* $\eta(h,k) \in (0,1)$ *s.t. if* $\rho^{-1}\left\lceil \frac{m}{T} \right\rceil \leq \eta(\cdot)$ *then any action profile which is individually rational and locally efficient is pure absorbing.*

**Proof.** Appendix. ■

The condition on $\rho^{-1}\left\lceil \frac{m}{T} \right\rceil$ ensures that samples remain informative enough, as outlined already above.[15] $\left\lceil \frac{m}{T} \right\rceil$ is a measure of the maximal number of "rare" events contained in an agent's memory. If $\rho$ now is too small compared to this expression then it is possible that such "rare" events are overrepresented in the sample on the basis of which agents form beliefs. This can destabilize the efficient absorbing profile. One role of the size of memory in our model is thus

[15] For any $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer larger than $x$.

to ensure that samples remain "informative". Unlike in Jehiel (1995) memory thus can be crucial in determining absorbing sets of the stochastic process.

If this condition is satisfied on the other hand then local efficiency (together with individual rationality) will also be sufficient, since for long enough time horizons (in periods $1, ..T-k+1$) agents will find it optimal to best respond with the efficient action to all beliefs that can be generated from such "informative" samples.

Note that the result in Lemma 4 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. Lemma 4 implies for example that paths involving cooperation in the Prisoner's Dilemma almost all the time can be absorbing.

We have seen that profiles which are not Nash equilibria can be induced at an absorbing state, provided they are individually rational and locally efficient. Next we want to answer the question whether all Nash equilibria can be induced at an absorbing state. It turns out that this is not the case and that we have to impose an - albeit very weak - condition on the Nash equilibrium. Consider the following condition.

**Definition (C1)** An action profile $\overrightarrow{a}^*$ satisfies C1 if $\forall i$ and $a_i' \neq a_i^* : \exists a_{-i} \in A_{-i}$ s.t. $\pi^i(a_i', a_{-i}) < \pi^i(\overrightarrow{a}^*)$.

Condition 1 is an extremely weak requirement. It only says that - starting from an action profile $\overrightarrow{a}^*$ - there should not exist an action that yields always (weakly) larger payoffs then $\pi^i(\overrightarrow{a}^*)$ irrespective of what the opponent chooses. Obviously strict Nash equilibria satisfy this requirement. 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* $(h, k) \gg (0, 1)$. *A profile is pure absorbing if and only if it is either (i) a Nash equilibrium satisfying C1 or (ii) individually rational and locally efficient satisfying the conditions from Lemma 4.*

**Proof.** Appendix. ■

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. The intuition simply is that if agents experience "bad" actions by their opponent with higher probability after a history of Nash play than after a history of efficient (but possibly non Nash) play and form the corresponding beliefs, then they will have incentives to refrain from choosing (myopic) best responses. More loosely speaking agents will 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 rounds of the repeated interaction.

One might ask whether there are other absorbing sets then the pure absorbing sets characterized above or whether cycles are possible. The following result shows that in acyclic games the process converges with probability one to one of the pure absorbing sets.

**Proposition 2** *Assume the game is acyclic. Then there exists a finite* $\underline{m}(h,k,\rho)$ *s.t. whenever* $m \geq \underline{m}(h,k,\rho)$, *the process converges almost surely to a pure absorbing set.*

**Proof.** Appendix. ■

The intuition is essentially that since beliefs are formed by drawing imperfect samples from the past there is always positive probability to draw "favorable" beliefs enabling convergence. This is only true for acyclic games, though. In games with best response cycles, such as e.g. the matching pennies game convergence to a pure absorbing state cannot be ensured. What about the bound $\underline{m}(h,k,\rho)$ ? The bound given in Young's (1993) theorem is clearly sufficient also here. Typically, though, the conditions will be weaker than in Young, since history-dependent memories are not always uniform. To obtain a tight bound is difficult.

Proposition 2 establishes that the stochastic process converges with probability one to a pure absorbing set. A natural question that arises is whether some of these absorbing sets are more likely to be observed in the long run than others. The previous results suggest that this might be the case.

Also note that the freedom to choose "off equilibrium" beliefs freely may sometimes be crucial. For example in order to sustain defection in the Prisoner's Dilemma as an absorbing profile the "off equilibrium belief" that the partner in the T-period interaction will cooperate after the history ("cooperate, defect") should not be too high, because else players can induce joint cooperation by switching once unilaterally. Depending on the game parameters and the degree of forward looking this may maximize $V(\mu^{it}(H),(a^{\tau}))$. This freedom to select off-equilibrium beliefs freely thus may question the robustness of some absorbing states. In the next subsection we will perturb the process a little and study which of the absorbing states survive under these conditions. More precisely we will check which of the absorbing states are also stochastically stable.

## 3.2 Stochastically Stable States in $2 \times 2$ Games

For our analysis of stochastically stable states we will focus for simplicity on $2 \times 2$ games. Consider the following payoff matrix

$$\begin{array}{|c|c|c|} \hline & z_1 & z_2 \\ \hline z_1 & \lambda, \lambda & 0, \theta \\ \hline z_2 & \theta, 0 & \delta, \delta \\ \hline \end{array} . \tag{3}$$

If $\theta > \lambda > \delta > 0$ this matrix represents a Prisoner's Dilemma. If $\lambda > \theta$ and $\delta > 0$ it represents a Coordination game. (If in addition $\theta = 0$, we have a pure Coordination game and if $\lambda > \max\{\theta, \delta\}$ the game is one of Common Interest).

We will focus on the different cases in turn. We adopt the notational convention that $\overrightarrow{a}$ denotes any action profile as before and $\overrightarrow{z}_j = (z_j, z_j), j = 1, 2$ is the profile where action $z_j$ was chosen by both agents.

### 3.2.1 Prisoner's Dilemma

Before we start our analysis of stochastically stable states, let us first describe the entire set of absorbing states for this game. It is quite obvious that states involving defection ($z_2$) in all periods can be absorbing (since $(z_2, z_2)$ is a strict NE this follows from Proposition 1). The more interesting question, though, is under which conditions states involving cooperation in some periods can be absorbing and how such states will look like. Note that, since cooperation is pareto efficient we know from Lemma 4 that such conditions will exist. Our first observation is the following.

**Claim 1** *The paths of play induced by absorbing sets involving cooperation satisfy non-increasing cooperation (NIC), i.e. they are such that - within any $T-$period interaction - if $a_i(t) = z_1$ then also $a_i(t-1) = z_1$.*

**Proof.** Appendix. ■

The Claim states that the probability to observe cooperation within a given $T-$period game is non-increasing in $t$. 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. For any given degree of forward-looking $k$ the (negative) effect of on total future payoffs 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).

Let us now characterize the absorbing sets more directly. The set

$$X_2 = \{s | M(H_i^s) = \{z_2, ..z_2\}, \forall i\} \tag{4}$$

is always absorbing. In addition sets

$$X_1 \subseteq \left\{ s \text{ s.t. } \begin{array}{c} \text{if } \overrightarrow{z_2} \notin H_i^s \Longrightarrow \frac{\left|\left\{\overrightarrow{z_1} \in M_i(H^s)\right\}\right|}{\left|\left\{\overrightarrow{z_2} \in M_i(H^s)\right\}\right|} \geq \frac{T-k}{2}, \text{ if } M_i(\varnothing) = \{z_1, ...z_1\} \\ \text{and if } \overrightarrow{z_2} \in H_i^s \text{ then } M(H_i^s) = \{z_2, ..z_2\}, \forall i \end{array} \right\} \tag{5}$$

are absorbing whenever $\rho^{-1} \left\lceil \frac{m}{T} \right\rceil < \frac{\lambda - \delta}{\lambda}$ (Condition 2 (C2)).[16] This is a sufficient condition that guarantees that $X_1$ is non-empty for $k > 1$, which is also necessary if $k = 2$. Condition 2 is the one from Lemma 4, ensuring that samples are informative enough. The set $X_1$ in (5) is directly characterized through the memory of agents as follows. Conditional on the empty history memory contains only elements $z_1$, after a history containing $\overrightarrow{z_2}$ the memory contains only $z_2$ and conditional on a history containing only $\overrightarrow{z_1}$ the memory contains at least $\frac{T-k}{2}$ elements $z_1$. (Note that as $T \to \infty$, the memory contains infinitely more

[16] For a proof of why this is the bound see the Appendix.

elements $\overrightarrow{z_1}$ than $\overrightarrow{z_2}$). Together with the assumption of best responses and C2 the property of non-increasing cooperation is implied.

Inspection of the first row in (5) may suggest that higher values of $k$ lead to "less" cooperation in absorbing states (once $k > 1$). This is not quite true, though. The reason is that with higher $k$ the payoff conditions that make a particular set $X_1$ absorbing become weaker. Note also that the share of $z_1$ entries in $M(H^s)$ is strictly increasing with $T$. Thus as $T \to \infty$ induced paths will be almost entirely cooperative. (Note that joint cooperation is efficient and thus pure absorbing under the conditions of Proposition 1).

Assume now that Condition 2 holds s.t. both sets $X_1$ and $X_2$ are absorbing. Then we can state the following proposition.

**Proposition 3** *If* $(h, k) >> (0, 1)$, *C2 holds and* $\frac{\rho-1}{\rho} \in (\frac{\delta}{\theta-\delta}, \frac{\lambda+2\theta+\delta}{\lambda+2(\theta+\delta)}]$, *then all stochastically stable states are contained in* $X_1$. *Else stochastically stable states can be contained in either* $X_1$ *or* $X_2$.

**Proof.** Appendix. ■

Two conditions are needed for this result. Condition C2 ensures that samples are "informative" enough s.t. agents beliefs conditional on histories containing only $\overrightarrow{z_1}$ place high enough probability on the opponent choosing cooperation again. C2 is necessary condition. The condition $\frac{\rho-1}{\rho} > \frac{\delta}{\theta-2\delta}$ on the other is sufficient to prevent too "easy" transitions from any state in $X_1$ to a state in $X_2$ by ensuring that few trembles to defection are never enough to infect the whole population. $\frac{\rho-1}{\rho} \leq \frac{\lambda+2\theta+\delta}{\lambda+2(\theta+\delta)}$ on the other hand is sufficient to enable relatively more "easy" transitions from any state characterized by defection to a state characterized by cooperation. A relatively weaker bound on $\rho$, thus suffices to enable transitions to cooperative states rather than vice versa. The intuitive reason is that, since cooperation is efficient, the range of beliefs sustaining cooperative outcomes is larger than that sustaining outcomes characterized by full defection.

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 $X_1$ to a state in $X_2$ requires less transition than the minimal number of transitions needed from any state in $X_2$ to $X_1$. Since this kind of computation includes all the states, even those through which no minimal mutation passes, the bounds are generally not tight (i.e. the interval generally even larger). Also they do not depend on $h$ or $k$ since they are sufficient conditions only.

### 3.2.2 Coordination Game

Since in this $2 \times 2$ game the only pareto efficient point is a Nash equilibrium and since both Nash equilibria satisfy C1, the pure absorbing sets are simply given by

$$X_j = \{s | M(H_i^s) = \{z_j, ..z_j\}, \forall i\}.$$

Note that no other set can be absorbing, since if $z_j$ is a best response to $\mu(b|\overrightarrow{z_j})$ whenever agents maximize over a horizon corresponding to their degree of forward-looking $k$, it must also be so for a horizon of any length $1, ..k-1$. (Holding fixed the agents beliefs, the longer the horizon over which agents maximize the stronger are the incentives to forego best responding in the current period in order to achieve a better outcome in the future). Note also that $X_1$ and $X_2$ are in general not singleton sets. The reason is that $M(H)$ is not uniquely determined for histories which are "off the equilibrium path".

To make the problem more interesting, let us assume that additionally $\theta+\delta > \lambda > \delta$, implying that $(z_1, z_1)$ is the efficient Nash equilibrium in the one-shot game and $(z_2, z_2)$ the risk-dominant equilibrium. 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 ? Again Young (1993) has analyzed this question for $2 \times 2$ 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 4** *There exists $\widehat{\rho}(\theta, \lambda, \delta)$ s.t. whenever $\rho \geq \widehat{\rho}(\cdot)$ and $(h, k) >> (0, 1)$ all stochastically stable states are contained in $X_1$. Else stochastically stable states can be contained in either $X_1$ or $X_2$.*

**Proof.** Appendix. ■

The exact value of $\widehat{\rho}(\theta, \lambda, \delta)$ is derived (implicitly) in the proof. The intuition is again quite simple. In the myopic case a unilateral tremble starting from the risk dominant equilibrium is not as detrimental (yielding a payoff of $\theta > 0$) as a tremble starting from the efficient equilibrium (yielding a payoff of zero). Heuristically speaking then less trembles will typically be needed to reach the risk-dominant equilibrium than to leave it. This continues to be true in the forward looking case only if it does not lead to a change in the opponent's behavior. 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 actually be less detrimental than those starting from the risk dominant action.

Note that the threshold value $\widehat{\rho}(\theta, \lambda, \delta)$ does neither depend on $h$ nor $k$. The reason is that the Proposition describes a sufficient condition. Even in the case least favorable to the efficient convention (the case $(h, k) = (1, 2)$), the threshold value derived in the Appendix will suffice to single out efficient states as stochastically stable. The more forward looking agents are the weaker the conditions on $\rho$ will be that suffice to get this result.

## 3.3 Application to Experimental Results

In this subsection we want to illustrate how the results from the previous subsection (in particular 3.2.1) can provide an alternative explanation for experimental data from finitely repeated Prisoner's dilemma games. An experiment that is relatively well suited to test our theory was conducted by Andreoni and Miller (1993).

The treatment that is most closely related to our theoretical set-up is their "Partner treatment". In this 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 10-period games, i.e. for a total of 200 rounds of the prisoner's dilemma. Their main results can be summarized as follows. There is significantly more cooperation in the first 5 rounds of each game than in the last five rounds. In the last two 10 period interactions the percentage of cooperation ranges from 60% to 85% until round 6 roughly. Afterwards cooperation breaks down (to 10%).

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 in any 10-period game programed 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 round, i.e. chooses $z_j$ at $t$ whenever $a_{-i}(t-1) = z_j$. In this treatment there is still significantly more cooperation in the first 5 rounds of each game than in the last five rounds. The percentage of cooperation now ranges between 60% and 70% until round 8 roughly. Afterwards cooperation breaks down (to 10%). In this treatment, thus, cooperation is sustained two periods longer on average.

The payoffs in the Prisoner's Dilemma in their experiment were given by

$$\begin{array}{|c|c|c|}\hline & z_1 & z_2 \\ \hline z_1 & 7,7 & 0,12 \\ \hline z_2 & 12,0 & 4,4 \\ \hline \end{array}. \tag{6}$$

Can we explain their findings with our model ? First note that our sufficient condition to rule out defection as a stochastically stable state yields $\rho \in (2, 9]$ and $\rho^{-1}\left\lceil \frac{m}{10} \right\rceil < \frac{3}{7}$. This is satisfied e.g. if $\rho = 5$ and $m = 10$. But since we do not know $\rho$ and $m$, in principle, both sets $X_1$ and $X_2$ as defined in (4) - (5) might be stochastically stable. Observe also that the experimental evidence (not only in their experiment) largely satisfies the property of non-increasing cooperation rates over time, that we stated in Claim 1 above. We can say much more, though. Focus on the "Partner"-treatment first. If we assume for simplicity that $h = 1$ for all agents, we can state the following result.[17]

**Claim 2** *If $(h, k) = (1, 5)$ and $\rho^{-1}\left\lceil \frac{m}{10} \right\rceil < \frac{3}{7}$ the path of play were agents cooperate in the first six rounds of all $T-$period interactions and defect afterwards is induced in the unique stochastically stable state.*

**Proof.** Appendix. ■

If $m$ is not too large (in fact $m \leq 13$), this path of play induces beliefs $\mu(z_1|(z_1, z_1)) \geq 5/6$ and $\mu(z_1|(z_2, z_2)) = 0$. Given these beliefs off-equilibrium

[17] Of course it is also possible to induce this path with higher values of $h$, since more sophistication allows also for behavior as if $h = 1$.

beliefs have to satisfy $\mu(z_1|(z_2,z_1)) \in [0.42, 0.49]$ in order for such a path to be part of a stable state.[18] Obviously we do not know what beliefs of the participants in the experiment were. We can look, though at actual play in the first 100 rounds of the experiment. We find that for the partner-treatment the probability to observe cooperation ($z_1$) after a round of mutual cooperation is roughly 0.83 and after a round of mutual defection is roughly 0.1. Given this we would need $\mu(z_1|(z_2,z_1)) \in [0.32, 0.41]$ for the above to be a stable state. Again we can't observe beliefs but average play shows roughly 30% cooperation after a history of oneself defecting and the opponent cooperating. This is at the lower bound of permissible beliefs. Thus if we think that the actual path of play is roughly consistent with the beliefs of the agents, our learning process can provide an explanation for their results.

What happens now if agents know that there is 50% chance of meeting a tit-for-tat player in each given $T-$period interaction ? Holding fixed the degree of forward looking for all agents, it is intuitive to expect that agents will have stronger incentives to cooperate in this case. The following Claim confirms this intuition.

**Claim 3** *If $(h,k) = (1,5)$, $\rho^{-1}\left\lceil \frac{m}{10} \right\rceil < \frac{3}{7}$ 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 rounds of all $T-$period interactions and defect afterwards is induced in the unique stochastically stable state.*

**Proof.** Appendix. ■

Note that $(h,k) = (1,5)$ here obviously is a condition on the human players only, since the computer players are pre-programmed to tit-for-tat as explained above. Obviously we would expect more cooperation in the presence of such players and this is indeed what we find. For the induced beliefs the following holds. If $m \leq 19$, on the equilibrium path $\mu(z_1|(z_1,z_1)) \geq 8/9$ and $\mu(z_1|(z_2,z_2)) = 0$ have to be satisfied. Given this, off-equilibrium beliefs would have to satisfy $\mu(z_1|(z_2,z_1)) \geq 0.12$ in order for such a path to be part of a stable state. In the data we find indeed that after a history of joint cooperation the probability that one's opponent will cooperate again is roughly 95%. After a history of joint defection this probability is roughly 11% and after a history $(z_2,z_1)$ this probability is roughly 13%. (Note that this substantially lower value compared to the first treatment is due to the presence of the tit-for-tat players who always defect after observing a defection by the opponent). If the agent's beliefs are only roughly consistent with their experience, then again our learning process again can explain their results.

[18] It has to be optimal for a player to cooperate in rounds $2, ...6$, and to defect in rounds $7, ..10$ given the induced beliefs. From this set of inequalities the feasible interval can be calculated. Note also that this interval is much larger if we relax the condition that all agents should defect in round $7$. See also the Appendix.

# 4 Extensions

## 4.1 Heterogenous Agents

A natural question that arises is whether agents with a higher degree of forward-looking ($k$) will always be able to exploit others with a lower degree of forward looking. We will see that this is not always the case. To see this consider the following example. Assume that there are two types. $k_1$ is a myopic type with $(h,k)=(1,1)$ and $k_2$ is forward-looking characterized by $(h,k)=(1,2)$. Denote the share of $k_1$ agents by $\sigma$. Irrespective of their type, agents are randomly matched to play a $4-$period repeated Prisoner's Dilemma. (Since the game is symmetric we simply assume that agents are matched randomly within $C_1 \cup C_2$.) The stage game payoffs are given by (3). We want to consider two different scenarios. In the first agents know that the population is heterogenous 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 homogenous or that they are simply never able to observe (or infer) the type of their opponent.

**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 rounds) depending on the behavior they observe in earlier rounds.

**Claim 4** *If $\sigma < \frac{3\lambda-\theta-2\delta}{3\lambda-\theta-\delta}$, then forward looking agents ($k_2$) obtain higher average payoffs in all absorbing states. If $\sigma \in \left[\frac{3\lambda-\theta-2\delta}{3\lambda-\theta-\delta}, \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}\right]$ then myopic agents ($k_1$) obtain higher average payoffs in all absorbing states and if $\sigma > \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}$ all agents obtain the same average payoff in all states.*

**Proof.** Appendix. ■

The condition $\sigma < \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}$ is simply necessary for absorbing states with cooperation to exist at all. Given that they do exist, forward looking agents do only make higher profits in expectation if $\sigma$ 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 rounds of of their horizon $(t+1,...t+k)$. 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 $\sigma$ where they should be choosing defection.

These results have natural implications in terms of evolution. More precisely, one could argue that two outcomes could be identified as stable given standard Replicator Dynamics (possibly with some drift). The state where all agents are forward looking ($\sigma = 0$) would be an attractor of such a system, since for $\sigma$ small enough forward looking agents always make higher profits. The other state that would be stable is the state where $\sigma = \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}$, i.e. a state where both myopic and forward looking agents are present.[19]

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. Whether they would always have higher payoffs will obviously depend on the degree to which matching is assortative.[20]

**Unconditional Beliefs**

Let us focus next on the case where agents are not able to infer the type of their opponents (or simply assume that the population is homogenous) 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.

**Claim 5** *If beliefs are unconditional all absorbing states involve full defection and all agents obtain the same payoff in expectation.*

**Proof.** Appendix. ■

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 empty 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.

## 4.2 Node-dependent Beliefs ($h \geq T-1$)

Note that in our setting agents are generally not aware of (or do not take into account) which round (decision node) they are currently in.[21] They only care about the current history when making their decisions. Conditioning on the decision node in addition certainly involves more computational resources. In fact node-dependent beliefs emerge as the sophistication of agents increases, in particular whenever $h \geq T-1$. This suggests that node-dependent beliefs may be a reasonable model only whenever $T$ is small. Still we want to investigate

[19] This point would be Lyapunov stable in a model without drift and asymptotically stable in a model with drift (where drift pushes the dynamics towards the interior of the state space).

[20] See e.g. Myerson, Pollock and Swinkels (1991) or Mengel (2007,2008).

[21] An exception is the first decision node in each $T-$period interaction which is always preceded by the empty history.

whether (and which) results would change, if agents formed beliefs $\mu(a|H,\tau)$ that depend on the current round of play $\tau = 1, ...T$. The following Proposition summarizes the main differences.

**Proposition 5** *If agents have node dependent beliefs then the size of memory $m$ will not affect absorbing states. In $2 \times 2$ games outcomes will always induce Nash equilibria, but in larger games locally efficient outcomes can also be sustained.*

**Proof.** Appendix. ■

Remember that the reason why memory affects the set of absorbing states in the general case, was that too large values of $m/\rho$ tend to make samples uninformative, since behavior that is optimal only for some rounds $T, T-1, ..$ might be dramatically overrepresented in any given sample with positive probability. This can imply that convergence to some efficient states is not possible. With node-dependent beliefs this is obviously not the case anymore, since if memory and thus beliefs are conditioned on the decision nodes then samples will always be informative of the behavior at that node.

Still though, the Proposition shows that this need not be unambiguously good for efficient outcomes to arise. The reason simply being that if beliefs are conditioned on the decision node agents at $T-1$ will eventually learn that whatever they choose, their opponent will best respond in $T$. But then for $2 \times 2$ games the typical backward induction logic kicks in and we will end up with Nash behavior at each node. In larger games, though individually rational and locally efficient outcomes can still be sustained even if they are not Nash. the reason is that agents can still believe that deviating from the efficient action will induce their opponent to choose a third (and possibly worse) action with high probability. Note again that this is true irrespective of whether there is a discrepancy between the Nash and the minmax payoff and irrespective of whether the game in question has a unique Nash equilibrium or not.

When can we expect node - independence of beliefs or a small degree of sophistication of $h$ ? This will certainly depend on whether a) the cognitive cost of holding memories is large and b) whether the time horizon $T$ (while finite) is large.

## 4.3 Anticipating Belief Changes ($k \to \infty$)

Finally let us also discuss which changes a "large" degree of forward looking (in the extreme case $k \to \infty$) would imply. First it follows quite immediately from our previous results that a larger $k$ will tend to make it more easy to sustain efficient outcomes.

One might also wonder, though, what would happen if agents anticipated the effect their action choice at $t$ has on their opponent's beliefs in future rounds. Such belief changes should not be taken into account by adaptive agents will explicitly, but what if they were somewhat more strategic ? Note though, that since we assume that agents care only about $k < T$ periods, but typically have a relatively large memory these effects will be negligible. They could become important, though, if agents had a larger horizon or even unlimited foresight (i.e. $k \to \infty$). To analyze these effects is beyond the scope of this paper. See Blume (2004) for a model of unlimited forward looking players focusing on these effects.

## 5 Conclusions

In this paper we have studied agents interacting in finitely repeated games. Agents are adaptive learners, but also forward-looking to some degree. We have shown that in a pure absorbing set either Nash equilibria satisfying a very weak conditions or individually rational and locally efficient profiles can be induced. In $2 \times 2$ 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.

Further research could extend on Section 4.1 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, since it seems to provide a very intuitive foundation of many other experimental results.

## References

[1] Andreoni, J. and J.Miller (1993), Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence, The Economic Journal 103, 570-585.

[2] Basu, K. and J. Weibull (1991), Strategy Subsets Closed Under Rational Behavior, Economics Letters 36, 141-146.

[3] Binmore, K., J. Mc Carthy, G. Ponti, L. Samuelson and A. Shaked (2001), "A Backward Induction Experiment", Journal of Economic Theory, 104(1), 48-88.

[4] Blume, L. (2004), "Evolutionary Equilibrium with Forward-Looking Players", working paper Santa Fe Institute.

[5] Boyd, R. and P. Richerson (2005), The Origin and Evolution of Cultures, Oxford University Press.

[6] Camerer, C.F., T-H. Ho, and J-K, Chong (2002), "Sophisticated Experience-Weighted Attraction Learning and Strategic Teaching in Repeated Games", Journal of Economic Theory, 1-52.

[7] Ehrblatt, W.Z., K. Hyndman, E. Oezbay and A. Schotter (2008), "Convergence: An Experimental Study of Teaching and Learning in Repeated Games", working paper NYU.

[8] Freidlin, M. I. and A. D. Wentzell (1984), Random Perturbations of Dynamical Systems, New York: Springer-Verlag.

[9] Fudenberg, D. and D. Levine (1989), "Reputation and Equilibrium Selection in Games with a Patient Player", Econometrica 57, 759-778.

[10] Fudenberg, D. and D. Levine (1998), "The Theory of Learning in Games", MIT-Press, Cambridge.

[11] Fujiwara-Greve, T. and C.Krabbe-Nielsen (1999), "Learning to Coordinate by Forward Looking Players", Rivista Internazionale di Scienze Sociali CXIII(3), 413-437.

[12] Gueth, W., R. Schmittberger and B. Schwarze (1982), "An experimental analysis of ultimatum bargaining", Journal of Economic Behavior and Organization 3(4), 367-388.

[13] Jehiel, P. (1995), "Limited Horizon Forecast in Repeated Alternate Games", Journal of Economic Theory 67, 497-519.

[14] Jehiel, P. (1998), "Learning to play Limited Forecast Equilibria", Games and Economic Behavior 22, 274-298.

[15] Jehiel, P. (2001), "Limited Foresight may Force Cooperation", Review of Economic Studies 68, 369-391.

[16] Karlin, S. and H. M. Taylor (1975), A first course in stochastic processes, San Diego: Academic Press.

[17] Kandori, M., G. Mailath and S. Rob (1993), "Learning, mutation, and long run equilibria in games," Econometrica 61, 29-56.

[18] Mengel, F. (2007), "The Evolution of Function-Valued Traits for Conditional Cooperation", Journal of Theoretical Biology 245, 564-575.

[19] Mengel, F. (2008), "Matching Structure and the Cultural Transmission of Social Norms", Journal of Economic Behavior and Organization 67, 608-623.

[20] Myerson, R.B., G.B. Pollock and J.M. Swinkels (1991), Viscous population equilibria, Games and Economic Behavior 3, pp. 101–109.

[21] Selten, R. (1991), "Anticipatory learning in two-person games" in: Game Equilibrium Models I (R.Selten ed.), 98-154. Springer-Verlag Berlin.

[22] Stahl, D.O. (1993), "Evolution of Smart$_n$ Players", Games and Economic Behavior 5, 604-617.

[23] Terracol, A. and J. Vaksman (2008), "Dumbing Down Rational Players: Adaptive Learning and Teaching in an Experimental Game", Journal of Economic Behavior and Organization, in press.

[24] Ule, A. (2005), "Exclusion and Cooperation in Networks", PhD thesis, Tinbergen Institute.

[25] Watson, J. (1993), "A reputation refinement without equilibrium", Econometrica 61, 199-205.

[26] Young, P. (1993), "The Evolution of Conventions", Econometrica 61(1), 57-84.

[27] Young, P. (1998), "Individual Strategy and Social Structure", Princeton University Press.

# A Appendix: Proofs

Denote by $BR_i(\cdot)$ player $i$'s best response correspondence for the one shot game.

**Proof of Lemma 1**

**Proof.** Consider an absorbing action profile $((a,b)^*_\tau)_{\tau=t-m+1,...t.}$ where the same actions are chosen at each time $t$ by both players. Focus wlg on player 1. Either $a^* \in BR_1(b^*)$. But then $a^*$ must guarantee the maxmin payoff $\widehat{\pi}$ to player 1. If $a^* \notin BR_1(b^*) \wedge \pi(a^*,b^*) < \widehat{\pi}$ then this must be because player 1 believes that a deviation (to say $a'$) yields a higher payoff in the future, i.e. for some $\tau \in [t+1, t+\min\{k,h\}]$ within the same $T-$period interaction. But then again at $\tau$ the same argument holds, i.e. there must be a $\tau' \in [\tau+1, t+\min\{k,h\}]$ for which this is true. Applying this argument recursively shows that she can guarantee herself the maxmin payoff at $\tau \in [t+1, \tau+k]$ and thus at $t$. ■

**Proof of Lemma 2**

**Proof.** As the profile $(a^*,b^*)$ is not a Nash equilibrium, there must exist at least one player $i$ s.t. $a^* \notin BR_i(b^*)$ at some $t$. Thus $(a^*,b^*)$ can only be optimal for player $i$ if she believes that deviating will reduce her payoff in some periods $\tau \in [t+1,..,t+k]$. But if $(a^*,b^*)$ is not pareto efficient then either $(a',b^*)$ or $(a',b')$ must yield a higher payoff to both players for $a',b' \neq a^*,b^*$.[22] But since the game is $2\times 2 : a' \in BR_i(b^*)$. Hence the previous statement cannot be true. ■

**Proof of Lemma 3**

**Proof.** Denote the efficient profile $\overrightarrow{a}^*$ and the history consisting of the efficient profile only by $H(\overrightarrow{a}^*)$. Since in some rounds $\tau \in [T-k+2,...,T]$ players will deviate from $\overrightarrow{a}^*$ some beliefs $\mu(a^*_{-i}|H(\overrightarrow{a}^*)) \in \text{rint}\,\Delta A_i$ can be drawn with positive probability. (In fact all those that are multiples of $\frac{1}{\rho}$). But then if the set $A'$ (containing both $a^*$ and the best response chosen in $T$) is not curb divergence may occur with positive probability. Finally $\overrightarrow{a}^*$ has to be efficient in $A'$ by Lemma 2. ■

**Proof of Lemma 4**

**Proof.** We will show both necessity and sufficiency, starting with necessity. First note that as $a^*$ is not a Nash equilibrium it cannot be a singleton curb

[22] If this is not true for player $i$ it must be true for player $-i$.

set and thus Lemma 3 states that it must be pareto efficient in a set of at least cardinality 4.[23] If $a^*$ is pareto efficient in a set of cardinality $\geq 4$ then it must be so also in a set of cardinality 4. But if it is not pareto efficient in a set of cardinality 4 it cannot be induced by Lemma 2. Now we will show that this set, denote by $A'$ has to be $\underline{\mu}(a^*)-$curb. If the set $A' = A'_1 \times A'_2$ is not curb $\Rightarrow \exists \mu^{;}_i \in \Delta A'$ where $\mu^{;}_i(a^*) \geq \underline{\mu}(a^*)$ s.t. $BR_{-i}(\mu^{;}_i) \notin \Delta A'$. Furthermore as $a^*$ is not a Nash equilibrium, some player $i$ must have a better response $a$", which will be chosen in a T-period interaction for some $\tau \in [T-\tau, T]$ after a history $(a^*, ...a^*)$. But then there is strictly positive probability that at some point $t$ player $i$ will hold a belief $\mu^{;}_i \in \Delta A'$ where $\mu^{;}_i(a^*) \geq \underline{\mu}(a^*)$ s.t. $BR_{-i}(\mu^{;}_i) \notin \Delta A'$.

Next we will show sufficiency. Denote by $\overrightarrow{a}^* = (a^*, b^*)$ an individually rational and locally efficient action profile. We want beliefs to satisfy a) $\mu(b, (\overrightarrow{a}^*, ..., \overrightarrow{a}^*))$ is s.t. $BR^t\left[\mu(b, (\overrightarrow{a}^*, ..., \overrightarrow{a}^*))\right] = a^*$, $\forall t \leq T-1$ and b) $\mu(b, (\overrightarrow{a}^*, ..., a'))$ is such that $BR^t\left[\mu(b, (\overrightarrow{a}^*, ..., a'))\right] = a'', \forall a', a'' \in A'$ where $A'$ contains $a^*$ and is $\underline{\mu}(a^*)-$curb. Now consider an absorbing state where all $T-$period interactions are identical and look as follows, $(\underbrace{\overrightarrow{a}^*, ..., \overrightarrow{a}^*}_{T-k+x \text{ Rounds}}, \overrightarrow{a}")$, where $x \in [1, k-1]$. Then $\mu(b^*, (\overrightarrow{a}^*, ..., \overrightarrow{a}^*)) \geq 1 - \rho^{-1}\left\lceil \frac{m}{T} \right\rceil$ and $\mu(b''|(\overrightarrow{a}^*, ..., \overrightarrow{a}^*)) \leq \rho^{-1}\left\lceil \frac{m}{T} \right\rceil$ since memory of size $m$ permits to draw $\overrightarrow{a}''$at most $\left\lceil \frac{m}{T} \right\rceil$ times in a sample of size $\rho$. (This is exactly true if $k = 2$ and a sufficient condition if $k > 2$. For the necessity part $\eta$ will depend on $k$ then). Also $\mu(b'|(\overrightarrow{a}^*, ..., \overrightarrow{a}^*)) = 0, \forall b' \neq b^*, b''$. On the other hand beliefs can be chosen s.t. $\mu(b''|(\overrightarrow{a}^*, ..., a')) > 0 \Rightarrow b'' \in A'_{-i}$ holds. Finally assuming that $\rho^{-1}\left\lceil \frac{m}{T} \right\rceil \leq \eta$ for $\eta$ small enough s.t. $BR\left[1 - \rho^{-1}\left\lceil \frac{m}{T} \right\rceil\right] = a^*, \forall t \leq T-k+x$ yields the result. ■

**Proof of Proposition 1**

**Proof.** Part (ii) follows directly from Lemma 1-4 as well as from the proof of Lemma 4. For part (i) the proof is as follows. Consider any state where the NE $\overrightarrow{a}^*$ is played at each $t$. We will first show that if C1 is satisfied such a state is absorbing. In order for such a state to be absorbing beliefs have to satisfy $\mu(a^*|(\overrightarrow{a}^*, ..., \overrightarrow{a}^*)) = 1$ and $\mu(b|(\overrightarrow{a}^*, ..., (a'_i, a^*_{-i}))$ is s.t. $\sum\limits_{\tau=t}^{t+k-1} \sum_{b \in A} \mu^{i\tau}(b|H(\tau - 1))\pi^i(a, b) - k\pi(\overrightarrow{a}^*) < 0$. But beliefs $\mu(b|(\overrightarrow{a}^*, ..., (a'_i, a^*_{-i}))$ that guarantee that the previous inequality holds always exist whenever C1 is satisfied. Next we prove necessity. Assume C1 is not satisfied, in particular assume that there exists $a'_i$ s.t. $\pi^i(a'_i, a_{-i}) \geq \pi^i(\overrightarrow{a}^*), \forall\ a_{-i} \in A_{-i}$. Then $\nexists \mu(b|(\overrightarrow{a}^*, ..., (a'_i, a^*_{-i}))$ for which player $i$ would strictly prefer to choose $a^*_i$ rather than $a'_i$. Hence divergence may occur. ■

**Proof of Proposition 2**

**Proof.** We will show that there exists a number $K \in \mathbb{N}$ and a probability $p$ s.t. 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 $a^*$ the

[23] Note that any set of cardinality 2 or 3 containing $(a^*, b^*)$ cannot be curb.

profile chosen at $t$. If $H(t+1) = H(t) = (a^*, ...a^*)$ then we can go to step (ii) of the proof (setting $t = \tau''$). Assume $H(t+1) \neq H(t)$. Then, since the set of all possible histories $\mathcal{H}$ is finite, $\exists \tau' > t$ s.t. $H(\tau') = H(\tau)$ 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 $H(\tau)$ any finite number of times. At history $H(\tau)$, there is positive probability, that each agent samples the last $\rho$ plays in $M(H(t))$. Denote this sample by $\xi$. There is also positive probability that the next $\rho$ times that the history is $H(t)$ the agent samples $\xi$ again and chooses the same best response.

(ii) Order the histories according to $\tau$ as follows: $H(\tau), H(\tau+1), ..H(\tau'-1)$. Now assume there exists $H(\tau'') \in [H(\tau), H(\tau'-1)]$ where $H(\tau'') =: (a^*, ..a^*)$ is part of an absorbing set. Then there is positive probability to sample only the last $\rho$ rounds for the next $m - \rho$ periods thereby creating a homogenous memory $M(H(\tau'')) = (a^*_{-i}, ..a^*_{-i})$. Since $a^*_i \in BR(a^*_{-i})$ 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 $\rho$ periods where the history was $H(\tau'')$, i.e. takes a homogenous sample $(a, ...a)$. The best response to $(a, ...a)$ for each agent lies on a directed path leading to an absorbing set since the game is acyclic. Again now $\exists \tau''' > \tau''$ s.t. $H(\tau''') = H(\tau^{iv})$ 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 $\rho$ periods $\forall H(\tau^{iv})...H(\tau''' - 1)$. If there is a history in $H(\tau^{iv})...H(\tau''' - 1)$ 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 $(a, ...a)$ to a history $(a^*, ..a^*)$ 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. In order for that to be possible $m$ needs to be big enough, since some agents have to be able to look back far enough.

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. ■

**Proof of Absorbing Sets Prisoner's Dilemma:**

**Proof.** That the set $X_2$ is absorbing follows directly from Proposition 1. The proof that $X_1$ is absorbing (under the conditions mentioned) follows from Lemma 4. It remains to show that the upper bound on $\rho^{-1} \left\lceil \frac{m}{T} \right\rceil$ is given by $\frac{\lambda - \delta}{\lambda}$. First note that the most restrictive conditions (for the efficient profile to be absorbing) are encountered in the case $k = 2$ and $h = 1$. In this case the condition is that both players have to find it advantageous to choose $z_1$ $(z_1)$ after a history of $\overrightarrow{a}_1$.

$$V(\mu(\overrightarrow{a}_1), z_1) > V(\mu(\overrightarrow{a}_1), z_2) \Leftrightarrow \mu(z_1|\overrightarrow{a}_1) > \frac{\delta}{\lambda},$$

where we have set $\mu(z_1|(z_2,z_1)) = 0$. But then since $M^s$ contains at most $\left\lceil \frac{m}{T} \right\rceil$ choices of $z_2$ and $\rho$ elements from $M^s$ are randomly drawn to form this belief. The inequality $\rho^{-1}\left\lceil \frac{m}{T} \right\rceil < 1 - \frac{\delta}{\lambda} = \frac{\lambda-\delta}{\lambda}$ follows. Also note that there can be no other absorbing states not contained in either $X_1$ or $X_2$. Since starting from any state outside $X_1$ involving some cooperation there is always positive probability to repeatedly draw beliefs which will lead to convergence to $X_2$. ■

**Proof of Claim 1**

**Proof.** Assume that at round $t$ (within a given $T-$period interaction) beliefs of agent $i$ are such that she finds it optimal to choose cooperation ($z_1$). If $\lceil t \rceil_T - t \geq k$ (where $\lceil t \rceil_T$ denotes the smallest multiple of $T$ larger than $t$), then the maximization problem at $t-1$ is identical to that at $t$. Now assume that beliefs at $t-1$ were such that the agent would find $z_2$ optimal. But then (since we are in an absorbing state) it cannot be that beliefs change in such a way that cooperation is optimal at $t$. What if $\lceil t \rceil_T - t < k$ ? Then at $t$ the agent will have strictly less "foresight" than at $t-1$. But then defection ($z_2$) will seem relatively better to cooperation ($z_2$) at $t$ compared to the situation at $t-1$ where the agent expects $k$ more periods. The reason is that choosing defection must always reduced 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 $t$). But given this again cooperation must follow at $t-1$. ■

**s-trees**

For most of the following proofs we will rely on the graph-theoretic techniques developed by Freidlin and Wentzell (1984).[24] They can be summarized as follows. For any state $s$ an $s-$tree is a directed network on the set of absorbing states $\Omega$, whose root is $s$ and such that there is a unique directed path joining any other $s' \in \Omega$ to $s$. For each arrow $s' \rightarrow s''$ in any given $s-$tree the "cost" of the arrow is defined as the minimum number of simultaneous trembles 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 3**

**Proof.** (i) Consider first transitions from $X_2 \rightarrow X_1$. 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 at each $t < T$. Note that $\kappa_{C(1)} > 1$ will hold for any $s \in X_2$, since otherwise $s$ couldn't have been absorbing in the first place.

Now assume that at $t$ player 1 trembles s.t. the action profile is $(z_1, z_2)$ and that then at $t+1$ player 2 trembles s.t. $\overrightarrow{a}(t+1) = (z_2, z_1)$. Consider choices at $t+2$. Now player 1 will choose $z_1$ whenever $\mu(z_1|(z_1,z_2)) > \frac{\delta}{\lambda+2(\theta-\delta)} =: \widehat{\mu}_1$ (see below). But then since beliefs are formed by drawing randomly $\rho$ out of the last $m$ observations, this implies that we need $\frac{1}{\rho} \geq \widehat{\mu}_1$ in order to have $\kappa_{C(1)} = 2$. Where does $\widehat{\mu}_1 = \frac{\delta}{\lambda+2(\theta-\delta)}$ come from ? First note that the least favorable case

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

for such a transition is the case with $(h,k) = (1,2)$. Then we observe that

$$\begin{aligned} V(\mu,(z_1,z_2)) &= \mu(z_1|(z_1,z_2))\left[\lambda + (\mu(z_1|\overrightarrow{z_1})\theta + (1-\mu(z_1|\overrightarrow{z_1}))\delta\right] \qquad (7) \\ &\quad +(1-\mu(z_1|(z_1,z_2))\left[\mu(z_1|(z_1,z_2)\theta + (1-\mu(z_1|(z_1,z_2))\delta\right] \text{ and} \\ V(\mu,(z_2,z_2)) &= \mu(z_1|(z_1,z_2))\left[\theta + \mu(z_1|(z_2,z_1))\theta + (1-\mu(z_1|(z_2,z_1)))\delta\right] \\ &\quad +(1-\mu(z_1|(z_1,z_2)))\left[\delta + \mu(z_1|\overrightarrow{z_2})\theta + (1-\mu(z_1|\overrightarrow{z_2}))\delta\right]. \end{aligned}$$

We then want to find conditions on $\mu(z_1|(z_1,z_2))$ such that $V(\mu(\cdot),(z_1,z_2)) > V(\mu(\cdot),(z_2,z_2))$ *for all* candidate states $s \in X_2$. Clearly $\mu(z_1|\overrightarrow{z_2}) = 0$ is determined "on the equilibrium path". By setting $\mu(z_1|(z_2,z_1)) = 0$ and $\mu(z_1|\overrightarrow{z_1})$ to either $\{0,1\}$ we obtain the threshold above. (We can set $\mu(z_1|(z_2,z_1)) = 0$ since the state with the corresponding memory can be reached from any other state in $X_2$ by a sequence of one-trembles which is not true for the reverse).

Finally note that after two agents have been infected (through $\kappa_{C(1)} = 2$ trembles) the whole population can be infected. The reason is that whenever the infected agents are rematched they will start to cooperate after the empty history since their beliefs $\mu(z_1|(z_2,z_1))$ are sufficiently high and since $k > 1$. The history $(z_2,z_1)$ will be repeated until there are sufficient draws for player 1 to optimally choose $z_1$. This is always possible if $T$ is "large enough" and since $\frac{\rho-1}{\rho} > \frac{\delta}{\theta-\delta}$.

(ii) Let us then turn to the reverse transitions $X_1 \rightarrow X_2$. 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 $t$. First assume that two players simultaneously make a mistake and choose $(z_2,z_2)$ at some time $t$. Then it can be shown by comparing the analogous expressions to (7) that a necessary condition for either player to choose $z_2$ ($z_2$) also at $t+1$ is that $2\delta > \theta$. Secondly assume that player 1 makes two mistakes and chooses $z_2$ at $t$ and $t+1$. 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 $z_2$ at $t+2$ is that $\mu(z_1|(z_2,z_1)) > \frac{\delta}{\theta-\delta}$. To see this compare

$$\begin{aligned} V(\mu,(z_1,z_2)) &= \mu(z_1|(z_2,z_1))\left[\lambda + \mu(z_1,\overrightarrow{z_1})\theta + (1-\mu(z_1,\overrightarrow{z_1}))\delta\right] \\ &\quad +(1-\mu(z_1|(z_2,z_1)))\left[\mu(z_1|(z_2,z_1))\theta + (1-\mu(z_1|(z_2,z_1)))\delta\right] \text{ and} \\ V(\mu,(z_2,z_2)) &= \mu(z_1|(z_2,z_1))\left[\theta + \mu(z_1|(z_1,z_2))\theta + (1-\mu(z_1|(z_1,z_2)))\delta\right] \\ &\quad +(1-\mu(z_1|(z_2,z_1))\left[\mu(z_1,\overrightarrow{z_2})\theta + (1-\mu(z_1,\overrightarrow{z_2}))\delta\right]. \end{aligned}$$

Then it can be seen that a necessary condition for a transition to be possible from *any* state in $X_1$ is that $\mu(z_1|(z_2,z_1)) > \frac{\delta}{\theta-\delta}$. Since $\rho$ rounds are drawn from the memory to form this belief we need$\frac{\rho-1}{\rho} > \frac{\delta}{\theta-\delta}$. By analyzing the analogous expressions for player 1 it can be shown that a transition cannot be induced by player 1 repeatedly choosing $z_2$ starting at $t+2$.

(iii) Combining the conditions found in (i) ad (ii) we first note that $\frac{\rho-1}{\rho} > \frac{\delta}{\theta-\delta} \Rightarrow 2\delta < \theta$. Together with $\frac{1}{\rho} \leq \frac{\delta}{\lambda+2(\theta-\delta)}$ a sufficient condition thus is $\frac{\rho-1}{\rho} \geq \max\{\frac{\delta}{\theta-\delta}, \frac{\lambda+2\theta+\delta}{\lambda+2(\theta+\delta)}\}$, the condition in Proposition 3.

(iv) To finish the proof take any state $s \in X_2$ and consider a minimal $s-$tree. Assume first that there exists a state $s' \in X_1$ s.t. 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. Finally if the shortest transition $s' \to s$ is indirect (passing through other states in $X_1$) 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 $s' \to s''$ s.t. reversing this link saves one "mutation". 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 4**

**Proof.** We will show that there exists $\widehat{\rho}(\theta, \lambda, \delta)$ s.t. whenever $\rho \geq \widehat{\rho}(\cdot)$ a transition from any state in $X_1$ to some state in $X_2$ involves more simultaneous mistakes than a transition from any state in $X_2$ to a state in $X_1$.

(i) Consider first transitions from $X_1$ to $X_2$. Assume that in a given $T-$period interaction one player makes a mistake and chooses $z_2$ in the first round. For some states in $X_1$ (e.g. states where $\mu(z_1|(z_2, z_1) = 0)$ this will be sufficient to induce a pair of agents to end up both choosing $z_2$ in this interaction with positive probability.[25] To induce a transition then one agent from such a pair has to observe $z_2$ often enough after the empty history in order to start choosing $z_2$ in each new interaction. Comparing expected payoffs yields $\mu(z_1|\varnothing) \leq \frac{\delta}{2\lambda+\theta}$ in the case where $(h, k) = (1, 2)$. For higher $(h, k)$ the conditions will be weaker. Since we are interested in a sufficient condition we focus on $(h, k) = (1, 2)$. But then since $\mu(z_1|\varnothing) \geq \frac{\rho-\kappa_2}{\rho}$ where $\kappa_2$ is the number of mutations, we have that at least

$$\widehat{\kappa}_2(\rho) = \left\lceil \frac{\rho(2\lambda+\theta-\delta)}{2\lambda+\theta} \right\rceil$$

mutations are necessary to induce such a transition.

(ii) On the other hand for the reverse transition from any state in $X_2$ to a state in $X_1$ the following number of mistakes $\kappa_1$ are sufficient. First it can be calculated that the number of mutations to ensure a pair of agents to converge to choosing $z_1$ from *any* state in $X_2$ is bound above by $\kappa' = \left\lceil \frac{2\delta}{\rho(\lambda+3\delta-\theta)} \right\rceil$ .(Again this can be found by comparing expected payoffs after assuming that one player of the pair makes $\kappa$ mistakes in a row). If this is true still one of the two players

[25] These states may seem inherently unstable, but remember that we only want to find a *sufficient* condition.

has to experience enough trembles after the the empty history in order to start choosing $z_1$ in each new interaction. If this is true, no additional trembles are needed, since for this player favorable beliefs have positive probability to be drawn, since she has already experienced convergence to $z_1$ once. How many additional trembles are needed can again be calculated comparing expected payoffs. Beliefs have to satisfy $\mu(z_1|\varnothing) \geq \frac{2\delta-\rho^{-1}\lambda}{(2-\rho^{-1})\lambda+\delta-\theta-\kappa'\rho^{-1}(\theta-\delta)}$ and since $\mu(z_1|\varnothing) \leq \frac{\kappa}{\rho}$ a sufficient condition is

$$\widehat{\kappa}_1(\rho) = \left\lceil \frac{2\delta\rho-\lambda}{(2-\rho^{-1})\lambda+\delta-\theta-\left\lceil\frac{2\delta}{\rho(\lambda+3\delta-\theta)}\right\rceil\rho^{-1}(\theta-\delta)} \right\rceil + \left\lceil \frac{2\delta}{\rho(\lambda+3\delta-\theta)} \right\rceil .$$

Now, if $\widehat{\kappa}_1 < \widehat{\kappa}_2$, then all stochastically stable states are contained in $X_1$. Note that $\partial\widehat{\kappa}_2/\partial\rho > \partial\widehat{\kappa}_1/\partial\rho, \forall\rho$ and $\widehat{\kappa}_2(0) = 0$. Hence a fixed point $\widehat{\rho}(\cdot)$ does exist. The proof is completed in analogy to part (iv) of the proof above. ■

**Proof of Claim 2:**

**Proof.** Assume that $\mu(z_1|(z_1,z_1)) = 5/6$, $\mu(z_1|\varnothing) = 1$ and $\mu(z_1|(z_2,z_2)) = 0$ and denote off-equilibrium beliefs $\mu(z_1|(z_2,z_1)) =: x$. By Claim 1, if an agent finds it optimal to cooperate in round 6, she will find it optimal to cooperate in round 1,..5. Also if an agent finds it optimal to defect in round 7, she will find it optimal to do so in rounds 8,..10. We will thus show that under the conditions of the Claim all agents will find it optimal to cooperate in round 6 and to defect in round 7. Denote the vectors $(z_1,z_1,z_1,z_1,z_2) =: \overrightarrow{a}(z_1)$ and $(z_2,z_2,z_2,z_2,z_2) =: \overrightarrow{a}(z_2)$. To show the first claim, it is then sufficient to verify that $V(\mu^{it}(z_1|\overrightarrow{z_1}), \overrightarrow{a}(z_1)) \simeq 33.73$ exceeds $V(\mu^{it}(z_1|\overrightarrow{z_1}), \overrightarrow{a}(z_2)) = 10.4 + 12\sum_{j=1}^{4} x^j + 4\sum_{j=1}^{4}(1-x^j)$. To show the second claim it is sufficient to establish that $V(\mu^{it}(z_1|\overrightarrow{z_1}), \overrightarrow{a}(z_1)') \simeq 27.9$ is smaller than $V(\mu^{it}(z_1|\overrightarrow{z_1}), \overrightarrow{a}(z_2)') = 10.4 + 12\sum_{j=1}^{3} x^j + 4\sum_{j=1}^{3}(1-x^j)$ where $\overrightarrow{a}(z_1)' := (z_1,z_1,z_1,z_2)$ and $\overrightarrow{a}(z_2)' := (z_2,z_2,z_2,z_2)$. Both inequalities are satisfied whenever $x \in [0.42, 0.49]$. Then whenever $m \leq 13$ beliefs will always lie in the relevant intervals and thus this will be absorbing. In fact we have shown that all absorbing states that involve any cooperation at all are characterized by this pattern. ■

**Proof of Claim 3:**

**Proof.** Assume that $\mu(z_1|(z_1,z_1)) = 7/8$, $\mu(z_1|\varnothing) = 1$ and $\mu(z_1|(z_2,z_2)) = 0$ and denote off-equilibrium beliefs $\mu(z_1|(z_2,z_1)) =: \frac{x}{2}$ and $\mu(z_1|(z_1,z_2)) = y$. (Note that now $\mu(z_1|(z_2,z_1))$ is denoted$\frac{x}{2}$, since with probability $\frac{1}{2}$ the agent faces a tit-for-tat player. In analogy to the proof of Claim 2, we will show that under the conditions of the Claim all agents will find it optimal to cooperate in round 8 and to defect in round 9. For this we verify that $V(\mu^{it}(z_1|\overrightarrow{z_1}), (z_1,z_1,z_2)) \simeq 21.9$ exceeds $V(\mu^{it}(z_1|\overrightarrow{z_1}), (z_2,z_2,z_2)) \simeq 16 + 4(x + x^2 + x^3), \forall x \in [0,1]$ and that $V(\mu^{it}(z_1|\overrightarrow{z_1}), (z_1,z_2)) \simeq 17.2 + \frac{2}{3}y$ is smaller than $V(\mu^{it}(z_1|\overrightarrow{z_1}), (z_2,z_2)) \simeq 17 + 6.6x$. Note that $y$ will be at least $\frac{1}{2}$ since a tit-for-tat player will always respond with cooperation to $(z_1,z_2)$. But then $\forall x > 0.12$ this inequality is satisfied. But then whenever $m \leq 19$ beliefs will always lie in the relevant intervals. ■

**Proof of Claim 4**

**Proof.** First note that absorbing states with full defection exist for all $\sigma$. Obviously in these states all agents will have the same average payoffs. Furthermore whenever $\sigma > \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}$ or whenever $3\lambda - \theta < 0$, all absorbing states will be characterized by full defection. Note also that myopic types will always choose defection since it is a dominant strategy. If $\sigma \leq \frac{3\lambda-\theta-3\delta}{3\lambda-\theta}$ types $k_2$ will find it always optimal to cooperate after the empty history (given all beliefs $\mu(z_1|\varnothing, k_2) = 1; \mu(z_1|\overrightarrow{z_1}, k_2) \geq \frac{2}{3}; \mu(z_1|\varnothing, k_1) = \mu(z_1|\overrightarrow{z_1}, k_1) = 0$). But then given that $k_2$ types cooperate in the first three and defect in the fourth round, $k_1$ types will make higher expected payoffs whenever

$$\begin{aligned} \Pi^e(k_1) &\geq \Pi^e(k_2) \Leftrightarrow \\ \sigma\delta + (1-\sigma)\theta + 3\delta &\geq (1-\sigma)[3\lambda + \delta] + \sigma 3\delta \Leftrightarrow \\ \sigma &\geq \frac{3\lambda - \theta - 2\delta}{3\lambda - \theta - \delta}. \end{aligned}$$

■

**Proof of Claim 5**

**Proof.** Note that whenever $\sigma > 0$ there is 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 $\mu(z_1|\varnothing) = 0$ (or at least will fall below the cooperation threshold) and they will start choosing defection at all rounds. 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. ■

**Proof of Proposition 5**

**Proof.** First note that at all (pure) absorbing states the same actions will be chosen at each given round $\tau$. But then - irrespective of $m$ - memories will be homogenous and beliefs will thus be point-beliefs no matter how large $m$ or $\rho$. The second assertion follows simply from a backward-induction argument. Remember from Lemma 2 that if a Non-Nash outcome is induced in an absorbing state it must be pareto-efficient. Such outcomes can be induced if agents anticipate a payoff loss in the future by choosing the Nash action at $t$. But since at time $T$ Nash actions will be chosen irrespective of the history (and agents have the corresponding round $T-$beliefs), there will be no incentive to choose the efficient action at $T-1$ etc..The last assertion follows from the following argument. Assume that at $T-1$ agents believe that if they choose the efficient action at $T-1$, the Nash action will be chosen at $T$ (with probability one), but if they do not a third action yielding even worse payoffs will be chosen. Then choosing the efficient action at $T-1$ (and thus at $T-2, T-3$...) can be optimal under certain payoff conditions and the corresponding state absorbing. ■