Learning models in interdependence situations

Learning models in interdependence situations

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-2462-4

Printed by CPI Wöhrmann Print Service

Cover designed by Coen van der Horst

Learning models in interdependence situations

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus, prof.dr.ir. C.J. van Duijn, voor een commissie aangewezen door het College voor Promoties in het openbaar te verdedigen op donderdag 14 april 2011 om 14.00 uur

door

Wouter van der Horst, bijgenaamd Linders

geboren te Oss

Dit proefschrift is goedgekeurd door de promotor:

prof.dr. C.C.P. Snijders

Copromotor:
dr. M.A.L.M. van Assen

Contents

1 Learning models in interdependence situations

Interesting forms of behavior in interdependent situations can be found throughout the history of mankind. A particularly remarkable example is that of soldiers engaged in Trench Warfare during World War I. Picture two battalions facing each other in France and Belgium across one hundred to four hundred meters of no-man's-land along an eight-hundred-kilometer line. The fundamental strategy of Trench Warfare was to defend one's own position while trying to achieve a break-through into the enemy's trench. Being inside the trench was dangerous because of the constant threat of enemy fire from the opposite trench. Remarkably, over time, soldiers from both sides were found to no longer shoot at enemy soldiers even when the enemies were walking within rifle range. While at first the soldiers were in full offense against their enemies, as time progressed the soldiers learned to live-and-let-live, and created a for-them sensible equilibrium in which neither party fired until fired upon. This equilibrium state was enduring because relieved troops would provide this information to the new soldiers. This socialization allowed one unit to pick up the situation right where the other left it (Ashworth, 1980, Axelrod, 1984).

The Trench Warfare example shows two features that are crucial to this thesis. People are interdependent and can in such situations learn to adapt their behavior over time. The interdependence is obvious: when considering to attack or not, the soldiers likely consider what the possible reaction of the other soldiers will be, and this affects their own evaluation of what to do next. Learning plays an obvious role as well: over time, soldiers find out what the results of their initial actions are, how the enemy soldiers react, and can adjust their behavior accordingly. This neatly fits the definition of learning as "an observed change in behavior owing to experience" (Camerer, 2003, p. 265). The soldiers experienced the result from firing at enemy soldiers (they probably get fired at in return) and given their experiences changed their behavior to a friendlier live-and-let-live strategy. Readers familiar with Axelrod's Trench Warfare example or game theory in general recognize the underlying structure of the soldiers' interaction as a "repeated game". At every encounter, a soldier can choose between shooting at the enemy or not, which results in four possible outcomes (both shooting, both not shooting, A shooting at B, and B shooting at A). We define the game of the Trench Warfare example below.

Table 1.1 The game of Trench Warfare during World War I

	Don't shoot	Fire!
Don't shoot	R, R	S, T
Fire!	T, S	P, P

where $T > R > P > S$ (meaning that T is preferred over R, R in his turn is preferred over P, and P is preferred over S).

Table 1.1 shows an abstract (and well-known) representation of the interaction. In line with the literature, we use letters to represent the different outcomes: **R**eward, **S**ucker, **T**emptation, and **P**unishment. Reward is the "payoff" of the strategy combination where both players leave each other alone, resulting in relative peace (and therefore obtaining their reward). Punishment is the payoff where both players fire at each other. The payoff in the case where one player fires while the other is not is called Temptation for the attacker (who shoots without being fired upon) and Sucker for the other player. From the example, we can conclude that the best result for a player is the Temptation payoff. Reward is preferred over Punishment and Sucker is the worst result for a player. The game is famous, both in game theory and beyond, as the Prisoner's Dilemma (PD). Typically, the non-shooting action is labeled "cooperation" whereas shooting is labeled "defection".

The curious aspect of a PD game is that while mutual cooperation is preferred over mutual defection by both parties, defecting is the dominant strategy for each player. Regardless of what the other player chooses, defecting is preferred over cooperation. That is, when the other player cooperates you gain the Temptation payoff by defecting which is more than the Reward payoff by cooperating; when the other player defects you gain the Punishment payoff by defecting which is better than getting the Sucker payoff by cooperating. In both cases, you obtain a higher payoff by defecting. Hence, a player cannot obtain a higher outcome by unilaterally changing his strategy from mutual defection. This is why mutual defection is a (and in fact the only) pure Nash equilibrium in this game. The dilemma stems from the fact that although mutual cooperation would be an improvement for both players over mutual defection, in the PD case the incentive structure of the game is geared against reaching this improvement. Many dilemmas in real-life can be represented by a Prisoner's Dilemma game (Binmore, 1992).

So far we have discussed a PD game being played just once. However, in our Trench Warfare example (and in many practical applications), PD games are played repeatedly. The same small units of soldiers face each other in immobile sectors for an extended period of time. This changes the game from a one-shot PD game to a repeated PD game, as is visualized in Table 1.2.

Table 1.2 The repeated game of Trench Warfare during World War I

	Don’t shoot	Fire!	
Don’t shoot	R, R	S, T	Round = 1
Fire!	T, S	P, P	

↓

	Don’t shoot	Fire!	
Don’t shoot	R, R	S, T	Round = 2
Fire!	T, S	P, P	

↓

	Don’t shoot	Fire!	
Don’t shoot	R, R	S, T	Round = 3
Fire!	T, S	P, P	

↓

⋮

where $T > R > P > S$.

Let us consider what game theoretical rationality prescribes in this case. In Round 1, both players choose to shoot or not and obtain a payoff. Then, they enter a new round and get to choose again. When the repeated PD game is played exactly N times and N is known to both players, the game theoretic equilibrium is similar to the equilibrium in a single-shot game: both players should defect all N times the game is played. This can be easily seen by starting to think about what could happen in the last round of the game. In round N, optimal behavior would be to defect, following the same logic as in the single-shot PD. In round $N - 1$, knowing that mutual defection will be optimal in round N, it is likewise optimal to choose defection. Continuing this argument all the way back until one reaches round 1 shows that mutual defection throughout is the only equilibrium in this game. Matters are different when the game is repeated but now for an unknown number of rounds. For instance, one can show that when there is a given probability (smaller than 1) to play the next round and this probability is sufficiently large, then equilibrium behavior can (but need not) lead to continuous mutual cooperation. The crucial argument is that both players playing conditionally cooperative strategies (“I will cooperate as long as the other one does so too”) can be shown to be in equilibrium in this case (Binmore, 1992). That mutual cooperation *can* occur is comforting, but unfortunately matters are not that straightforward. As the so-called Folk Theorem shows, in a repeated PD game there are infinitely many Nash equilibrium outcomes (for a proof, see e.g. Binmore, 1992). And not all of these equilibria contain high percentages of cooperative behavior. So mutual cooperation is possible in the sense that it is not completely at odds with game theoretical arguments, but that is as far as it goes.

This leads to three reasons why just using game-theoretical arguments in repeated games in general can be improved upon. First, game theory often does not give any definite predictions in repeated games. The indefinitely repeated PD game is one case in point, and there are many more examples. Second, as soon as the game gets even a bit more complicated (think about adding more options for each player, more players, or information that is available to some but not all players), calculating the game-theoretic equilibrium becomes much more difficult and/or more dependent on additional, often unrealistic assumptions. When finding the game-theoretic equilibrium is hard for trained mathematicians, it is less likely that regular people will be able to play it, which makes that game theory alone is less likely to predict the behavior of regular people well. Third, as can be seen from our example, game theory emphasizes players being (mainly) "forward looking": players think many steps ahead and forecast the expected consequences of their future behavior and that of the opponent. However, experiments have shown that players typically cannot think more than a few steps ahead (Camerer, 2003, Poundstone, 1992). In fact, players are typically found to be "backward looking" as well: they learn from their past behavior and adapt accordingly.

This has inspired the development of learning models to understand the behavior of players in interdependence situations (e.g., Roth & Erev, 1995). A learning model determines the probability of a future choice as a (usually relatively simple) function of historical information and other characteristics of the situation. Learning models assume less cognitive effort on the part of the player, assume less far-sightedness, and they also allow for updating behavior in ways that need not be rational when using game theoretical tools. This is a strong point, because experimental data strongly suggests that learning behavior of humans is far from the rational ideal. Moreover, learning models have provided substantially better descriptions of behavior in interdependence situations than standard game theoretical models, although the debate is still open on that issue (e.g., Camerer, 2003, Camerer & Ho, 1999, Erev & Roth, 1998, Roth & Erev, 1995).

The application of learning models to interdependent situations lies at the heart of this thesis. What are the predictions of learning models in interdependent situations, which kinds of learning lead to which kinds of predictions, and how can we recognize different kinds of learners? Finally, we also take a closer look at another, but related topic. Rather than examining learning in games, we examine how a player's social preferences can influence his behavior in these games. A player is said to have social preferences if the evaluation of that player's outcome also depends on the other player's outcome.

1.1 Learning models

Literature shows different learning models, corresponding to different approaches to learning (Becker, 1976). For an overview of different types of learning in relation to interdependent situations, we refer to Camerer (2003). Here, we restrict ourselves to the two most commonly used classes: reinforcement and belief learning models.

Reinforcement learning assumes that successful past actions have a higher probability to be played in the future. This approach therefore assumes that players are backward looking. For example, pigeons that peck at levers in a lab can learn that by doing so they obtain food (Skinner, 1938). Since they consider this to be a favorable outcome, they will be more likely to peck the lever again in the future. Reinforcement learning models have been popular in psychology (Bush & Mosteller, 1951), sociology (Flache & Macy, 2001, Macy & Flache, 2002), and economics (Camerer, 2003, Roth & Erev, 1995). Although predictions of reinforcement learning models clearly outperform predictions based on game-theoretical models, predictions are still not convincing in some types of games (Camerer, 2003).[1] For instance, the speed of learning predicted by these models is often too slow compared to human learning (Camerer, 2003, Camerer & Ho, 1999). One reason for this might be that reinforcement learning models assume that the behavior of a player is not affected by foregone payoffs: payoffs the player would have earned after choosing other strategies.

Foregone payoffs are assumed to have a large effect on behavior in what is called "belief learning". Belief learning assumes that players have beliefs about which action the opponent(s) will choose and that players determine their own choice of action by finding the action with the highest payoff given the beliefs about the actions of others. Hence, in belief learning players are actually looking forward, but only one step. In mathematical terms, their beliefs are based on the probability distribution of the available actions of the other players. The player then chooses the action with the highest expected value based upon this belief. Belief learning models have been more successful than reinforcement learning models in predicting behavior in some games, where reinforcement learning models outperform in others (Camerer, 2003, Cheung & Friedman, 1997).

To study both the information gained by foregone payoffs as well as information gained by the player's choice in the past, Camerer & Ho (1999) created a hybrid of reinforcement and belief models which also uses both types of information. This Experience-Weighted Attraction (EWA) model contains a parameter denoting the extent to which foregone payoffs reinforce unchosen strategies. Differently put: the parameter describes how much a player is a reinforcement learner or a belief learner. Therefore it makes it a informative model to determine the exact type of learning.

One of the crucial questions is whether it is possible to determine whether someone is a belief learner or a reinforcement learner by looking at his or her behavior (the choices in the game) alone. This is problematic, because we know from the literature that in experimental data of some games the two types of learning can hardly be distinguished. For example, Feltovich (2000) concluded for the games in his study that: "While quite different in rationale and in mathematical specification, the two models [belief and reinforcement learning] yield qualitatively similar patterns of behavior." Likewise, Hopkins (2002) derived analytically to what extent two specific reinforcement and belief learning models yield similar predictions in games with strictly positive outcomes. He concluded that in the special case where no forgetting of behavior in previous rounds takes place, specific kinds of reinforcement learning and belief learning will lead to the same behavior. Hopkins

[1] In defense of the game-theoretical predictions one could of course argue that more precise assumptions about the underlying game would improve the game-theoretical predictions, or that over time humans will converge to the game-theoretical predictions (either because they learn the game-theoretical optimum by playing, or because humans in the end will evolve in the game-theoretic direction). We do not want to take part in this discussion and instead restrict ourselves to analyzing how far one can get using standard learning models.

then concluded that the main identifiable difference between the two models was the speed of learning. Salmon (2001) demonstrated that commonly used econometric techniques were unable to accurately discriminate between belief, reinforcement, and EWA learning in the type of games he studied. Salmon simulated data from a given model and then estimated different models to the data to check whether indeed the original model would provide the best fit. While he concluded that it was typically difficult to identify the process that generated the data, the problem appeared most severe in games with only two strategies per player. Salmon (2001) states that: "overcoming this difficulty [discerning between different models] on a purely econometric basis will be difficult in the least". So on the one hand it makes sense that when one wants to understand learning in interdependent situations, starting with relatively small games is the obvious way to go. However, the literature suggests that especially in games where players have only two actions to choose from, the choices of the players do not carry enough information to reliably assess the underlying learning model. As we show in the following chapters, this general conclusion is too pessimistic or at least needs some conditioning. We show that it is possible to find differences between belief learners and reinforcement learners in relatively small and simple repeated games if you know where to look.

Let us return to our example, the PD game, to see whether we can reproduce this problem of not being able to discern between belief and reinforcement learning. In the PD a reinforcement learner will tend to choose an action that is positively reinforced. Whether reinforcement occurs depends on his payoffs, which are determined by his own and the other player's choice. If the obtained payoff is low, then a reinforcement learner would become more inclined to change from cooperation to defection or vice versa (the soldiers would be inclined to stop their truce when the enemy starts firing again). Now suppose that both players cooperate at some point in the repeated game and receive positive payoffs. When both players are reinforcement learners, cooperation would be considered by both to be a positive result and both players would therefore become more inclined to retry cooperation. Reinforcement learners could therefore easily end up playing mutual cooperation over and over again in a PD. However, if both players started out with defection that yielded positive payoffs (although, by definition, less than they could get under mutual cooperation), both players will be more inclined to retry defection. Hence reinforcement learning can predict both mutual cooperation and mutual defection in case of positive payoffs (reinforcement). For belief learners the prediction is completely different. Suppose that two belief learners end up in mutual cooperation at some point in the game. A belief learner would notice that he could have gotten a better payoff if he had chosen the other action. Hence, the belief learner, on his next move, would become more likely to choose defection. Over time, belief learners will eventually change their behavior towards defecting. Belief learners would never end up in playing mutual cooperation over and over again.[2]

A surprising result: in the literature there are several suggestions that it is hard to disentangle belief learning and reinforcement learning in interdependent games, but when we compare the two on the most well-known repeated game, we immediately see a clear difference! How can this be? This leads to the main question of this thesis:

[2] We postpone a formal derivation of this result to Chapter 2

> *Can we distinguish between different types of EWA-based learning, with reinforcement and belief learning as special cases, in repeated* 2×2 *games?*

To answer this question, we start in Chapter 2 with an analysis of the theoretical implications of the EWA model (with reinforcement and belief learning as special cases). We discuss the underlying assumptions of the model to get a better insight into reinforcement and belief learning. The results allow us to find repeated games in which we expect different behavior between reinforcement and belief learners. Although the results derived in Chapter 2 are general in the sense that they apply to a large class of games, we will demonstrate the crucial differences between belief learning and reinforcement learning with three simple types of 2×2 games. First, we use games of the Prisoner's Dilemma type, as the Trench Warfare example above. Our second set of games are those with a pure, Pareto-optimal Nash equilibrium (NE games).[3] Finally, we consider games with one mixed-strategy Nash equilibrium (ME games). In ME games, players maximize their expected payoff of the game by choosing each of their available actions with a certain probability, rather than choosing one strategy over the other (Tsebelis, 1989). Chapter 2 concludes that in all three cases, there are differences between belief and reinforcement learners.

The theoretical result of Chapter 2 alone is not sufficient. These results merely suggest that the learning models can be distinguished *after a sufficient number of rounds have been played*, but it is not clear how large that number needs to be and in general how likely it is that we find these differences. A logical next step, therefore, is to simulate play for reinforcement and belief learning models. So we simulate play in the same three sets of games for 10, 30, and 150 rounds of play. Then we compare the data of the reinforcement and the belief learners and try to see where we can find different play for belief versus reinforcement learning. The most obvious characteristic to consider are streaks of the same behavior in the data (mimicking a long period of live-and-let-live with our soldiers). The question is whether there are different streaks in the data of reinforcement learners when compared to the data of belief learners. The results of this analysis can be found in Chapter 3. In addition, we consider several other important characteristics of play, such as how often players change between their two strategies, how often different outcome combinations occur, and how soon players end up in a repeated behavior pattern. As Chapter 3 will show, for various values of parameters of the EWA model we can indeed find learning differences, even after only ten rounds of play. Not only differences in the occurrence of streaks, but other characteristics of the game play give us instruments to tell the differences between the two learning models. In some cases these other characteristics confirm the differences between reinforcement and belief learning even more strongly, but in some cases these differences can also be used to differentiate between the two learning models when streaks cannot tell them apart.

Chapter 2 and Chapter 3 enable researchers to construct games that can be used to differentiate reinforcement from belief learners, based on the assumption that players are of either class. In Chapter 4 we relax this assumption using the Experience-Weighted Attraction (EWA) model. The extent to which foregone payoffs are taken into account is expressed by the EWA parameter δ. For $\delta = 0$ we

[3] These are games in which the equilibrium yields in fact also the best payoff to both players, as opposed to the PD in which the equilibrium yields inferior payoffs.

have (a form of) reinforcement learning and for $\delta = 1$ we have (a form of) belief learning. Values for δ between zero and one represent a learner who incorporates both belief and reinforcement learning to a given extent. In Chapter 4 we simulate data generated by the EWA learning model for a given set of parameters and then try to re-estimate the parameters of the model. Is it possible to retrieve the original set of parameters and if so, how many rounds of play are necessary to be able to re-estimate the original set? In Chapter 4 we must find low rates of convergence of the estimation algorithm, and if the algorithm converges then biased estimates of the parameters are obtained most of the time. Hence, we must conclude that re-estimating the exact parameters in a quantitative manner is very difficult in most experimental setups, but qualitatively (as done in Chapter 3) we can find patterns that pinpoint in the direction of either belief or reinforcement learning

Finally, our last Chapter is related but different. We started out by introducing learning models because there are certain disadvantages to the strict game-theoretical approach in interdependence situations. In fact, game theory (in the strict sense of the term) has another disadvantage. It is widely accepted that the payoff or utility a player obtains might not only depend on his own monetary outcome, but also on the monetary outcomes of the other actors in the game (e.g. Bolton & Ockenfels, 2000, Camerer, 2003, Engelmann & Strobel, 2004, Fehr & Gintis, 2007, Rabin, 1993, 2006). The notion that players have social preferences (with corresponding feelings like "envy" and "spite") that could or perhaps should somehow be reflected in the formal decision making models is widespread (Snijders & Raub, 1996). These social preferences change and complicate the analysis of behavior in interdependent situations, especially in mixed-strategy equilibrium games without a pure Nash equilibrium (where it is not even clear what to play in a one-shot game). In Chapter 5 we look at the effect of introducing social preferences in mixed-strategy Nash equilibrium games. The idea is that, if evolution bestowed us with social preferences, then should they not be beneficial to those having social preferences? In terms of the game, the question we tackle is whether the expected monetary payoff for a player increases after introducing social preferences and if so, under what circumstances.

Chapter 5 concludes that introducing social preferences of a player on average actually increases the expected payoff of a game for that player, although the effect seems small. Moreover, the effect is not uniformly positive in all situations. The larger the difference in status (as in hierarchical relationships), the smaller the increase of the expected payoff of the game. The increase in expected payoff is highest when the "status" of the two players is the same (say, among colleagues or friends). In addition, we find that the increase of the expected payoff is largest in games where players have a high risk alternative (either a very high or a very low payoff, depending on what the other player does) and a low risk alternative (about the same payoff). A final and important result is that the effects of envy and spite are analogous. That is, the effect of a player's envy on his payoffs in a situation is equal to the effect of spite in another situation. Hence spite and envy are different sides of the same coin in mixed-strategy equilibrium games.

2 Analyzing behavior implied by EWA learning: an emphasis on distinguishing reinforcement from belief learning

An important issue in the field of learning is to what extent one can distinguish between behavior resulting from either belief or reinforcement learning. Previous research suggests that it is difficult or even impossible to distinguish belief from reinforcement learning: belief and reinforcement models often fit the empirical data equally well. However, previous research has been confined to specific games in specific settings. In the present study we derive predictions for behavior in games using the EWA learning model (e.g., Camerer & Ho, 1999), a model that includes belief learning and a specific type of reinforcement learning as special cases. We conclude that belief and reinforcement learning *can* be distinguished, even in 2×2 games. Maximum differentiation in behavior resulting from either belief or reinforcement learning is obtained in games with pure Nash equilibria with negative payoffs and at least one other strategy combination with only positive payoffs. Our results help researchers to identify games in which belief and reinforcement learning can be discerned easily.

2.1 Introduction

Many approaches to learning in games fall into one of two broad classes: reinforcement and belief learning models. Reinforcement learning assumes that successful past actions have a higher probability to be played in the future. Belief learning assumes that players have beliefs about which action the opponent(s) will choose and that players determine their own choice of action by finding the action with the highest payoff given the beliefs about the actions of others. Belief learning and (a specific type of) reinforcement learning are special cases of a hybrid learning model called Experience Weighted Attraction (EWA) (Camerer & Ho, 1999). The

EWA model contains a parameter δ denoting the extent to which foregone payoffs reinforce unchosen strategies. For $\delta = 1$ we obtain belief learning. For $\delta = 0$ we obtain a class of reinforcement learning models with a fixed reference point.[1] One of the main questions in the analysis of learning models is which model best describes the actual learning behavior of people. An important question is therefore under which conditions it is possible to discern between different kinds of learning. Answering this question is the focus of the present study.

Whether a player adopts belief or reinforcement learning is, ultimately, an empirical question. However, in some games the two types of learning can hardly be distinguished because they predict similar behavior. The problem we address is to identify conditions in which behavior of players adopting belief learning will be fundamentally different from those adopting reinforcement learning. As we will argue, many studies examined learning in conditions in which belief and reinforcement learning can hardly be distinguished. Our identification of conditions will help experimenters design experiments in which both learning types can be distinguished.

A substantial number of empirical studies were carried out to determine which model best describes the learning behavior of people in different interdependence situations. Camerer (2003) summarizes their main findings. He concluded, among other things, that belief learning generally fits behavior better than reinforcement learning in coordination games and in some other classes of games (such as market games and dominance-solvable games), whereas in games with mixed strategy Nash equilibria both models predict with about the same accuracy. However, Camerer (1999, pp. 304) added that "it is difficult to draw firm conclusions across studies because the games and details of model implementations differ."

Some previous studies explicitly state that it is difficult to determine the underlying process (either reinforcement learning, belief learning, or something else) that generated the data for several games. For example, Feltovich (2000, pp. 637) concluded for the multistage asymmetric-information games in his study that:"While quite different in rationale and in mathematical specification, the two models [belief and reinforcement learning] yield qualitatively similar patterns of behavior." Hopkins (2002) derived analytically to what extent two specific learning and belief models yield similar predictions in 2-person normal form games with strictly positive outcomes. He concluded that in the special case where no forgetting of behavior in previous rounds takes place, cumulative reinforcement learning and fictitious belief learning will have the same asymptotic behavior. Hopkins concluded that the main identifiable difference between the two models was speed: stochastic fictitious play results in faster learning.

Salmon (2001) demonstrated that commonly used econometric techniques were unable to accurately discriminate between belief, reinforcement, and EWA learning in the constant-sum normal form games with non-negative payoffs that he studied. He simulated data from a given model and then estimated different models to the data to check whether indeed the original model would provide the best fit. While he concluded that it was typically difficult to identify the process that generated the

[1] Some reinforcement models assume adjustable reference points (e.g., Erev & Roth, 1998 and Macy & Flache, 2002). Given that we use the EWA model our results do not directly apply to reinforcement learning with an adjustable reference point.

data, the problem appeared most severe in 2×2 games, less severe in 4×4 games, and least severe in 6×6 games. Salmon (2001, pp. 1626) states that "overcoming this difficulty [discerning between different models] on a purely econometric basis will be difficult in the least", and he suggested to design experiments in such a way that more than simply the observed choices can be assessed based on the general notion that especially in 2×2 games the choices of the players do not carry enough information to reliable assess the underlying learning model.

The conclusion that one might draw based on this research could well be that different learning models are difficult or even impossible to discern. However, these conclusions would be based on an empirical comparison of behavior in specific games (with non-negative payoffs) (Camerer, 2003; Feltovich, 2000) or on a theoretic comparison of learning models for specific games (with non-negative payoffs) (Feltovich, 2000; Hopkins, 2002; Salmon, 2001). As Salmon (2001, pp. 1625) suggested, the difficulties he observed in distinguishing learning models might only exist in constant-sum normal form games. In general, given that the results from these studies are confined to specific settings, it is not clear to what extent the difficulty to distinguish reinforcement from belief learning can be generalized to *any finite game*. In this paper we analyze finite games analytically using the EWA model to find conditions under which one can discern belief and reinforcement learning. This analysis is a necessary first step for setting up experiments in which one can determine on the basis of agents' behavior whether learning occurs according to belief or reinforcement learning.

We formally derive predictions on behavior as implied by EWA learning for any finite game, that is, any game with a finite number of players choosing from a finite number of strategies. Our focus is on *stable behavior*. We define stability as a stochastic variant of pure Nash equilibrium: there is a large probability (typically close to 1) that all players will make the same choice in round $t + 1$ as in t. We analyze which kinds of stable behavior can be predicted by EWA learning, with special attention to the comparison of stable behavior predicted by reinforcement and belief learning.

Contrary to the general gist of previous research, our main conclusion is that belief and reinforcement learning *can* yield very different predictions of stable behavior in finite games. Even in simple 2×2 games, reinforcement and belief learning can lead to completely different predictions about which kinds of behavior can be stable. While only pure Nash equilibria can be stable under belief learning, all strategy combinations can be stable in reinforcement learning if such a combination provides strictly positive payoffs to all players in the game. Hence, maximum differentiation in predictions of the two types of learning is obtained in games with pure Nash equilibria with negative payoffs, and at least one other strategy combination that yields positive payoffs to all players.

The goal of our analyses is twofold. First, our short term goal is to specify conditions under which belief and reinforcement learning can be distinguished. These conditions refer to the type of game and the (order in the) payoffs in these games. Our second and longer term goal is to generate recommendations with respect to experimental conditions (type of game, specific game payoffs, necessary number of rounds, etc.) under which discerning between belief and reinforcement learning is most likely and not difficult.

The setup of the paper is as follows. In Section 2 we describe the EWA learning model and introduce some notation. The analytical results of predictions by the EWA model on stable behavior are described in Section 3. We first derive results for 2×2 games, then we consider analytical results for any finite game. We end with conclusions and a discussion in Section 4.

2.2 The EWA learning model

The EWA learning model has been used to model subjects' behavior in many applications. It was developed by and first applied in Camerer & Ho (1999); see Camerer (2003: Chapter 6) for a description of applications of the EWA model. The notation we use is based on Camerer & Ho (1999). Players are indexed by i $(= 1, \ldots, n)$ and the strategy space S_i consists of m_i discrete choices, that is, $S_i = \{s_i^1, s_i^2, \ldots, s_i^{m_i-1}, s_i^{m_i}\}$. Furthermore, $S = S_1 \times \ldots \times S_n$ is the strategy space of the game. Then, $s_i \in S_i$ denotes a pure strategy of player i and $s = (s_1, \ldots, s_n) \in S$ is a pure strategy combination consisting of n strategies, one for each player; $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ is a strategy combination of all players except i. In our learning context a "strategy" simply refers to a choice in the constituent game, hence a strategy s leads to a particular outcome for all i. The term "strategy" does not refer to a general prescription of how a player will behave under all possible conditions, as is standard in game theoretical models. The outcome or scalar-valued payoff function of player i is denoted by $\pi_i(s_i, s_{-i})$. Denote the actual strategy chosen by player i in period t by $s_i(t)$, and the strategy chosen by all other players by $s_{-i}(t)$. Denote player i's payoff in a period t by $\pi_i(s_i(t), s_{-i}(t))$.

The core of the EWA model consists of two variables that are updated after each round. The first variable is $A_i^j(t)$, player i's attraction (also called "propensity", by, e.g., Erev & Roth, 1998) for strategy s_i^j *after* period t has taken place. The second variable is $N(t)$, which is related to the extent to which previous outcomes play a role (see below). These two variables begin with certain prior values at $t = 0$. These values at $t = 0$ can be thought of as reflecting pregame experience. Updating of these two variables is governed by two rules. The first one is

$$A_i^j(t) = \frac{\varphi N(t-1) A_i^j(t-1) + [\delta + (1-\delta) I(s_i^j, s_i(t))] \pi_i(s_i^j, s_{-i}(t))}{N(t)}, \tag{2.1}$$

where $\varphi \in [0, 1]$ is a recall parameter (or 'discount factor' or 'decay rate'), which depreciates the attraction value of the given strategy in the previous round. Furthermore, $\delta \in [0, 1]$ is the parameter that determines to what extent the foregone payoffs are taken into account. The $I(x, y)$ is an indicator function equal to 1 when $x = y$ and 0 if not. The second variable, $N(t)$, is updated by

$$N(t) = \rho N(t-1) + 1, \qquad t \geq 1, \tag{2.2}$$

where ρ is a depreciation rate or retrospective discount factor that measures the fractional impact of previous experience, compared to a new period. To guarantee that experience increases over time $(N(t) > N(t-1))$ it is assumed that $N(0) < \frac{1}{1-\rho}$.

Another restriction is that $\rho \in [0, \varphi]$.

The parameters δ, $N(0)$, φ, ρ, and $A_i^j(0)$ in the EWA model have psychological interpretations. The δ expresses to what extent foregone payoffs matter in comparison to currently obtained payoffs. For $\delta = 0$ only actual payoffs matter, as in reinforcement learning. If $\delta = 1$ actual payoffs matter as much as foregone payoffs, as in belief learning. One can interpret δ times the average foregone payoff in each period as a kind of aspiration level to which the actual payoff is compared in each period. With $N(0)$ one captures the idea that players have some prior familiarity with the game (Salmon, 2001, pp. 1609). The parameter φ is a discount factor of the past. It denotes the relative contribution of previous play to the attraction as compared to recent play. If $\varphi = 1$ previous play matters as much as recent play and is recalled perfectly, if $\varphi = 0$ then previous play does not matter. The ρ parameter symbolizes the importance of prior experiences relative to new experiences, allowing attractions to grow faster than a given average, but slower than a cumulative total (Camerer & Ho, 1999, pp. 839). If $\rho = 0$ the most recent experience gets equal weight relative to prior experience, and a strategy's attraction accumulates over time according to the reinforcement rule $[\delta + (1-\delta)I(s_i^j, s_i(t))]\pi_i(s_i^j, s_{-i}(t))$. At the other extreme, if $\rho = \varphi$ then a strategy's attraction is a weighted average of the payoffs one can obtain with that strategy, wich means that a strategy's attraction is bounded by the minimum and maximum payoff one can obtain by using that strategy. If $0 < \rho < \varphi$ then EWA learning models that players use something in between "lifetime performance" and "average performance" to evaluate strategies (Camerer & Ho, 1999, pp. 839). Finally, $A_i^j(0)$ can be interpreted as the initial preference for a given strategy.

Different known learning models can be obtained by using different values for the parameters. If $\rho = \varphi$ and $\delta = 0$ then EWA reduces to "averaged" reinforcement learning (Camerer & Ho, 1999). The averaged reinforcement learning specification of EWA is analogous to the model of Mookerjee & Sopher (1997). When instead $N(0) = 1$ and $\rho = 0$ and $\delta = 0$, EWA equals the "cumulative" reinforcement model of Erev & Roth (1998). Finally, if $\delta = 1$ as well as $\rho = \varphi$ then EWA reduces to the "weighted fictitious play" belief learning model of Fudenberg & Levine (1998).

Another parameter in the EWA model, λ, determines how the strategies' attractions are transformed into probabilities. The probability that player i plays strategy s_i^j at time $t+1$ in EWA learning is a function of the strategies' attractions at time t, using logit transformations:

$$P_i^j(t+1) = \frac{\exp^{\lambda A_i^j(t)}}{\sum_{k=1}^{m_i} \exp^{\lambda A_i^k(t)}} = \frac{1}{1 + \sum_{k=1, k \neq j}^{m_i} \exp^{\lambda (A_i^k(t) - A_i^j(t))}}, \tag{2.3}$$

where $\lambda \geq 0$ is called the payoff sensitivity parameter (e.g. Camerer & Ho, 1999). As one can see, for $\lambda = 0$ the probabilities for all strategies are equal, regardless of the values of the attractions, whereas for a large value of λ the strategy with the highest attraction is chosen almost certainly.

Besides the above-mentioned logit link between attractions and probabilities, some researchers have used the probit or power link function (see Camerer, 2003, pp. 834-836, for a discussion). The probit link replaces the cumulative logistic in

(2.3) by the cumulative normal distribution function. The power link assumes that the probability of choosing a strategy is equal to the ratio of its attraction raised to the power of λ, divided by the sum of attractions raised to the power of λ. That is:

$$P_i^j(t+1) = \frac{A_i^j(t)^{\lambda}}{\sum_{k=1}^{m_i} A_i^k(t)^{\lambda}}, \tag{2.4}$$

Although the differences between these transformations might seem relatively unimportant, they do have substantial implications (e.g., see also Flache & Macy, 2001). For instance, the logit form is invariant to adding a constant to all attractions, whereas the power form is invariant to multiplying all attractions with a constant. Moreover, if the constituent game has only positive payoffs, the speed of convergence does not depend on the number of previous periods played when using the logit link, but it decreases in the number of previous periods when using the power link. A typical characteristic of the logit link is that it allows negative attractions. Moreover, Camerer & Ho (1999, pp. 836) reported that, whereas previous studies showed roughly equal fits of logit and power link models, the logit link yields better fits to their data than the power form. Although the two forms provide roughly equal fits for some games, both forms can yield fundamentally different predictions in games with only a mixed strategy Nash equilibrium.[2]

2.3 Analytical results

As it turns out, the results for 2×2 games can be generalized to any finite game in a straightforward way. Nevertheless, for ease of exposition we derive the implications of the EWA model for learning behavior in two subsections: one on 2×2 games and one on any finite game. All results concern the conditions for pure strategy combinations s to be stable. We define $P(s(t))$ as the probability of players playing a strategy combination s in round t. Then, a strategy combination s is defined to be stable if

$$P(s(t)) \geq 1 - \varepsilon \tag{2.5}$$

after repeated play of s, for some $\varepsilon < 0.5$. For small ε,say $\varepsilon = 0.01$, (5) can be considered a stochastic variant of pure Nash equilibrium: with a probability of at least 0.99 s is played in the next round by all players. Our stability concept is similar to the concept of *local stability* as defined by Fudenberg and Kreps (1995, pp. 345) in the context of fictitious belief learning of mixed strategy equilibria.

We now turn to how the concept of stability can be of use for our purposes. Stability of s has two implications. First, it implies that the outcome of s is attractive for all i. In Section 4 we show that repeatedly playing an s with an attractive

[2] To see this, consider a mixed strategy as stable if (and only if) the probability distribution over the strategy space for each individual does not change (by more than some small epsilon) over time. It is easy to show that both the logit and probit link will not lead to stable mixed strategies (except in a number of trivial cases). This result follows from the fact that $P_i^j(t+1)$ is a function of the difference in attractions, and this difference does not converge. When the power link is used, stable mixed strategies are possible: probabilities may converge if $\varphi = 1$.

outcome for all players leads to $P(s(t)) > 0.5$, whereas if s contains an unattractive outcome for at least one i repeated play of s leads to $P(s(t)) < 0.5$. Hence an s that satisfies (5) at time t but not after repeated play is not stable. Second, stability is a sufficient condition for s to be played for a large number of rounds in a row with high probability. And if s is not stable, then it is unlikely to observe s to be played for a large number of rounds in a row. In the present paper stability is used to derive lemmas and a theorem on conditions for s to be stable, (i.e., on conditions for s to be played in many rounds in a row) for belief and reinforcement learning under the EWA model. In this way, stability helps to identify those s with a high probability to be played again in the subsequent rounds. Hence, streaks of s in a repeated game for an s that is stable under reinforcement but not under belief learning is empirical evidence in favor of reinforcement and against belief learning. In Section 4 we identify games for which streaks of play provide favoring evidence for one, and impeding evidence for the other type of learning.

We end with two clarifying remarks on stability. First, note that the potential stability of an s does not tell us much about whether and when stability in s will actually occur. Second, our stability concept is weaker than "convergence" in the mathematical sense of the word; convergence implies stability, but stability does not imply convergence. For example, at a certain round t a stable s might satisfy (2.5), but random shocks can lead the player to choose another strategy s^* at $t+1$, after which $P((s(t))$ might decrease substantially and play might never return to s again.

2.3.1 Results on 2×2 games

For the analysis it is fruitful to work out the expression in the exponent of the right hand of (2.3). Note that we are only considering the logit rule for probabilities. From now on when we use strategy k, we assume that $k \neq j$. Let $n = 2$ and $m_i = 2$ for $i \in 1, 2$, then the expression in the exponent of $P_i^1(t+1)$ can be written as

$$\begin{aligned}
&\lambda(A_i^2(t) - A_i^1(t)) = \\
&\frac{\lambda}{N(t)}\Big[\delta\{\pi_i(s_i^2, s_{-i}(t)) - \pi_i(s_i^1, s_{-i}(t))\} + \\
&\quad (1-\delta)\{I(s_i^2, s_i(t))\pi_i(s_i^2, s_{-i}(t)) - I(s_i^1, s_i(t))\pi_i(s_i^1, s_{-i}(t))\} + \\
&\quad \sum_{u=1}^{t-1} \varphi^{t-u}\big[\delta\{\pi(s_i^2, s_{-i}(u)) - \pi(s_i^1, s_{-i}(u))\} + \\
&\qquad\qquad (1-\delta)\{I(s_i^2, s_i(u))\pi_i(s_i^2, s_{-i}(u)) - I(s_i^1, s_i(u))\pi_i(s_i^1, s_{-i}(u))\}\big] + \\
&\quad \varphi^t N(0)\{A_i^2(0) - A_i^1(0)\}\Big]. \qquad (2.6)
\end{aligned}$$

On the right hand side in (2.6) we see terms representing the effects of choices at time t (first two lines), the sum of the effects of previous trials (third and fourth line), and the effect of the initial conditions (last line). It is also useful to derive an expression for the condition under which $P_i^j(t+1) > P_i^j(t)$ or $A_i^2(t) - A_i^1(t) < A_i^2(t-1) - A_i^1(t-1)$. This expression is

$$[\delta + (1-\delta)I(s_i^1, s_i(t))]\pi_i(s_i^1, s_{-i}(t)) - [\delta + (1-\delta)I(s_i^2, s_i(t))]\pi_i(s_i^2, s_{-i}(t)) > \\ [1-(\varphi-\rho)N(t-1)](A_i^1(t-1) - A_i^2(t-1)). \quad (2.7)$$

where

$$N(t-1) = \rho^{t-1}N(0) + \frac{1-\rho^{t-1}}{1-\rho}, \quad (2.8)$$

for $\rho < 1$, and $N(t-1) = N(0) + t - 1$ for $\rho = 1$. The left part of (2.7) shows the difference in reinforcement in favor of strategy 1. The right part shows a depreciation of that difference one round earlier. Now consider the following two extreme cases. If $\varphi = \rho$ (averaged reinforcement learning) then $P_i^1(t+1) > P_i^1(t)$ if the difference in reinforcement at $t+1$ is larger than the difference in attractions at t. If both $\rho = 0$ and $N(0) = 1$ (cumulative reinforcement learning) then the probability to play strategy 1 is increasing if the difference in reinforcement is larger than the difference between the actual and the remembered attraction. Note that Eq. (2.7) implies that a probability to play a certain strategy can increase, even when its reinforcement is lower than that of another strategy, as long as this difference in current reinforcements is compensated for by the depreciated difference in reinforcement.

Let us now turn to stable strategy combinations. A necessary and sufficient condition for s to be stable is if (2.5) holds after infinitely repeated play of s. Two cases have to be distinguished: $\varphi = 1$, and $\varphi < 1$, that is, perfect recall versus discounting of payoffs. We assume that there are no correlated strategies. Then, after infinite play of s $(= (s_1^1, s_2^1))$ in a 2 × 2 game, s can be stable if and only if

$$\lim_{t\to\infty} P_1^1(t+1)P_2^1(t+1) \geq 1 - \varepsilon. \quad (2.9)$$

Upon substituting (2.6), this is true in case of imperfect recall if

$$\frac{1}{1 + \exp\left(\frac{\lambda(1-\rho)}{1-\varphi}[\delta\pi_i(s_i^2, s_{-i}(t)) - \pi_i(s_i^1, s_{-i}(t))]\right)} \geq \sqrt{1-\varepsilon}. \quad (2.10)$$

for both players i.[3] We see that the initial attractions and the history of play before the time that s was starting to be played becomes irrelevant because $\varphi < 1$. This does not become irrelevant in the case of $\varphi = 1$ (perfect recall).

It follows from (10) that for an s to be stable under imperfect recall the fraction $\frac{\lambda(1-\rho)}{1-\varphi}$ must be large enough. Increasing λ and φ has the same effect on this fraction, the effect of ρ is opposite. Note that this fraction implies that, whatever the outcomes and the values of the other parameters, there always exist low values of λ such that no outcome is stable under any type of learning.

Assume the payoffs in the game are fixed, and that the parameter values of the model can be chosen (or fitted) freely, as in empirical applications, with the restriction that $\delta \in [0, 1]$. Then the following results directly follow from (2.10).

[3] Note that this is a stronger statement than (2.9).

Theorem 2.3.1 (continued). Strategy combination (s_i^j, s_{-i}) in a 2×2 game can not be stable in EWA learning if it yields an outcome $\pi_i(s_i^j, s_{-i}) < 0$ and $\pi_i(s_i^j, s_{-i}) < \pi_i(s_i^k, s_{-i})$. In case of imperfect recall (s_i^j, s_{-i}) in a 2×2 game cannot be stable in EWA learning if it yields an outcome $\pi_i(s_i^j, s_{-i}) < 0$ and $\pi_i(s_i^j, s_{-i}) \leq \pi_i(s_i^k, s_{-i})$. In all other cases (s_i^j, s_{-i}) can be stable.

The proof of Theorem 2.3.1 is straightforward. In the case of perfect recall, if $\pi_i(s_i^j, s_{-i}) < 0$ and $\pi_i(s_i^j, s_{-i}) < \pi_i(s_i^k, s_{-i})$, then $\delta\pi_i(s_i^k, s_{-i}) - \pi_i(s_i^j, s_{-i}) > 0$, and $P_i^j(t+1) < 0.5$ for all parameter values combinations. Hence $(s_i^j, s_{-i}(t))$ cannot be stable. In words, suppose that a player plays a strategy that results in a negative payoff and that he would have obtained a larger payoff if he had played the other strategy. Then it must follow that his play can never be stable. This is because in the case of imperfect recall, if $\pi_i(s_i^j, s_{-i}) < 0$ and $\pi_i(s_i^j, s_{-i}) \leq \pi_i(s_i^k, s_{-i})$, (2.10). This gives us that $\delta\pi_i(s_i^k, s_{-i}) - \pi_i(s_i^j, s_{-i}) \geq 0$, resulting in $\lim_{t\to\infty} P_i^j(t+1) \leq 0.5$ for all parameter value combinations. This means that in the case of imperfect recall a strategy cannot be stable even if it obtained the same negative payoff after playing an alternative strategy.

If $\pi_i(s_i^j, s_{-i}) < 0$ and $\pi_i(s_i^j, s_{-i}) \leq \pi_i(s_i^k, s_{-i})$ and perfect recall s can still be stable for the right choice of parameters. In that case, initial attractions and the history of play before the time that s was starting to be played are relevant again. Strategy s can then be stable in several ways. For example, suppose that $\delta = 1$. Then s can be stable by choosing initial attractions in such a way that $P(s(t)) > 1 - \varepsilon$ at the start of the game. It can also be stable by a history of play that increases the attractions of s more than the attractions of another strategy.

Consider the following example. Player 1 and player 2 have a large initial tendency to play strategy 1 and strategy 2, respectively (initial attractions $A_1^1(0) = A_2^2(0) = 100$, $A_1^2(0) = A_2^1(0) = 0$). We do not consider a depreciation rate and we assume belief learning ($\rho = 0$, $\delta = 1$). Furthermore we assume perfect recall ($\varphi = 1$) and $\lambda = 1$. Finally, the payoffs are given by

Table 2.1
A game with no stable strategy combinations

	s_2^1	s_2^2
s_1^1	$-2, -3$	$-4, -2$
s_1^2	$-2, -2$	$-3, -3$

Note that (s_1^2, s_2^1) is a weak Nash equilibrium. First the players play (s_1^1, s_2^2) with probability close to 1. Playing this strategy relatively increases the attraction of the weak Nash equilibrium for player 1. Then play switches to (s_1^2, s_2^2), also with probability close to 1. But then strategy 1's attraction increases relatively more for player 2. Finally, player 2 switches to strategy 1 with probability close to 1. Hence thereafter they play the weak Nash equilibrium with probability close to 1.

Strategy s can be stable in all other cases than those stated in Theorem 1. If $\pi_i(s_i^j, s_{-i}) > 0$ for both i, then $\delta\pi_i(s_i^k, s_{-i}) - \pi_i(s_i^j, s_{-i}) < 0$ in case of reinforcement learning. If $\pi_i(s_i^j, s_{-i}) \geq \pi_i(s_i^k, s_{-i})$, then s can be stable in case of belief learning. This s can be stable by choosing parameter values such that $\frac{\lambda(1-\rho)}{1-\varphi}$ is high enough.

Theorem 2.3.1 has a few interesting implications:

Lemma 2.3.2 (continued). Strict pure Nash equilibria of a 2×2 game can be stable strategy combinations in the EWA learning model. Weak Nash equilibria cannot be stable when there is imperfect recall and when there are negative payoffs in the Nash equilibrium for an indifferent player.

To see this, note that in case of a strict pure Nash equilibrium $\pi_i(s_i^k, s_{-i}) < \pi_i(s_i^j, s_{-i})$ we can always choose the parameter values in such a way that $P_i^j(t+1) > 1 - \varepsilon$ and (s_i^k, s_{-i}) is stable. The observation that strict Nash equilibria can be stable under belief learning (for large values of $\frac{\lambda(1-\rho)}{1-\varphi}$) is a well-known fact (Camerer, 2003, and also Fudenberg & Levine, 1998). Note that this result also follows directly from Theorem 2.3.1. For the weak pure Nash equilibria, suppose we play strategy s_i^j being the weak pure Nash equilibrium with $\pi_i(s_i^j, s_{-i}) < 0$ and imperfect recall, then we obtain $\delta\pi_i(s_i^k, s_{-i}) - \pi_i(s_i^j, s_{-i}) = -\pi_i(s_i^j, s_{-i})(1-\delta) \geq 0$. Hence $\lim_{t\to\infty} P_i^j(t+1) \leq 0.5$, and it cannot be stable.

Consider the game in Table 2.2 to illustrate Lemma 2.3.2.

Table 2.2
A game with three weak Nash equilibria

	s_2^1	s_2^2
s_1^1	$100+\nu, 90+\nu$	ν, ν
s_1^2	$100+\nu, 100+\nu$	$90+\nu, 100+\nu$

The game in Table 2.2 has three weak pure Nash equilibria, and no mixed strategy Nash equilibrium. Consider belief learning. In this case there is no combination of parameter values such that any s can be stable in case of imperfect recall, because $\delta[\pi_i(s_i^j, s_i(t)) - \pi_i(s_i^{3-j}, s_i(t))] = 0$ for at least one player for each of the three equilibria. Hence the game in Table 2.2 is quite a challenge for belief learning with discounting; it cannot explain convergence to any s although it has three pure Nash equilibria. Now let $\delta = 0$ and $\nu = -110$. Then no s can be stable under reinforcement learning, since all outcomes of this game are negative. More generally, if $\nu = -110$ and we have imperfect recall, there is no $\delta \in [0,1]$ and no combination of values for the other parameters for which any of the four strategy combinations can be stable. To conclude, when we assume discounting the EWA model cannot lead to strategy combinations that are stable in the 2×2 game in Table 2.2 if all outcomes are negative, even though the game has three pure Nash equilibria.

Two other interesting implications of Theorem 2.3.1 are Lemmas 2.3.3 and 2.3.4.

Lemma 2.3.3 (continued). In a 2×2 game with only negative payoffs and no pure-strategy Nash equilibrium, no pure strategy can be made stable.

Lemma 2.3.3 is true because in 2×2 games with only negative payoffs $\pi_i(s_i^1, s_{-i}(t)) < \pi_i(s_k^1, s_{-i}(t))$ for at least one player (otherwise s would be a pure Nash equilibrium), and hence $P(s_i^k, s_{-i}(t)) < 0.5$ for at least one player i.

Lemma 2.3.4 (continued). Any strategy combination in a 2×2 game with positive payoffs for both players can be stable in EWA learning.

Lemma 2.3.4 follows from the fact that if $s = (s_i^j, s_{-i}(t))$ yields positive outcomes to all players, $\delta(\pi_i(s_i^k, s_{-i}(t))) - \pi_i(s_i^j, s_{-i}(t))) < 0$ in case of reinforcement learning.

Theorem 2.3.1 and the lemmas derived from it also have implications for the conditions to distinguish between reinforcement and belief learning. The two most conspicuous implications are:

Lemma 2.3.5 (continued). Under belief learning only pure Nash equilibria can be stable, independent of the sign of the payoffs.

This holds because if $s = (s_i^j, s_{-i}(t))$ is not Nash, then $\delta(\pi_i(s_i^k, s_{-i}(t))) - \pi_i(s_i^j, s_{-i}(t)) > 0$ and $P < 0.5$ for at least one player.

Lemma 2.3.6 (continued). Under reinforcement learning only strategy combinations yielding positive outcomes to both players can be stable.

Lemma 2.3.6 follows directly from the proof of Lemma 2.3.4.

Using Lemma 2.3.5 and Lemma 2.3.6, games can easily be constructed so that belief and reinforcement learning lead to fundamentally different predictions. One example was the game in Table 2.1 with $\nu = 0$: none of the strategy combinations can be stable under belief learning with discounting, but the three Nash equilibria can all be stable under reinforcement learning. Another example is a Prisoner's Dilemma with both positive and negative payoffs, but with mutual defection yielding negative payoffs to both players.

Table 2.3
A Prisoner's Dilemma game

	s_2^1	s_2^2
s_1^1	$10+\nu, 10+\nu$	$-20+\nu, 20+\nu$
s_1^2	$20+\nu, -20+\nu$	$-10+\nu, -10+\nu$

Consider the Prisoner's Dilemma game in Table 2.3, with $\nu = 0$. Under belief learning only the Nash equilibrium (the bottom-right cell) can be stable, whereas under reinforcement learning only the Pareto-optimal but dominated outcome (the top-left cell) can be stable. Note that this difference in prediction is dependent on the value of ν. While the prediction of belief learning is independent of ν, under reinforcement learning either zero ($\nu \leq -10$), one ($-10 < \nu \leq 10$), two ($10 < \nu \leq 20$) or four ($20 < \nu$) strategy combinations can be stable.

The last two lemmas and the Prisoner's Dilemma with mixed outcomes demonstrate that even 2×2 games can be used to differentiate between belief and reinforcement learning. Hence, the suggestions one could get from the literature that it can be difficult or impossible to discern between both models is false. In fact, our results clearly show why previous research has had trouble to distinguish the two empirically. For instance, as far as we know, all previous studies were restricted to games with solely non-negative outcomes. In these games all strategy combinations can be stable, including the pure Nash equilibria that can be stable under belief learning. Only when the pure Nash equilibrium yields a negative outcome for at least one of the players, both types of learning cannot yield the same stable strategy combinations. Moreover, Salmon (2001) used a game with a 0 outcome for at least one player in each of the strategy combinations, which makes it particularly unlikely to be able to distinguish different learning models. This is because a 0 outcome for a strategy does not differentiate between EWA-based reinforcement and belief learning if another strategy than this particular strategy happens to be chosen. Finally, since our results show that payoffs and in particular the sign of the payoffs determine which kind of behavior can be predicted by both types of learning, we can draw two additional conclusions. First, shifting the outcomes in a game has an effect on which s can become stable under reinforcement learning and EWA learning with positive δ (not under belief learning). This implies that if one wants to distinguish the two types of learning and estimate the parameters of a learning model one can also choose to analyze the stable s or streaks of play in sets of 'shifted games'.

It is important to realize that our analysis only shows which strategy combinations can be stable, not with what probability this actually occurs. The probability that s is stable, if it can be, is dependent on the values of all the parameters. Since here our focus is on the possibility of stable behavior we only briefly comment upon this probability. Note that the probability that s is stable can be directly manipulated by choosing skewed initial attractions. A stable strategy will then soon be reached if λ is large. However, note that the stable s need not be a pure Nash equilibrium (for instance, in case of reinforcement learning in a game with only positive outcomes). A slow tendency to a stable s can be modeled with a combination of a low λ and a high φ. Finally, note that an increase in δ can lower the probability that a pure Nash equilibrium is stable, or might slow down the path towards it. For instance, this is true if all outcomes in the game are positive, because then $\delta(\pi_i(s_i^j, s_{-i}(t)) - \pi_i(s_i^k, s_{-i}(t))) < \pi_i(s_i^k, s_{-i}(t))$. That is, by taking the foregone payoff into account more, the strategy combination that is not a Nash equilibrium is also reinforced more, which decreases the speed of convergence to the equilibrium.

2.3.2 Results on finite games

The results of the previous subsection can easily be extended to finite games with an arbitrary number of players with an arbitrary number of strategies. The probability that player i plays $(s_i^j, s_{-i}(t))$ at time t is

$$P_i^j(t+1) = \frac{1}{1+\sum_{k\neq j} \exp^{\lambda(A_i^k(t)-A_i^j(t))}}, \tag{2.11}$$

where $A_i^k(t) - A_i^j(t)$ is equal to (2.6) in the 2×2 case. Strategy s is stable if

$$P_i^j(t+1) = \frac{1}{1+\sum_{k\neq j} \exp^{\lambda(A_i^k(t)-A_i^j(t))}} \geq (1-\varepsilon)^{1/n} \tag{2.12}$$

holds for all players, with ε close to 0, and in the case of imperfect recall. In the limit, this gives

$$\frac{1}{1+\sum_{k\neq j} \exp\left(\frac{\lambda(1-\rho)}{1-\varphi}[\delta\pi_i(s_i^k, s_{-i}(t)) - \pi_i(s_i^j, s_{-i}(t))]\right)} \geq (1-\varepsilon)^{1/n}. \tag{2.13}$$

Note that there is a strategy for which player i obtains his maximum payoff $\pi_{s_{-i}}^{max}$ given the strategies of the other players. Then Theorem 2.3.1 can be generalized to Theorem 2.3.7 as follows:

Theorem 2.3.7 (continued). Strategy $s = (s_i^j, s_{-i}(t))$ in a finite game cannot be stable in EWA learning if it yields for at least one player i an outcome $\pi_i(s_i^j, s_{-i}(t)) < 0$ and $\pi_i(s_i^j, s_{-i}(t)) < \pi_{s_{-i}}^{max}$. In case of imperfect recall, strategy s in any finite game cannot be stable in EWA learning if it yields an outcome $\pi_i(s_i^j, s_{-i}(t)) < 0$ and $\pi_i(s_i^j, s_{-i}(t)) \leq \pi_{s_{-i}}^{max}$. In all other cases s can be stable.

The proof of Theorem 2.3.7 follows the same reasoning as the proof for Theorem 2.3.1, and is therefore omitted.

All the results for the 2×2 game can now be generalized straightforwardly to finite games: Pure Nash equilibria can always be stable (e.g., in case of belief learning), whereas weak pure Nash equilibria cannot be stable when there is imperfect recall and when there are negative payoffs in the Nash equilibrium for the indifferent player (Lemma 2.3.2). And finite games with only negative payoffs and only a mixed strategy Nash equilibrium cannot be stable under EWA learning (Lemma 2.3.3). However, in games with only positive payoffs *any* s can be made stable (e.g., in case of reinforcement learning) (Lemma 2.3.4). Hence also Lemma 2.3.5 and Lemma 2.3.6 distinguishing belief and reinforcement learning still hold.

The fact that our results also hold for games with an arbitrary number of players with a finite strategy set implies that we can likewise construct games so that belief and reinforcement learning imply fundamentally different predictions. Again, games with pure Nash equilibria with negative outcomes for at least some players will differentiate between the two since these equilibria cannot be stable under reinforcement learning, but are the only combinations that can be stable under belief learning.

2.4 Conclusions

An important issue in the field of learning is to what extent one can distinguish behavior resulting from belief and reinforcement learning. Thus far, research suggests that it is difficult or impossible to discern between belief and reinforcement learning models for several sets of games. In the present study we derived predictions for behavior in any finite game on the basis of the EWA learning model, a model that includes both reinforcement and belief learning as special cases. Our main conclusion is that belief and reinforcement learning *can* yield very different predictions as to which behavior can be stable in finite games. Even in simple 2 × 2 games, reinforcement and belief learning derived from the EWA learning model can lead to completely different predictions about which strategy combinations can be stable. A second conclusion, pertaining to EWA learning in general, is that in some games no strategy combination can be stable at all. In particular, games with a negative payoff in each cell for at least one player and only mixed strategy Nash equilibria can never lead to stable behavior. The same holds for games with negative payoffs in each cell for at least one player, one or more weak Nash equilibria and imperfect recall of subjects.

Our results can help identify games in which it should be easy to discern belief and reinforcement learning. This is an important first step in order to differentiate between these two types of learning in empirical data. In games in which all pure Nash equilibria have at least one negative payoff and some strategy combinations yield strictly positive outcomes to all players, there is no overlap at all in the sets of strategy combinations that can be stable under both types of learning. Since predictions of stable behavior do not overlap, distinguishing both types is straightforward, as long as stable behavior does occur. As an example of such a game one can think of the standard Prisoner's Dilemma game with mixed payoffs (see Table 2.3).[4]

However, we also might be able to distinguish both types of learning in games with nonnegative outcomes. In games with one pure Nash equilibrium that is Pareto inferior to some other outcomes in the game, only the Nash equilibrium is stable under belief learning. Yet under reinforcement learning the outcomes of the other strategy combinations can be constructed so that they are stronger attractors than the Nash equilibrium. On the other hand, since both types of learning consider the Nash equilibrium to be a stable strategy combination, it is hard to discern between reinforcement and belief learning.

Our result that belief and reinforcement learning can be distinguished evokes the question why previous studies have concluded that it is difficult to discern between them. The answer is obviously because these other studies have used games that are typically not the games where one would expect to find a difference. Previous studies concentrated on games with solely positive payoffs, games that contain some zero outcomes, and games with a mixed strategy Nash equilibrium.

An implication of our results is that seemingly innocent shifts of payoffs or reinforcements can have far-reaching consequences on the predictions of behavior. That

[4] One way to implement games with negative payoffs in experiments is to give a large initial payoff to participants, e.g., as a show up fee, and make sure that after playing the repeated game the participants go home without a loss. To prevent a "gambling with the house money effect" (Thaler and Johnson, 1990), we suggest to design the experiment such that participants put the show up fee in their pocket and some time (or even a task) is between showing up and playing the game.

is, in case of reinforcement learning, or more in general in case of EWA learning with $\delta < 1$, predictions are affected by shifting the outcomes. Hence in our case it cannot be assumed without loss of generality that payoffs are nonnegative (cf. Rustichini, 1999, p. 249), or that the reinforcement of a strategy $[R(x)]$ is equal to its outcome minus the minimum outcome obtained $[x - x_{min}]$ Erev & Roth (1998, p. 860). Interestingly, empirical studies have indeed shown that learning is affected by adding a constant to all payoffs. For instance, Erev, Bereby-Meyer, and Roth (1999) and Bereby-Meyer & Erev (1998) found that learning was faster in the loss domain. Van Assen, Snijders, and Weesie (2006) found that people behave differently in repeated PDs with negative, mixed, and positive outcomes. See also Cachon & Camerer (1996), Rydval & Ortmann (2005), and Feltovich, Iwasaki, & Oda (2008) for further evidence on the effect of adding a constant to payoffs on learning.

Note that studying behavior of individuals in structurally identical games with shifted outcomes provides a powerful tool to distinguish belief and reinforcement learning; belief learning predicts the same stable behavior in these games, while reinforcement learning predicts that stable behavior is different, even if the same people play these games.

Some studies on learning make explicit use of an aspiration level, recognizing the effect of a shift of payoffs on predicted behavior (e.g., Borgers & Sarin (2000), Erev, Bereby-Meyer & Roth (1999), Macy & Flache (2002), Sarin & Vahid (1999)). An aspiration level or reference point is the value with which each outcome is compared. When an outcome is above the aspiration level, then reinforcement is positive. Macy & Flache (2002) demonstrated through simulation that the value of the aspiration level has a strong effect on a reinforcement model's predictions of the likelihood of cooperative behavior in the Prisoner's Dilemma, Chicken game, and Stag Hunt games, a result that also follows from our formal analysis. Since the aspiration level has such a strong effect on the predictions, the model's flexibility can be increased in future work by incorporating a dynamic aspiration level as in the models that were used in the aforementioned studies.

A standard point of criticism on the EWA model is that it includes so many parameters that it is no wonder that it can fit learning behavior in many games rather well (e.g., Ho, Camerer & Chong (2002)). This criticism motivated Ho, Camerer, and Chong to develop a one-parameter EWA model, called f(unctional)EWA, that replaces fixed parameters with functions of experience, allowing both individual differences in "learning styles" and endogenous cross-game differences (Ho, Camerer & Chong, 2002, p. 194). Our analysis suggests that there might be games where EWA fails to fit. That is, there might be games where no parameter value combination exists that can reasonably predict behavior. Games for which EWAs predictions might fail are games that lead to a negative outcome for at least one player in all strategy combinations, and with only weak pure Nash equilibria. If recall is not perfect, EWA cannot predict stable behavior in such games. As far as we know, perfect recall is not typically found in learning. Hence perfect recall would not be the most obvious explanation of stable behavior in this game, a more likely explanation is that EWA failed in these cases and something else is going on. A direct test of EWA learning would be to fit the EWA model simultaneously to such a game (for instance, a game as in Table 2.1) and to the same game with the outcomes shifted so that all outcomes are positive. A bad fit or very different parameter estimates for both games would be evidence against the EWA model. Note, however, that whether stable behavior occurs in a game with only weak Nash equilibria is an empirical question, as well as the fit of the EWA learning model for such a game.

Note also that the conclusions of this chapter are not in line with Hopkins' claim (2002). He states that "They [reinforcement learning and stochastic fictitious play] embody quite different assumptions about the processing of information and optimization. This paper compares their properties and finds that they are more similar than were thought." (p. 2141). One of the main differences is the assumptions in Hopkins' model. For starters, he chose to accept negative payoffs but always update the propensities with positive reinforcements, resulting in a different model. For instance, we state that in a Prisoner's Dilemma game with only negative payoffs, reinforcement learning is never able to find a stable strategy combination, because all propensities keep decreasing, resulting each time in a switch of strategy. For EWA-based belief learning the sign of the payoffs is irrelevant, resulting in a stable strategy combination: the Nash equilibrium. Another difference is that he uses "perturbed" versions of the two learning models, resulting in different equilibria than the Nash equilibria for stochastic fictitious play. Our belief learning model excludes this option in a theoretical analysis.

Our study has at least three limitations. First, our analysis assumes the EWA learning model. Although the EWA model is well-known and frequently applied, not all types of reinforcement and belief learning are special cases of the EWA model. Hence our results on the conditions when reinforcement and belief learning can be distinguished do not imply that all reinforcement learning can be distinguished from all belief learning under these conditions. Second, we have only identified conditions under which strategy combinations can be stable. We did not derive results on possible differences in predicted behavior of belief and reinforcement learning in cases where no stable behavior occurs. Third, we did not address possible differences in the probability distributions of stable behavior predicted by both learning types. That is, even though both learning types might predict the same possible stable strategy combinations for some games, it might still be that they predict a different probability distribution of stable behavior. So we analyzed only one aspect with respect to which the predictions of both types of learning might differ, but there are certainly more aspects on which both kinds of learning might differ.

As stated in the introduction, the second and longer term goal of our analysis is to come up with a sensible experimental design (type of game, payoffs, number of rounds, etc) to be able to distinguish between belief and reinforcement learning. Our analysis is just a first step to reach this goal. More specifically, we did neither analyze the probability that a stable strategy combination will actually be played in many subsequent rounds depending on the values of the parameters, nor the minimum number of rounds that is required before such stable states are reached. Finally, a related issue are the conditions under which one can successfully estimate the parameters of the learning model with which the data were simulated. A considerable amount of studies focus on this issue, for instance, Salmon (2001), Wilcox (2006), and Camerer (2003) (see also Camerer, 2003 for a discussion of this branch of research). We plan to investigate in future research to what extent unbiased estimates of the parameters can be obtained, and to what extent any possible bias is a function of the experimental conditions. The analysis reported in the present study can guide us and others in this future research.

3 Discerning Reinforcement and Belief Learning in 2×2 games. A Simulation Study

We consider if and when one can distinguish between behavior resulting from either belief or reinforcement learning in repeated 2×2 games. Previous research has suggested that it is difficult or even impossible to decide which kind of learning the empirical data are most consistent with (e.g. Salmon, 2001). We examine this issue by simulating data from learning models in three common but very different types of 2×2 games and investigate whether we can discern between reinforcement and belief learning in an experimental setup. Our conclusion is, contrary to the intuition of previous scholars, that this is possible, especially in games with positive payoffs and in the repeated Prisoner's Dilemma game, even when the repeated game has a relatively small number of rounds. We also show that other characteristics of the players' behavior, such as the number of times a player changes strategy and the number of strategy combinations the player uses, can help differentiate between the two learning models.

3.1 Introduction

In experimental repeated games it can be difficult to determine the way in which players learn and adapt their behavior during the course of the game. The archetypical learning models that have been compared in the literature are reinforcement learning and belief learning. Often both kinds of models fit empirical data equally well (e.g., Camerer, 2003; Feltovich, 2000; Hopkins, 2002). As, among others, Salmon (2001) has argued, the difficulty to distinguish learning models is most apparent in 2×2 games. In this case, the players communicate with each other only by choosing between two options, so not much information goes back and forth (Salmon, 2001). The present study focuses on discerning reinforcement and belief learning models in 2×2 games.

The previous chapter showed that belief and reinforcement learning can be distinguished in many repeated games, including several kinds of 2 × 2 games, when one uses the theoretical concept of "stability". Loosely speaking, a strategy combination (defined as the combined choice of both players) is stable if the probability that this combination is played is close enough to 1 after repeatedly playing this combination in the previous rounds. However, the theoretical result alone is not sufficient. First of all, our theoretical results imply that the learning models can be distinguished *after a sufficient number of rounds have been played*, but it is not clear how large that number needs to be. It is also not clear how likely it is that stability actually occurs in game play: the theoretical results show that it is possible that different learning models lead to different stable strategy combinations, but does this happen often enough to be empirically verifiable? Moreover, the theoretical analyses in the previous chapter did not focus on other possibly important characteristics of play in games that might be used to discern between different kinds of learning. Even if the data showed no signs of players repeatedly choosing the same strategy combination, we might be able to find differences between belief and reinforcement learning by using other characteristics of the game play. In this paper we considered as possible characteristics: how often players change between their alternatives, how often different strategy combinations occur, and how soon players end up in a stable state.

We used the theoretical results of last chapter as a starting point, and simulate data from learning models in repeated 2 × 2 games to find out whether discerning between reinforcement and belief learning can indeed be feasible in an experimental setup, as the theoretical analysis suggested. Our conclusion, corroborating the findings of the last chapter, is that —contrary to the intuition of previous studies (e.g., Salmon, 2001)— discerning between different kinds of learning *is* possible, even in repeated 2 × 2 games with a relatively small number of rounds. Both the stability concept and the other characteristics we examined help us to differentiate reinforcement and belief learning models.

The paper is structured as follows. We first start by describing the learning models as special cases of the Experience Weighted Attraction (EWA) model (e.g., Camerer & Ho, 1999), and then describe some important lemmas from the previous chapter in Section 3.2. Then, we outline some implications of these results for three kinds of repeated games in Section 3.3. The implications of the other characteristics are then discussed in Section 3.4. In Section 3.5, we outline the design of our simulations and we discuss the results in Section 3.6. The three sections after that describe the results for each of the games. In the final section, we conclude with some thoughts about the implications for applying our results to the design and the analysis of behavior in experimental games.

3.2 Theoretical background

The EWA model, which includes belief and reinforcement learning as special cases, was developed and first applied in Camerer & Ho (1999) and used in a large number of papers on interdependent learning (see Camerer (2003), for an overview). The notation used is based on Camerer & Ho (1999). Players are indexed by i $(= 1, 2)$ and the strategy space S_i consists of two discrete choices. Player i's scalar valued payoff in a round t equals $\pi_i(s_i(t), s_{-i}(t))$, with $s(t) = (s_i(t), s_{-i}(t)$ denoting the strategy combination played in round t by player i and the other player. Note that we used the term "strategy" and "strategy combination" for the choice(s) of a player in a given round, and not for the general set of rules that describes a player's behavior across all possible states of the game (as is common in game theory).

The core of the EWA model consists of two variables that are updated after each round. The first variable is $A_i^j(t)$, which is player i's attraction (sometimes called "propensity", e.g., in Erev & Roth, 1998) for strategy $s_i(t) = j$ *after* round t has taken place:

$$A_i^j(t) = \frac{\varphi N(t-1) A_i^j(t-1) + [\delta + (1-\delta) I(s_i(t), j)] \pi_i(s_i^j, s_{-i}(t))}{N(t)}, \tag{3.1}$$

where $\varphi \in [0, 1]$ is a recall parameter (or "discount factor" or "decay rate") that depreciates the attraction value of the given strategy in the previous round. Furthermore, $\delta \in [0, 1]$ is the parameter that determines to what extent the player takes foregone payoffs into account. The $I(s_i(t), j)$ is an indicator function equal to 1 when the played strategy is j and 0 if not. $A_i^j(0)$ is called i's initial propensity to play j. In this study, we assumed that all players have the same initial propensities for all strategies. The second core variable of the EWA model, $N(t)$, determines the extent to which previous outcomes play a role in the current choice of behavior. It is updated as follows:

$$N(t) = \rho N(t-1) + 1, \qquad t \geq 1, \tag{3.2}$$

where ρ is a depreciation rate or retrospective discount factor that measures the fractional impact of previous experience, compared to a new round. To guarantee that experience increases over time ($N(t) > N(t-1)$) it is assumed that $N(0) < \frac{1}{1-\rho}$. Another restriction is that $\rho \in [0, \varphi]$.

Another parameter in the EWA model, λ, determines how the strategies' attractions are transformed into probabilities. The probability that player i plays strategy j at time $t + 1$ in EWA learning is a logit transformation of the strategies' attractions at time t:

$$P_i^j(t+1) = \frac{1}{1 + \exp^{\lambda(A_i^k(t) - A_i^j(t))}}, \tag{3.3}$$

where $k \neq j$ and $\lambda \geq 0$ is called the sensitivity parameter.

From here on, we considered $\rho = 0$ so that $N(t) = 1$ for all t. Hence, the most recent experience gets equal weight relative to prior experience. We justified fixing $\rho = 0$ on several grounds. First, we showed in the previous chapter that the effect of ρ on results concerning stability is precisely opposite to that of λ and φ. Modeling stability can therefore also be carried out with only the latter two parameters. Second, empirical results suggest that ρ and φ are fundamentally related, as indicated by the strong correlation (0.88) between their estimates (Camerer & Ho, 1999, p. 867) when fitting EWA to actual play in several games. Since in particular ρ is difficult to recover when estimating the EWA model to data (Camerer & Ho, 1999, p. 333), it makes most sense to omit ρ from the model. Finally, whereas the scaling parameter λ and recall parameter φ are commonly used in learning models, many models do not consider the depreciation rate ρ at all (Camerer & Ho, 1999).

We first focused on differences with respect to stability in order to distinguish the two types of learning. Stability is defined as a stochastic variant of pure Nash equilibrium: there is a large probability (close to 1) that all players make the same choice in two consecutive rounds. Our concept for stability is similar to what is called *local stability* defined by Fudenberg and Kreps (1995, p. 345) in the context of fictitious belief learning of mixed strategy equilibria. We defined $P(s(t))$ as the probability of players playing a strategy combination $s(t)$ in round t. Then, a strategy combination s is defined to "be stable" in a 2 × 2 game when

$$P(s(t)) \geq (1-\varepsilon)^2 \tag{3.4}$$

after repeated play of s, for some ε close to zero (more details on this definition can be found in the previous chapter). The extra stipulation 'after repeated play of s' is crucial. The idea of stability is that the players continue playing s with a large probability in the next round(s), and as such can be considered to be a probabilistic equilibrium concept for learning models. Using this definition of stability and assuming EWA learning, the following lemmas hold:

Lemma 3.2.1 (continued). Under reinforcement learning, only strategy combinations that yield positive payoffs to both players can be stable.

Lemma 3.2.2 (continued). Under belief learning only pure Nash equilibria of the constituent game can be stable, independent of the sign of the payoffs.

These differences between the learning models have interesting implications. For example, consider a Prisoner's Dilemma game with positive payoffs only. A Prisoner's Dilemma game has one (pure) sub-optimal Nash equilibrium. Under reinforcement learning all strategy combinations can be stable (because all have positive payoffs), whereas under belief learning only the Nash equilibrium can be stable. Combining the two lemmas leads to a third one:

Lemma 3.2.3 (continued). In a 2 × 2 game with only negative payoffs and no pure-strategy Nash equilibrium, no pure strategy combination can be stable.

Under belief learning only the pure Nash equilibrium can be stable (and we did not have one in a game with only a mixed strategy equilibrium), and under reinforcement learning only strategy combinations with positive payoffs can be stable. Hence, observing stable states in experimental data in mixed strategy equilibrium with negative outcomes would argue against both belief and reinforcement learning.

3.3 Theoretical implications on stability in three 2×2 games

The implications of the theoretical results are discussed for three different 2 × 2 games, inspired by Lemma 1 to Lemma 3: the Prisoner's Dilemma game, a Pareto-optimal Nash equilibrium game, and a game with only a mixed strategy equilibrium. Moreover, in these games we varied the payoffs from all positive, all negative, to mixed. The three games and the payoff combinations are selected to create conditions in which behavior implied by the learning models should be either different or similar. Moreover, note that the three games cover an interesting diversity of situations: a situation with a pure equilibrium that is not Pareto optimal and a situation where it is, and a situation where there is just a single mixed-strategy equilibrium.

Table 3.1. The Prisoner's Dilemma game. The abbreviations of the strategies used in the table correspond to *T*op and *B*ottom for player 1, and *L*eft and *R*ight for player 2.

	L	R
T	$5+\Delta, 5+\Delta$	$1+\Delta, 7+\Delta$
B	$7+\Delta, 1+\Delta$	$3+\Delta, 3+\Delta$

We used Δ as a means to create a game with positive, mixed, and negative payoffs with Δ equal to 0, -4, and -8, respectively. A Prisoner's Dilemma game has only one Nash equilibrium, here the bottom-right, which is not Pareto optimal. Our lemmas imply that in belief learning only the bottom-right strategy combination can be stable. In reinforcement learning all strategy combinations with positive payoffs can be stable. The strategy combination with the highest payoff to both players, which is the payoff corresponding to the top-left corner, satisfies the stability condition most easily. In the Prisoner's Dilemma game with mixed payoffs ($\Delta = -4$), top-left is the only strategy combination with positive payoffs for both players. Hence, under reinforcement learning top-left is the only strategy combination that can be stable if payoffs are mixed. With negative payoffs, no strategy combinations are stable under reinforcement learning. Our predictions are summarized in the PD row of Table 3.2.

Table 3.2. Strategy combinations for different games with different payoffs that can be stable.

		Reinforcement	Belief
PD	positive payoffs	All	2
	mixed payoffs	1	2
	negative payoffs	None	2
NE	positive payoffs	All	1
	mixed payoffs	1,3,4	1
	negative payoffs	None	1
ME	positive payoffs	All	None
	mixed payoffs	None	None
	negative payoffs	None	None

1 = top-left; 2 = bottom-right; 3 = top-right; and 4 = bottom-left.

Table 3.3. The Pareto-optimal Nash equilibrium game.

	L	R
T	$7+\Delta, 7+\Delta$	$5+\Delta, 5+\Delta$
B	$5+\Delta, 5+\Delta$	$3+\Delta, 3+\Delta$

Table 3.3 presents the Pareto-optimal Nash equilibrium game, where Δ again shifts the payoff with the values 0, -4, or -8. Belief learning can only be stable in the Nash equilibrium (top-left strategy combination). Under reinforcement learning with positive payoffs *all* strategy combinations can be stable, whereas with mixed payoffs all except the bottom-right can be stable. With negative payoffs no combinations are stable (this can be seen in the NE row of Table 3.2).

Table 3.4. The game with only a mixed strategy equilibrium.

	L	R
T	$5+\Delta, 2+\Delta$	$1+\Delta, 4+\Delta$
B	$2+\Delta, 5+\Delta$	$4+\Delta, 1+\Delta$

Table 3.4 presents the game with only a mixed strategy equilibrium, with Δ equal to 0, -3, or -6. Note that the mixed strategy equilibrium has probabilities $(\frac{2}{3}, \frac{1}{2})$ for strategies (*T*op, *B*ottom) and (*L*eft, *R*ight), and does not depend on Δ. For belief learning no strategy combination is stable. For reinforcement learning and positive payoffs all strategy combinations can be stable. For mixed or negative payoffs no stability can occur. Table 3.2 again summarizes our predictions.

There are two more parameters that we varied: the recall parameter φ and the sensitivity parameter λ. We chose the values of the parameters such that stability is satisfied for different values of ε. By combining equations (3.1), (3.3), and (3.4) the stability condition can be written as:

$$\frac{\lambda}{1-\varphi}[\pi_i(s_i^j, s_{-i}) - \delta\pi_i(s_i^k, s_{-i})] = \ln(\frac{1-\varepsilon}{\varepsilon}). \tag{3.5}$$

Equation (3.5) shows that two factors affect stability; the ratio $\frac{\lambda}{1-\varphi}$, and the payoff $\pi_i(s_i^j, s_{-i})$ (in case of reinforcement learning) or payoff difference $\pi_i(s_i^j, s_{-i}) - \pi_i(s_i^k, s_{-i})$ (in case of belief learning). Using these two factors, we chose three ratios $\frac{\lambda}{1-\varphi}$ to obtain three different stability conditions. First, we chose $\frac{\lambda}{1-\varphi} = 6.9$ to obtain stability with $\varepsilon = 0.001$ for a payoff (difference) of 1. Using this ratio, all strategies are stable as long as the payoff (difference) is at least 1. This implies that a player's probability of repeatedly playing a strategy can be very high, since ε is less than 0.001 for a payoff (difference) of 1. Consequently, for this ratio the pure Nash equilibria are stable under belief learning, and all strategy combinations with positive payoffs are stable under reinforcement learning. Second, we chose $\frac{\lambda}{1-\varphi} = 2.3$ to obtain stability with $\varepsilon = 0.01$ for a payoff (difference) of 2. For this ratio, strategy combinations are stable if the payoff (difference) is not too small. At the same time the probability of repeatedly playing a strategy can be high, but not higher than .99 if the payoff (difference) is 2. As a result, the pure Nash equilibrium is stable under belief learning since the payoff difference equals 2, and under reinforcement learning only if combinations result in a payoff of at least 2. Finally, we chose $\frac{\lambda}{1-\varphi} = 0.314$ such that no strategy combination is stable with $\varepsilon = 0.1$ and a payoff (difference) less than 7. The implication is that no strategy is stable, and that the probability of one player repeatedly choosing a strategy is never larger than 0.9.

For stability only the ratio $\frac{\lambda}{1-\varphi}$ matters, and not the individual values of the two parameters λ and φ. However, the individual values matter for actual play and can lead to other characteristics to distinguish the two types of learning. Hence we chose six different parameter combinations, two for each ratio. The values we chose are:

$$(\varphi, \lambda) \in \{(0.5, 0.157), (0.5, 1.15), (0.5, 3.45), (0.95, 0.0157), (0.95, 0.115), (0.95, 0.345)\}. \tag{3.6}$$

A value of φ equal to 0.5 signifies that the choices in all rounds up to the penultimate round do not affect the propensity as much as the last round, whereas a value equal to 0.95 means that the outcomes in the past exert a strong influence on current behavior. The largest value of λ, 3.45 represents a fast way of learning. That is, one shift of strategy can drastically alter response probabilities, especially in combination with a low value of φ. The two lower values of λ are chosen to make the response probabilities less sensitive to changes in strategy.

Table 3.5 extends Table 3.2 and shows which strategy combinations can be stable, now including the different parameter values λ and φ. Hence, we could find in Table 3.5 for what parameter combination and for what game, which strategy combination is stable for reinforcement and belief learning, with $\varepsilon = 0.01$. For instance, **R**12**B**2 implies that for reinforcement learning both top-left and bottom-right are stable, whereas for belief learning only bottom-right is stable.

Table 3.5. Results regarding stability of strategy combinations (with $\varepsilon = 0.01$) for the Prisoner's Dilemma game (PD), Pareto-optimal Nash equilibrium game (NE), and game with only a mixed strategy equilibrium (ME), different payoffs (positive, mixed, negative), and different values of (φ,λ).

φ	0.95	0.50	0.95	0.50	0.95	0.50
λ	0.0157	0.157	0.115	1.15	0.345	3.45
$\frac{\lambda}{1-\varphi}$	0.314	0.314	2.300	2.300	6.900	6.900
$PD+$	.	.	**R**12**B**2	**R**12**B**2	**R**1234**B**2	**R**1234**B**2
$PD+/-$	.	.	**B**2	**B**2	**R**1**B**2	**R**1**B**2
$PD-$	.	.	**B**2	**B**2	**B**2	**B**2
$NE+$	.	.	**R**1234**B**1	**R**1234**B**1	**R**1234**B**1	**R**1234**B**1
$NE+/-$	.	.	**R**1**B**1	**R**1**B**1	**R**134**B**1	**R**134**B**1
$NE-$	.	.	**B**1	**B**1	**B**1	**B**1
$ME+$	.	.	**R**14	**R**14	**R**1234	**R**1234
$ME+/-$	.	.	.	.	.	.
$ME-$	.	.	.	.	.	.

1 = top-left; 2 = bottom-right; 3 = top-right; 4 = bottom-left. R = reinforcement learning and B = belief learning. +, +/-, and - correspond to positive, mixed, and negative payoffs, respectively.

3.4 Theoretical implications on other characteristics than stability

Theoretical analysis based on equation (3.5) suggests that the speed of entering a streak (repeated play of same strategy combination) is dependent on (i) the payoff (difference) (the larger, the faster), (ii) the ratio $\frac{\lambda}{1-\varphi}$ (the higher the ratio, the faster), In addition, another result is that (iii) λ (the higher, the faster) has a stronger effect on speed than φ. More specifically, we expected higher speed for reinforcement learning with positive payoffs than for belief learning, but lower speed for other cases of reinforcement learning, for the same values of Δ, λ, and φ. Since the payoff differences are the same in all games, the speed of learning under belief learning should be independent of the value of Δ.

Similar analyses suggest that how often the four strategy combinations are played and how often a player changes strategy are affected by the same three variables. We expected the proportion of play of unstable combinations to decrease in the number of rounds played. We also expected that strategy combinations that are not stable are played less frequently if stronger stability conditions are satisfied (with smaller values of ε), that is, with a large payoff (difference) and large ratio $\frac{\lambda}{1-\varphi}$, and for large values of λ. Similarly, less changes of strategy were expected in these cases. Note that results for belief learning are expected to be independent of δ. Moreover, since for belief learning only the pure strategy Nash equilibrium is stable, we expected a particular high probability of play of this equilibrium and not much changes of strategy if the equilibrium is stable.

3.5 Simulating learning models: operationalisation of variables

Our theoretical analyses and their implications led us to simulate learning models in three different games (one mixed strategy equilibrium and two with a pure Nash equilibrium), two types of learning (reinforcement and belief learning), with three different types of payoffs (positive, mixed, and negative), with six different values for φ and λ. This gives us, thus far, $3 \times 2 \times 3 \times 6 = 108$ different scenarios.

Because an important issue in our study is whether players actually choose the same strategy combination repeatedly after playing a reasonable amount of rounds (as opposed to whether a strategy combination can be played repeatedly with a high probability in theory but in practice will occur only after many rounds), we also varied the number of rounds a game is played. In a review study of experimental repeated Prisoner's Dilemma games Sally (1995) found that the average number of rounds per game in an experimental setup equals 30. The largest number of rounds used in a repeated game was 150. We therefore chose to create simulation runs of 10, 30, and 150 rounds per game. We did not use the runs of 150 rounds to compute statistics for the games consisting of 10 and 30 rounds, but ran independent simulations for each. This leads to $108 \times 3 = 324$ scenarios. In our simulations we used the EWA model with parameter values as described above. To minimize sampling error in the results we used $10,000$ simulations per scenario (which amounts to $3,240,000$ repeated games in total). From these simulation runs, we examined under which conditions (if at all) we could determine with which learning model the data were generated.

In the data, we considered stability of behavior by looking at where in the repeated games we found "streaks". A streak was defined as 8 consecutive rounds of ending up in the same strategy combination. We chose to look at 8 consecutive rounds because the probability of randomly choosing one and the same strategy combination in 8 consecutive rounds is very small ($(\frac{1}{4})^{8-1} = 0.00006$). That is, if we found such a streak then this is likely the result of satisfying the stability condition rather than just a coincidence. If a strategy combination is stable with $\varepsilon = 0.01$, the probability that a specific sequence of eight rounds is a streak is higher than 0.85. Obviously, it would have been possible to examine streaks of other sizes, or examine streaks of different size at the same time.

We examined how often streaks of a strategy combination occur in one play of the game, and how soon in the repeated game the streaks occur. We chose to present the frequency of streaks of a combination instead of the probability that this streak is observed in the game, since more interesting information is provided by the frequency. A frequency larger than one implies that players have stopped playing the stable strategy combination at least once, indicating that he combination is not that stable (that is: ε is probably not very small). The speed of entering a streak was operationalized as the round in which the first streak (of any strategy combination) starts. We also kept track of how often the four different strategy combinations are played in the game by a population of players, and how often a player changes strategy during the game.

3.6 Simulation results: restricting the parameter space

We started by discussing the results that hold across all games. We then discussed the Prisoner's Dilemma game, the Pareto-optimal Nash equilibrium game, and the mixed strategy equilibrium in separate sections.

The first thing to note is that, as expected, the results for the two conditions with $\frac{\lambda}{1-\varphi} = 0.314$ do not give us much information regarding streaks; hardly any streaks were found (frequency was less than 1%). Differences between reinforcement and belief learning with respect to how often combinations are played and how often players change strategy in these conditions were also very small (in the order of magnitude of 1%), and in any case not of much use in an experimental setting. Therefore, we omitted the results for $\lambda = 0.0157$ and $\varphi = 0.95$ and those for $\lambda = 0.157$ and $\varphi = 0.50$ in the results sections below.

The second result that holds across all our games is that for any given number of rounds there were no significant differences between the simulations with belief learning with positive, mixed, or negative payoffs. Again, this is consistent with the theoretical prediction. Therefore, for belief learning we restricted ourselves to the results for positive payoffs for each game with a given number of rounds.

3.7 Results for the Prisoner's Dilemma game

Number of streaks

Table 3.6 presents the results regarding the average number of streaks of each strategy combination per simulation of the PD game for all conditions. The numbers that are expected to be larger than zero for some value(s) of $\frac{\lambda}{1-\varphi}$ are presented in italics.[1]

Table 3.6. Average number of streaks of each strategy combination per simulation of the PD as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	BR	TR	BL	TL	BR	TR	BL	TL	BR	TR	BL
$\frac{0.115}{1-0.95}$	R+	*0.03*	*0.02*	0.00	0.00	*0.13*	*0.73*	0.05	0.04	*0.12*	*1.37*	0.07	0.07
$\frac{1.15}{1-0.50}$	R+	*0.25*	*0.39*	0.09	0.10	*0.24*	*0.71*	0.09	0.10	*0.24*	*0.94*	0.10	0.10
$\frac{0.345}{1-0.95}$	R+	*0.19*	*0.27*	*0.05*	*0.05*	*0.20*	*0.64*	*0.08*	*0.09*	*0.20*	*0.64*	*0.11*	*0.11*
$\frac{3.45}{1-0.50}$	R+	*0.24*	*0.27*	*0.23*	*0.23*	*0.24*	*0.27*	*0.24*	*0.24*	*0.25*	*0.32*	*0.24*	*0.24*
$\frac{0.115}{1-0.95}$	R+/-	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.20	0.00	0.00
$\frac{1.15}{1-0.50}$	R+/-	0.07	0.00	0.00	0.00	0.40	0.00	0.00	0.00	2.31	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	R+/-	*0.01*	0.00	0.00	0.00	*0.05*	0.02	0.00	0.00	*0.22*	0.19	0.00	0.00
$\frac{3.45}{1-0.50}$	R+/-	*0.46*	0.00	0.00	0.00	*1.01*	0.00	0.00	0.00	*1.25*	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	B	0.00	*0.00*	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00	*3.57*	0.00	0.00
$\frac{1.15}{1-0.50}$	B	0.00	*0.81*	0.00	0.00	0.00	*1.20*	0.00	0.00	0.00	*3.23*	0.00	0.00
$\frac{0.345}{1-0.95}$	B	0.00	*0.34*	0.00	0.00	0.00	*1.01*	0.00	0.00	0.00	*1.01*	0.00	0.00
$\frac{3.45}{1-0.50}$	B	0.00	*1.00*	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00	*1.00*	0.00	0.00

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. R = reinforcement learning; B = belief learning. +, +/-, - refer to positive, mixed, negative payoffs, respectively.

In general, the simulation results capture the theoretical predictions. The numbers in italics in Table 3.1 are all larger than 0 and the other numbers are in general close(r) to 0. We first consider the predictions for reinforcement with positive payoffs. In this case we observed streaks in the first ten rounds of the game (as can

[1] Sometimes we used 'mutual cooperation' and 'mutual defection' for top-left and bottom-right, respectively, in the Prisoner's Dilemma game.

be seen in the third column of the table). Increasing the number of rounds obviously resulted in an increase in the number of observed streaks (for those cases where streaks were expected). An interesting observation is that in particular the number of streaks for the Nash equilibrium (bottom-right) increased in the number of rounds, but this did not happen for the Pareto optimal (top-left strategy combination) and other combinations. This implies that if a streak of these other three strategy combinations was found, then it most likely started in the first ten rounds of the game. Related to this observation is the effect of λ. For large values of λ (e.g., 3.45), the initial strategy combination often led to a streak of that combination, as can be seen in the last row of the R+ block. Because the probability of each initial choice is about 0.25 here, the frequency of streaks is approximately equal to 0.25 for each of the four combinations. Since higher values of λ tend to 'lock in' behavior in stable combinations after a choice for that combination, the effect of λ is to spread out streaks more evenly over stable combinations. Finally, a considerable and unexpected number of streaks (around 0.10) was found for the top-right and bottom-left strategies for λ equal to 0.115 and 1.15. These streaks occur after an initial choice for that strategy resulting in a high probability of repeating this choice, although not as high as to satisfy the stability condition.

For reinforcement with mixed payoffs streaks of mutual cooperation (top-left strategy combination) were observed as expected, but not many of them (0.22 or less) were found for $\lambda = 0.345$. Streaks were also observed for top-left when λ was 1.15 (0.07, 0.40, 2.31 for 10, 30, 150 rounds, respectively), and for the bottom-right for the two lowest values of λ (around 0.2 for 150 rounds). Again, these streaks occurred since a repeated choice of that strategy results in a high probability of repeating this choice again, although this probability is not high enough to satisfy the stability criterion.

Finally, the results for reinforcement with negative payoffs and belief learning are easily summarized. As expected, no streaks were observed for reinforcement learning, and only streaks were observed for the Nash equilibrium in case of belief learning. Consequently, streaks of other strategy combinations than the pure Nash equilibria provide strong evidence against belief learning.

Time until streaks occur

Table 3.7 presents the average number of the starting round of the first streak in the PD for all conditions, except for reinforcement learning with negative payoffs where no streaks were observed.

Table 3.7. Average number of the starting round of the first streak in the PD game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	1.80	11.51	14.26
$\frac{1.15}{1-0.50}$	R+	1.27	1.96	2.02
$\frac{0.345}{1-0.95}$	R+	1.64	3.96	4.06
$\frac{3.45}{1-0.50}$	R+	1.01	1.03	1.03
$\frac{0.115}{1-0.95}$	R+/-	2.25	16.37	76.11
$\frac{1.15}{1-0.50}$	R+/-	1.65	10.51	40.90
$\frac{0.345}{1-0.95}$	R+/-	1.87	10.59	65.72
$\frac{3.45}{1-0.50}$	R+/-	1.52	3.71	3.76
$\frac{0.115}{1-0.95}$	B	2.47	13.92	15.84
$\frac{1.15}{1-0.50}$	B	1.94	2.88	2.87
$\frac{0.345}{1-0.95}$	B	2.42	4.64	4.64
$\frac{3.45}{1-0.50}$	B	1.75	1.74	1.75

R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

Consider the first row of Table 3.7. The increase from 1.80 to 11.51 combines two pieces of information. First, the proportion of simulations in which a streak is found, and second, the round at which the streak starts. Obviously, the smaller the proportion of simulations with a streak in the first ten rounds, the larger the difference between values in consecutive columns. Since the previous subsection focuses on the occurrence of streaks, and given that we were interested in when streaks start, we focused on comparisons within columns and not on comparisons within rows. Our expectations are generally confirmed. The onset of a streak was faster the higher the payoff (differences), the higher the ratio $\frac{\lambda}{1-\varphi}$, and the higher λ. The only exception was the faster onset for reinforcement than for belief learning, in case of mixed payoffs (for all λ's and 10 rounds).

Played strategy combinations

Table 3.8 presents the percentages of how often a strategy combination was played in the PD game. The strategy combinations that are stable according to the theoretical predictions in Table 3.5 are again in italics.

Table 3.8. Percentages of how often a strategy combination was played in the PD game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	*BR*	*TR*	*BL*	*TL*	*BR*	*TR*	*BL*	*TL*	*BR*	*TR*	*BL*
$\frac{0.115}{1-0.95}$	R+	*18*	*40*	21	21	*14*	*59*	13	13	*12*	*81*	4	3
$\frac{1.15}{1-0.50}$	R+	*25*	*46*	14	14	*25*	*61*	7	7	*25*	*72*	2	2
$\frac{0.345}{1-0.95}$	R+	*21*	*49*	*15*	*15*	*20*	*59*	*10*	*10*	*21*	*63*	*8*	*8*
$\frac{3.45}{1-0.50}$	R+	*25*	*27*	*24*	*24*	*24*	*27*	*24*	*24*	*25*	*29*	*23*	*22*
$\frac{0.115}{1-0.95}$	R+/-	17	36	23	23	12	47	21	21	8	54	19	19
$\frac{1.15}{1-0.50}$	R+/-	35	37	14	14	38	36	13	13	39	35	13	13
$\frac{0.345}{1-0.95}$	R+/-	*16*	48	19	18	*13*	56	16	16	*14*	57	14	14
$\frac{3.45}{1-0.50}$	R+/-	*74*	17	4	4	*91*	6	2	2	*97*	2	0	0
$\frac{0.115}{1-0.95}$	R-	20	32	24	24	19	34	24	24	18	35	24	24
$\frac{1.15}{1-0.50}$	R-	34	36	15	15	36	37	14	14	36	37	13	13
$\frac{0.345}{1-0.95}$	R-	20	35	22	22	20	36	22	22	19	36	22	22
$\frac{3.45}{1-0.50}$	R-	42	42	8	8	43	43	7	7	44	44	6	6
$\frac{0.115}{1-0.95}$	B	10	*50*	20	20	4	*76*	10	10	1	*93*	3	3
$\frac{1.15}{1-0.50}$	B	3	*89*	4	4	1	*95*	2	2	0	*97*	1	1
$\frac{0.345}{1-0.95}$	B	4	*78*	9	9	1	*92*	3	3	0	*98*	1	1
$\frac{3.45}{1-0.50}$	B	2	*92*	3	3	1	*97*	1	1	0	*99*	0	0

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

We expected to find that play of strategy combinations that are not stable decreases in payoff (differences), decreases in the ratio $\frac{\lambda}{1-\varphi}$, and decreases in λ, and moreover that the proportion of play of these strategies decreases in the number of rounds. For reinforcement learning with positive payoffs these expectations are confirmed: play of top-right and bottom-left decreased in λ and rounds whenever these combinations are not stable for $\lambda = 0.115$ and 1.15. Note that the proportion of play of the Nash equilibrium increased over rounds (except for $\lambda = 3.45$, which 'locks in' behavior), whereas play of the top-left did not increase. Apparently, players do not tend to change to repeatedly playing mutual cooperation after a given amount of time.

For reinforcement learning with mixed payoffs and $\lambda = 3.45$ our expectations are also confirmed; play of top-left increased in the number of rounds. Unexpectedly, play of the top-left when $\lambda = 0.45$ was infrequent (less than 16%) and did not increase in the number of rounds. Apparently, the pull towards the top-left was not strong enough in that condition. Note that the reinforcement of top-left was only 1, whereas unilateral cooperation yields a reinforcement of -3. That is, an occasional deviation from top-left by one player will offset the positive reinforcement of at most three subsequent plays of top-left depending on the value of φ.

For reinforcement learning with negative payoffs, mutual defection was played most often, accompanied by mutual cooperation for $\lambda = 1.15$ and $\lambda = 3.45$. In case of belief learning Nash-equilibrium was played most of the time, and our expectations are confirmed that its play increased in ratio $\frac{\lambda}{1-\varphi}$, λ, and number of rounds. The most important result to discern reinforcement and belief learning is that the relative frequency of bottom-right over top-left was much larger for belief learning than for reinforcement learning, and the lowest relative frequency of the top-left was for belief learning.

Changing strategy

Table 3.9 presents the average number of strategy changes per player of the PD game per simulation.

Table 3.9. Average number of strategy changes per player of the PD game per simulation as a function of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	2.67	4.46	5.55
$\frac{1.15}{1-0.50}$	R+	0.21	0.31	0.52
$\frac{0.345}{1-0.95}$	R+	0.87	1.03	1.09
$\frac{3.45}{1-0.50}$	R+	0.01	0.01	0.04
$\frac{0.115}{1-0.95}$	R+/-	4.22	12.29	58.43
$\frac{1.15}{1-0.50}$	R+/-	3.41	10.46	52.31
$\frac{0.345}{1-0.95}$	R+/-	3.55	10.08	47.11
$\frac{3.45}{1-0.50}$	R+/-	1.37	1.52	2.23
$\frac{0.115}{1-0.95}$	R-	5.02	15.77	80.46
$\frac{1.15}{1-0.50}$	R-	7.51	23.99	123.04
$\frac{0.345}{1-0.95}$	R-	5.53	17.45	89.13
$\frac{3.45}{1-0.50}$	R-	8.40	27.14	139.73
$\frac{0.115}{1-0.95}$	B	3.57	5.98	8.86
$\frac{1.15}{1-0.50}$	B	0.80	1.19	3.55
$\frac{0.345}{1-0.95}$	B	1.60	1.64	1.64
$\frac{3.45}{1-0.50}$	B	0.49	0.50	0.50

R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

We expected less changes if more strategies are stable, and in case of reinforcement learning also for larger values of λ and larger positive payoffs. All these expectations are confirmed by the data. For reinforcement learning with positive payoffs very few changes were observed. For instance, in 150 rounds at most 5.55 changes of strategy were observed. Increasing $\frac{\lambda}{1-\varphi}$ and λ decreased the average number of changes. Reinforcement learning with mixed payoffs resulted in a much higher average number of strategy changes, even though the top-left is stable. Since no combination can be stable under reinforcement learning and negative payoffs, more changes were observed than half of the number of rounds in reinforcement learning with negative payoffs. The latter is a distinguishing feature of reinforcement learning with negative payoffs. Finally, few changes were observed for belief learning.

Implications for distinguishing learning models in the repeated PD

It is not straightforward to derive implications for distinguishing learning models that hold in general, as opposed to only for the conditions that we studied. For example, we found that the onset of a streak under belief learning is faster than under reinforcement learning with mixed payoffs given the same values of the parameters. Does this result suggest that early onsets of streaks in mixed payoff PD implies belief learning rather than reinforcement learning? Unfortunately, no. Both learning models can predict the same onset of a streak using *different* values of the parameters. Hence we focus here on results that we argue hold in general, that is: independent of the payoffs in the game and the values of the parameters.

We found four such results. First, if streaks are found of other strategy combinations than the Nash equilibrium, this is direct evidence for reinforcement learning. Under belief learning only streaks of Nash equilibrium play occur. If streaks occur early, this difference between the two types of learning can already be found in a setup with only 10 rounds of play. Notice that under reinforcement learning streaks of combinations that are not Nash equilibria, such as mutual cooperation, are typically found early in the game (before 30 rounds have passed), whereas streaks of the Nash equilibrium can also start later in the game.

Although we found no general results on the onset of streaks, we did find results for the played strategy combinations and the number of strategy changes. Our second general result is that under belief learning mutual defection is always an attractor, and mutual cooperation is played less than any other strategy combination. Hence if mutual cooperation is played more than unilateral cooperation, this is strong evidence in favor of reinforcement learning. Interestingly, also under reinforcement learning the play of mutual defection increased in the number of rounds, which reveals that an increase of mutual defection play in itself is not a discerning feature. Third, the number of strategy changes under reinforcement learning in games with negative payoffs is at least 0.5, that is, higher than predicted by random choice, whereas under belief learning strategies are changing much less frequently. Finally, our fourth result is that under belief learning the behavior is independent of the absolute values of the payoffs (only payoff differences matter), whereas under reinforcement learning behavior is different for positive, mixed, and negative payoffs.

3.8 Results for the Pareto-optimal Nash equilibrium game

Number of streaks

The results regarding the number of streaks per simulation can be found in Table 3.10.

Table 3.10. Average number of streaks of each strategy combination per simulation of the NE as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	*BR*	*TR*	*BL*	*TL*	*BR*	*TR*	*BL*	*TL*	*BR*	*TR*	*BL*
$\frac{0.115}{1-0.95}$	R+	*0.11*	*0.00*	*0.04*	*0.04*	*0.45*	*0.07*	*0.24*	*0.25*	*0.46*	*0.10*	*0.26*	*0.25*
$\frac{1.15}{1-0.50}$	R+	*0.25*	*0.23*	*0.25*	*0.25*	*0.24*	*0.23*	*0.26*	*0.26*	*0.24*	*0.23*	*0.28*	*0.29*
$\frac{0.345}{1-0.95}$	R+	*0.28*	*0.10*	*0.22*	*0.23*	*0.31*	*0.14*	*0.26*	*0.27*	*0.31*	*0.14*	*0.27*	*0.26*
$\frac{3.45}{1-0.50}$	R+	*0.25*	*0.24*	*0.26*	*0.24*	*0.24*	*0.25*	*0.25*	*0.25*	*0.25*	*0.25*	*0.24*	*0.25*
$\frac{0.115}{1-0.95}$	R+/-	*0.01*	0.00	0.00	0.00	*0.84*	0.00	0.00	0.00	*1.64*	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	R+/-	*0.57*	0.00	0.05	0.04	*1.00*	0.00	0.06	0.05	*1.26*	0.00	0.06	0.06
$\frac{0.345}{1-0.95}$	R+/-	*0.28*	0.00	*0.00*	*0.00*	*0.95*	0.00	*0.03*	*0.02*	*0.96*	0.00	*0.04*	*0.04*
$\frac{3.45}{1-0.50}$	R+/-	*0.51*	0.00	*0.23*	*0.23*	*0.53*	0.00	*0.24*	*0.23*	*0.61*	0.00	*0.25*	*0.24*
$\frac{0.115}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	B	*0.00*	0.00	0.00	0.00	*0.99*	0.00	0.00	0.00	*3.57*	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	B	*0.81*	0.00	0.00	0.00	*1.20*	0.00	0.00	0.00	*3.21*	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	B	*0.34*	0.00	0.00	0.00	*1.01*	0.00	0.00	0.00	*1.01*	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	B	*0.99*	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00	*1.00*	0.00	0.00	0.00

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game)

In general the simulation results reflect our theoretical predictions: the numbers in italics are larger than 0 in simulations with more than 10 rounds, and the other numbers are in general close to 0. As expected, for reinforcement with positive payoffs, streaks were observed for all strategies, with most streaks observed for Nash equilibrium play for the two lowest values of λ. A substantial number of streaks were found in the first ten rounds (as can be seen from the third column).

Note that increasing the number of rounds from 30 to 150 hardly increased the number of observed streaks.

For reinforcement learning with mixed payoffs we observed the most streaks for Nash equilibrium play in all conditions, but streaks for other stable strategy combinations were observed as well in case of more than 10 rounds of play. Table 3.10 reveals that if streaks were observed for other than stable strategy combinations, they typically occurred in the first thirty rounds. Noteworthy is that more streaks of top-right and bottom-left were observed if they are not stable in case of $\lambda = 1.15$, than if they are stable but $\lambda = 0.345$. The streaks of unstable strategy combinations occurred after an initial choice for that strategy resulting in a high probability of repeating this choice because of a high value of λ, although not as high as to satisfy the stability condition.

The results for reinforcement with negative payoffs and belief learning are again easily summarized. As expected, no or hardly any streaks were observed for reinforcement learning with negative payoffs, and only streaks were observed for the Nash equilibrium in case of belief learning. Note that for $\frac{\lambda}{1-\varphi} = 6.3$ less streaks were observed since players often end up in a streak and do not change strategy anymore.

Time until streaks

Table 3.11 Average number of the starting round of the first streak in the NE game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	1.79	7.69	8.01
$\frac{1.15}{1-0.50}$	R+	1.02	1.04	1.04
$\frac{0.345}{1-0.95}$	R+	1.39	1.92	1.90
$\frac{3.45}{1-0.50}$	R+	1.00	1.00	1.00
$\frac{0.115}{1-0.95}$	R+/-	2.07	14.86	17.97
$\frac{1.15}{1-0.50}$	R+/-	1.74	3.29	3.36
$\frac{0.345}{1-0.95}$	R+/-	2.02	5.84	5.89
$\frac{3.45}{1-0.50}$	R+/-	1.29	1.34	1.34
$\frac{0.115}{1-0.95}$	R-	.	14.57	76.70
$\frac{1.15}{1-0.50}$	R-	.	.	.
$\frac{0.345}{1-0.95}$	R-	3.00	12.07	71.12
$\frac{3.45}{1-0.50}$	R-	.	.	.
$\frac{0.115}{1-0.95}$	B	2.38	13.95	15.74
$\frac{1.15}{1-0.50}$	B	1.95	2.86	2.85
$\frac{0.345}{1-0.95}$	B	2.43	4.60	4.67
$\frac{3.45}{1-0.50}$	B	1.75	1.75	1.75

R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

Table 3.11 shows the results on the number of rounds until a streak is observed. Again, we focus on comparisons within columns and not on comparisons within rows. No values are presented for those conditions in which the average number of streaks in Table 3.10 was smaller than 0.01, since few data are available for these conditions. Our expectations are generally confirmed. The onset of a streak was faster for higher payoff differences, for higher ratios $\frac{\lambda}{1-\varphi}$, and for higher λ. An interesting observation is that in case of 10 rounds of play, the onset of streaks under reinforcement learning was faster than under belief learning, but slower for 30 and 150 rounds. The explanation for this is that the early streaks were mostly of the strategy combination top-left, which occurred indeed faster for reinforcement learning than for belief learning (but not for the top-right and bottom-left). Note that the average number of starting round is slightly smaller for 150 rounds than for 30 rounds for $\lambda = .345$ under reinforcement learning with positive payoffs. This is peculiar, but this small difference falls well within the margins of simulation error.

Played strategy combinations

Table 3.12. Percentage of how often a strategy combination was played in the NE game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	BR	TR	BL	TL	BR	TR	BL	TL	BR	TR	BL
$\frac{0.115}{1-0.95}$	R+	*35*	*17*	*24*	*24*	*42*	*10*	*24*	*24*	*46*	*6*	*25*	*24*
$\frac{1.15}{1-0.50}$	R+	*26*	*24*	*26*	*25*	*24*	*23*	*27*	*26*	*25*	*20*	*27*	*28*
$\frac{0.345}{1-0.95}$	R+	*30*	*16*	*27*	*27*	*31*	*15*	*26*	*27*	*32*	*14*	*27*	*27*
$\frac{3.45}{1-0.50}$	R+	*25*	*25*	*26*	*24*	*25*	*25*	*25*	*25*	*25*	*25*	*25*	*25*
$\frac{0.115}{1-0.95}$	R+/-	*40*	15	23	23	*68*	6	13	13	*93*	1	3	3
$\frac{1.15}{1-0.50}$	R+/-	*69*	6	13	13	*87*	2	6	5	*97*	0	1	1
$\frac{0.345}{1-0.95}$	R+/-	*63*	8	*14*	*14*	*84*	3	*7*	*6*	*93*	1	*3*	*3*
$\frac{3.45}{1-0.50}$	R+/-	*49*	3	*24*	*24*	*52*	1	*24*	*23*	*57*	0	*22*	*21*
$\frac{0.115}{1-0.95}$	R-	33	18	24	24	37	15	24	24	40	14	23	23
$\frac{1.15}{1-0.50}$	R-	27	13	30	30	26	12	31	31	26	11	31	31
$\frac{0.345}{1-0.95}$	R-	39	15	23	23	43	12	23	23	44	11	22	22
$\frac{3.45}{1-0.50}$	R-	23	10	33	33	10	4	43	43	3	1	48	48
$\frac{0.115}{1-0.95}$	B	*50*	10	20	20	*75*	4	10	10	*93*	1	3	3
$\frac{1.15}{1-0.50}$	B	*89*	3	4	4	*95*	1	2	2	*97*	0	1	1
$\frac{0.345}{1-0.95}$	B	*78*	4	9	9	*92*	1	3	3	*98*	0	1	1
$\frac{3.45}{1-0.50}$	B	*92*	2	3	2	*97*	1	1	1	*99*	0	0	0

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

The expectation that strategy combinations occur less frequently if they are not stable is confirmed by all data of all Pareto optimal Nash games in all conditions. Although all stable combinations were played under reinforcement learning with positive payoffs, frequency of play of bottom-right decreased in the number of rounds. For reinforcement learning with mixed payoffs, we also observed a decrease in play of the stable combinations top-right and bottom-left. These results agree with the result that if other streaks than Nash equilibrium play occur, they occur early in the game. For reinforcement learning with negative games the top-right and the bottom-left were played most often for higher values of $\lambda (> 0.345)$, since the players switched between these two strategy combinations, not being able to find a mutually attractive combination. Finally, the data suggest that the relative frequency of the top-left over other strategy combinations is larger for belief learning than for reinforcement learning.

Changing strategy

Table 3.13. Average number of strategy changes per player of the NE game per simulation as a function of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	2.09	2.63	2.66
$\frac{1.15}{1-0.50}$	R+	0.01	0.02	0.04
$\frac{0.345}{1-0.95}$	R+	0.38	0.38	0.38
$\frac{3.45}{1-0.50}$	R+	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	R+/-	3.78	6.92	7.48
$\frac{1.15}{1-0.50}$	R+/-	0.85	1.04	1.33
$\frac{0.345}{1-0.95}$	R+/-	1.94	2.13	2.15
$\frac{3.45}{1-0.50}$	R+/-	0.27	0.30	0.39
$\frac{0.115}{1-0.95}$	R-	4.83	14.92	74.82
$\frac{1.15}{1-0.50}$	R-	7.16	23.11	118.85
$\frac{0.345}{1-0.95}$	R-	5.03	15.35	77.79
$\frac{3.45}{1-0.50}$	R-	7.78	27.34	146.64
$\frac{0.115}{1-0.95}$	B	3.58	5.98	8.85
$\frac{1.15}{1-0.50}$	B	0.79	1.19	3.52
$\frac{0.345}{1-0.95}$	B	1.60	1.63	1.65
$\frac{3.45}{1-0.50}$	B	0.48	0.50	0.49

R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

Table 3.13 shows the results on strategy changes. Our expectations are confirmed that there are less changes if more strategies are stable, for larger values of λ, and for larger positive payoffs in case of reinforcement learning. The number of changes under reinforcement learning with negative payoffs was again more than half the number of possible changes that can occur. Finally, the results on belief learning and reinforcement learning with positive and mixed payoffs reveal that, if changes occurred, they occurred most often early in the game.

Implications for distinguishing learning models in a repeated NE

Again we only focus on results that hold in general and not only for the conditions that we studied. Differences in behavior between both types of learning can already be observed in games of 30, or even 10 rounds of play. The four general results we found for PD games also hold or the NE game. First, if streaks are found of other strategy combinations than the Nash equilibrium, this is direct evidence for reinforcement learning, since under belief learning only streaks of Nash equilibrium play occur. Streaks other than Nash equilibrium play are typically found early in the game. Even when other strategy combinations show streaks, most streaks are still with the Nash equilibrium.

Second, under belief learning the Nash equilibrium is always an attractor, and playing it increases during the course of the game, whereas play of other combinations decreases. If the frequency of play of other combinations does not decrease in the NE game with negative or positive payoffs, this is an indication of reinforcement learning. Notice, however, that equally frequent play of all strategies can also be expected from belief learners with a very low $\frac{\lambda}{1-\varphi}$. But in such cases one can use streaks or strategy changes to discern the two types of learning.

Third, the proportion of strategy changes under reinforcement learning in games with negative payoffs is at least 0.5 (so higher than predicted by random choice), whereas under belief learning the strategy combinations are changing much less frequently. Finally, belief learning behavior is independent of the absolute values of the payoffs (only payoff differences matter), unlike reinforcement learning behavior.

3.9 Results for the game with only a mixed strategy equilibrium

Number of streaks

Table 3.14 presents the results on the number of streaks.

Table 3.14. Average number of streaks of each strategy combination per simulation of the ME as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	BR	TR	BL	TL	BR	TR	BL	TL	BR	TR	BL
$\frac{0.115}{1-0.95}$	R+	*0.01*	0.00	0.00	*0.01*	*0.33*	0.08	0.11	*0.27*	*1.02*	0.50	0.33	*1.12*
$\frac{1.15}{1-0.50}$	R+	*0.23*	0.15	0.10	*0.28*	*0.32*	0.26	0.14	*0.57*	*0.89*	0.55	0.43	*1.05*
$\frac{0.345}{1-0.95}$	R+	*0.16*	*0.08*	*0.07*	*0.17*	*0.29*	*0.22*	*0.17*	*0.35*	*0.28*	*0.26*	*0.19*	*0.37*
$\frac{3.45}{1-0.50}$	R+	*0.24*	*0.24*	*0.23*	*0.26*	*0.25*	*0.25*	*0.23*	*0.26*	*0.24*	*0.27*	*0.24*	*0.30*
$\frac{0.115}{1-0.95}$	R+/-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	R+/-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	R+/-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	R+/-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	R-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.115}{1-0.95}$	B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{1.15}{1-0.50}$	B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\frac{0.345}{1-0.95}$	B	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
$\frac{3.45}{1-0.50}$	B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

As expected, only streaks were found for reinforcement learning with positive payoffs.[2] Interestingly, these streaks were not mainly found at the start of the game, but also after round 30. A considerable and unexpected number of streaks was found for the top-right and bottom-right strategies when $\frac{\lambda}{1-\varphi} = 2.3$. These streaks occurred after an initial choice for that strategy resulting in a high probability of repeating this choice, although not as high as to satisfy the stability condition.

[2] The occurence of some streaks for belief learning in the case of $\lambda = 0.345$ for 30 and 150 rounds were due to chance.

Time until streaks

Table 3.15. Average number of the starting round of the first streak in the ME game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	2.02	13.40	18.70
$\frac{1.15}{1-0.50}$	R+	1.31	2.51	2.46
$\frac{0.345}{1-0.95}$	R+	1.84	4.45	4.40
$\frac{3.45}{1-0.50}$	R+	1.02	1.03	1.03

R+ = reinforcement learning with positive payoffs; B = belief learning.

Table 3.15 summarizes the results on the onset of streaks for reinforcement learning with positive payoffs only, since streaks were observed only very infrequently in the other conditions. When we focus again on comparisons within columns, we observed as expected a faster onset of a streak the higher ratio $\frac{\lambda}{1-\varphi}$, and λ. Note that the average number of starting round was slightly smaller for 150 rounds than for 30 rounds for two conditions. This is peculiar, but these small difference are negligible.

Played strategy combinations

Table 3.16. Percentages of how often a strategy combination was played in the ME game as a function of the number of rounds, values of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds											
		10				30				150			
		TL	BR	TR	BL	TL	BR	TR	BL	TL	BR	TR	BL
$\frac{0.115}{1-0.95}$	R+	*28*	22	22	*28*	*34*	17	19	*29*	*39*	12	9	*40*
$\frac{1.15}{1-0.50}$	R+	*25*	24	16	*34*	*28*	17	10	*45*	*38*	8	6	*49*
$\frac{0.345}{1-0.95}$	R+	*28*	*22*	*20*	*31*	*29*	*21*	*17*	*34*	*28*	*20*	*14*	*37*
$\frac{3.45}{1-0.50}$	R+	*25*	*25*	*24*	*26*	*26*	*25*	*24*	*26*	*25*	*24*	*23*	*28*
$\frac{0.115}{1-0.95}$	R+/-	28	22	23	27	34	18	23	26	35	16	26	22
$\frac{1.15}{1-0.50}$	R+/-	32	19	23	26	33	18	24	25	33	18	24	25
$\frac{0.345}{1-0.95}$	R+/-	34	17	22	27	36	15	28	21	36	16	29	20
$\frac{3.45}{1-0.50}$	R+/-	34	20	21	25	35	19	22	24	35	19	22	24
$\frac{0.115}{1-0.95}$	R-	27	23	23	26	28	22	23	27	29	21	23	27
$\frac{1.15}{1-0.50}$	R-	27	21	26	26	27	20	26	26	27	20	26	26
$\frac{0.345}{1-0.95}$	R-	29	22	23	27	29	21	23	27	29	21	23	27
$\frac{3.45}{1-0.50}$	R-	26	18	29	27	24	16	30	30	24	15	31	31
$\frac{0.115}{1-0.95}$	B	32	19	22	27	37	15	26	22	36	16	28	20
$\frac{1.15}{1-0.50}$	B	33	19	24	24	34	19	24	23	34	19	25	23
$\frac{0.345}{1-0.95}$	B	38	15	23	23	36	16	28	20	35	16	30	19
$\frac{3.45}{1-0.50}$	B	37	20	21	22	37	20	22	21	38	20	22	20

TL = top-left; BR = bottom-right; TR = top-right; BL = bottom-left; R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

As expected, we observed more frequent play of the stable strategy combination and less play of other combinations for reinforcement learning with positive payoffs. We had no expectations on played strategy combinations for belief learning and reinforcement learning with mixed and negative payoffs, and did not find any systematic pattern therein.

Changing strategy

Table 3.17. Average number of strategy changes per player of the ME game per simulation as a function of $\frac{\lambda}{1-\varphi}$, and scenario.

$\frac{\lambda}{1-\varphi}$	Scenario	Rounds		
		10	30	150
$\frac{0.115}{1-0.95}$	R+	3.13	6.20	10.16
$\frac{1.15}{1-0.50}$	R+	0.26	0.48	1.51
$\frac{0.345}{1-0.95}$	R+	1.06	1.22	1.27
$\frac{3.45}{1-0.50}$	R+	0.01	0.01	0.04
$\frac{0.115}{1-0.95}$	R+/-	4.39	13.39	65.52
$\frac{1.15}{1-0.50}$	R+/-	3.62	11.43	58.21
$\frac{0.345}{1-0.95}$	R+/-	3.70	9.79	46.37
$\frac{3.45}{1-0.50}$	R+/-	3.56	11.14	56.71
$\frac{0.115}{1-0.95}$	R-	4.96	15.71	80.45
$\frac{1.15}{1-0.50}$	R-	7.19	22.94	117.27
$\frac{0.345}{1-0.95}$	R-	5.40	17.03	87.06
$\frac{3.45}{1-0.50}$	R-	7.86	24.95	127.36
$\frac{0.115}{1-0.95}$	B	4.14	11.56	56.58
$\frac{1.15}{1-0.50}$	B	3.54	11.12	56.57
$\frac{0.345}{1-0.95}$	B	2.93	7.47	32.68
$\frac{3.45}{1-0.50}$	B	3.71	11.65	59.30

R = reinforcement learning; B = belief learning. +, +/-, - refer to the outcomes in the game.

The results concerning strategy changes are summarized in Table 3.17. As expected, the number of strategy changes decreased with $\frac{\lambda}{1-\varphi}$ and λ if at least one strategy combination is stable as in reinforcement learning with positive payoffs. Second, also as expected, the number of changes was larger for reinforcement learning with negative payoffs (at least half of the possible number of changes) than for belief learning (less than half of the possible number of changes).

Implications for distinguishing learning models in a repeated ME

Three general results are obtained to discern reinforcement from belief learning. First, if streaks are observed in ME games with positive payoffs, this is direct evidence that both players use reinforcement learning. Notice that neither of the two types of learning can explain the occurrence of streaks in ME games with a negative payoff for at least one player in all cells of the game. Second, the number of times players change strategy is a helpful way to distinguish the two types of learning;

if payoffs are positive, less changes are expected for reinforcement learning, and if payoffs are negative, more changes are expected for reinforcement learning than for belief learning. Finally, behavior is independent of the absolute values of the payoffs for belief learning (only payoff differences matter), unlike for reinforcement learning behavior.

Notice that the absence of a pure Nash equilibrium makes discerning the two learning models using played strategy combinations more difficult. Only (much) less frequent play of one of the strategy combinations hints at reinforcement learning.

3.10 General discussion and conclusion

The present paper addresses when and to what extent reinforcement and belief learning can be discerned. Previous research suggests that this is difficult or even impossible, in particular in 2 × 2 games. In the previous chapter we showed, however, that in theory the two types of learning can be distinguished quite easily in many 2 × 2 games, using the concept of stability. If a strategy combination is stable, a streak of this strategy combination is likely, otherwise it is not. This paper put the stability concept into practice to examine whether differences between the two types of learning can be observed in the onset and number of streaks for different games of different length. Additionally, we investigated whether the two types of learning can be discerned empirically by examining how often players change between their alternatives and how often different strategy combinations occur. For this purpose we simulated reinforcement and belief learning in three important and well-known 2 × 2 games; the Prisoner's Dilemma game with one Pareto dominated pure Nash equilibrium, a game with one Pareto optimal pure Nash equilibrium, and a game with only a mixed strategy equilibrium.

Our predictions on the onset and number of streaks derived from the notion of stability were generally corroborated. That is, streaks were found when expected. Sometimes streaks were observed for strategy combinations that were not stable for the parameters under investigation. However, these strategy combinations were stable for other values of parameters φ and λ, indicating that these streaks were likely but not likely enough to satisfy the stability condition. Hence stability is a useful concept to discern reinforcement from belief learning: if streaks of a strategy combination are observed that cannot be stable under belief learning (reinforcement learning), then it is very unlikely that both players' behavior is based on this type of learning. The onset of a streak turned out not to be useful to discern the two types of learning; faster onset of a streak of a stable strategy combination can be predicted by increasing $\frac{\lambda}{1-\varphi}$, or more particular by increasing λ.

This brings us to some general conclusions on which characteristics of behavior to use to discern reinforcement from belief learning. Note that the two types of learning can only be distinguished for large enough values of $\frac{\lambda}{1-\varphi}$. For very low values of this ratio, such as $\frac{\lambda}{1-\varphi} = 0.314$ as in some of our simulations or even lower, behavior generated by both types of learning resembles random behavior, and both types cannot be distinguished. Our results show, however, that for not too low values of this ratio, reinforcement and belief learning can be discerned in all games, depending on the values of the payoffs. Four general results were found.

First, if streaks are found of other strategy combinations than the Nash equilibrium then this is direct evidence for reinforcement learning relative to belief learning. In this way we could for instance differentiate reinforcement and belief learning with the Prisoner's Dilemma game (for positive, mixed, and negative payoffs), the Pareto-optimal Nash equilibrium game (always for positive and negative payoffs, often with mixed payoffs), and the game with only a mixed strategy equilibrium (with positive, but not with mixed or negative payoffs). Second, under belief learning play of the Nash equilibrium increases during the course of the game, whereas play of the other strategy combinations decreases. If the frequency of play of a combination that is not an equilibrium does not decrease, then this is evidence against belief learning. Third, the proportion of strategy changes of a player is at least .5 under reinforcement learning if all payoffs in the game are negative, whereas this proportion is small under belief learning if there is a pure Nash equilibrium. Finally, the fourth and arguably the most powerful tool to discern the two types of learning is that behavior is not affected by a shift Δ of the payoffs under belief learning, but changes dramatically under reinforcement learning. All these differences between the two types of learning can in principle already be observed after ten rounds, although, obviously, games consisting of more rounds provide higher statistical power to distinguish the two types. Hence, reinforcement and belief learning can be distinguished empirically in the 2 × 2 games studied using streaks, played strategy combinations, and strategy changes, even after ten rounds, provided that behavior does not resemble random behavior.

To conclude, the two types of learning can not only be discerned in theory (based on the concept of stability after a sufficient number of rounds), but also in empirical practice using data of repeated games with only a small number of rounds. Even after 10 rounds of play we could differentiate the models in different games, a result that did not follow from our theoretical analysis. In addition, the number of times a player changes strategy and the number of strategy combinations played gives us a tool (even after 10 rounds) to differentiate the two learning models.

Although we studied reinforcement and belief learning in 2 × 2 games, some of our results generalize in a straightforward way to m by n and N person games. For instance, using results from the previous chapter, it is easy to see that the finding that streaks of strategy combinations other than the Nash equilibrium are direct evidence for reinforcement learning, can be generalized to any finite game. In addition, in the previous chapter the result that behavior is not affected by a shift of Δ of the payoffs under belief learning, but is affected under reinforcement learning. Part of the result that the probability of play of the Nash equilibrium in any game with one pure equilibrium should increase can be proved using the fact that the attraction of the equilibrium grows faster than that of any other strategy combination. Finally, the result about the more than 50% strategy changes in case of only negative payoffs can be generalized as well.

Our four general results provide guidelines for experimental researchers to establish how people learn. More specifically, the general results aid in optimally designing the game and the experiment to discern reinforcement and belief learning. First, we did show that one does not need a large number of rounds to distinguish the different learning models. Furthermore, we recommended designing games with strictly positive payoffs to all players in all strategy combinations (which is convenient from a practical point of view as well). As a result, all combinations can be stable under reinforcement learning. Including games with negative payoffs for some strategy combinations can also distinguish the two types of behavior (i.e., under reinforcement learning the probability of a strategy change is high for the player(s) experiencing a loss). However, games in which players can lose money are less attractive for both researchers and participants of experiments. We do recommend not to use zero-payoffs as outcomes of the games. Using zero-payoffs has two disadvantages. First, the strategy leading to this payoff does not change its attraction, and second, the difference in attraction between this strategy and all other strategies is the same for reinforcement and belief learning. Hence using zero payoffs substantially reduces the power to distinguish the two types of learning. Finally, it is useful to let players play the same game, preferably within-subjects, but with changing values of Δ. This gives information about the change in strategy within-person and highlights the differences between a player using a reinforcement or a belief learning model. It must be noted that this would differentiate between pure reinforcement learners and pure belief learners. No intermediate states are considered in this study.

As our results show, belief and reinforcement learning can be distinguished. How can this be, given that previous research has suggested that this is not or hardly possible? Closer inspection of the games previous researchers have used reveals the answer. Other authors have for instance studied games in which discerning the two types is difficult, such as 2×2 games with only a mixed strategy equilibrium, or games with many zero payoffs (Salmon, 2001). Note that our fourth result on the effect of a shift Δ of the payoffs on behavior has previously received some attention in the literature (cf. Feltovich, 2000) but only for reinforcement learning.

We focused on discerning the two archetypical forms of learning, belief learning and reinforcement learning, using four characteristics of behavior. However, it is perfectly reasonable to expect that learning is a mixture of both types of learning. In terms of the model we used, this implies considering learning with $0 < \delta < 1$. An interesting venue for future research is the issue under what conditions (characteristics of the game including payoffs, and parameter values) these more subtle differences in learning behavior can be discerned.

4 Re-estimating parameters of learning models

We consider under which conditions we can re-estimate three parameters of Camerer & Ho's (1999) EWA learning model from simulated data, generated for different games and scenarios. Previous research has suggested that it is difficult or even impossible to decide which kind of learning model the empirical data are most consistent with (e.g. Salmon, 2001). We examine this issue by simulating data from learning models in three common but very different types of 2 × 2 games and investigate whether we can accurately re-estimate the parameters that were used to generate the data. The results show low rates of convergence of the estimation algorithm, and even if the algorithm converges then biased estimates of the parameters are obtained most of the time. Hence, we must conclude that re-estimating the exact parameters in a quantitative manner is difficult in most experimental setups. However, qualitatively we can find patterns that pinpoint in the direction of either belief or reinforcement learning.

4.1 Introduction

Trying to figure out how players behave in interdependent situations has been a topic of research for many years. In the literature, many examples can be found where players behave systematically different from what game theory prescribes or predicts (Camerer, 2003). A probable and well-known source of this difference can be that game theory emphasizes agents being (mainly) "forward looking": agents can and do think many steps ahead and forecast the expected consequences of their behavior for themselves and for the behavior of others. Experiments have shown that agents typically cannot think more than a few steps ahead (Bosch-Domenech et al., 2002). At best, one could then argue that although agents do not live up to the assumptions of game theory, they might still behave in accordance with the main gist of the game theoretical predictions ("a baseball does not need to know anything about differential equations"). However, experimental results suggest otherwise. It seems that agents are often "backward looking" instead of "forward looking": they learn from past behavior and adapt accordingly. The difference between the game theoretical model and the empirical results has lead to the development of learning models to understand the behavior of agents in interdependence situations, and a growing literature is concerned with trying to identify which among different learning rules appear to be the most in accordance with human behavior (Chapter 6 of Camerer, 2003 presents an overview of this literature).

To realize the magnitude of the problem, it is important to realize that there are many different learning models that can be fitted to experimental data. Even given the large variety, a consensus is starting to develop about the fact that simple learning rules can approximate learning better than equilibrium (Camerer, 2003). If we want to make further progress in research on how players learn in interdependent situations, studies are crucial in which the performance of different learning models are compared and in which the accuracy of the estimation of these learning models is examined.

Previous research on estimating the proper learning model from experimental data has proven to be an arduous endeavor. Several papers are committed to determining how well various learning models can be estimated and re-estimated[1] on different kinds of data sets. One well-known paper is the work by Salmon (2001). His paper examines how accurately four learning models (two learning models from Mookerjee and Sopher (Mookerjee & Sopher, 1994 & Mookerjee & Sopher, 1997), one from Cheung and Friedman (Cheung & Friedman, 1997), and Camerer & Ho's Experience Weighted Attraction (EWA) approach (Camerer & Ho, 1999)) can be identified from the data after having been generated by these four models. Salmon's results revealed severe difficulties in identifying which model was used to generate the data. These difficulties were most prominent in 2×2 games.

But not only Salmon experienced problems in identifying which learning model was used to generate the data. For instance, Feltovich (2000) performed an experiment involving multistage asymmetric-information games. The results of the experiment were compared with the predictions from both reinforcement as well as belief learning models. These two models performed better than the Nash equilibrium in most cases, but he also concluded that the relative performance of the two learning models depended on conditions of the experiment and varied widely according to which criterion of success is used. He stated: "It may well be that there is no single model that is able to describe behavior well in all situations; much more experimental work is necessary before such a question can be answered." (Feltovich, 2000, p. 638). Wilcox (2006) studied the effect of adding a parameter reflecting heterogeneity to reinforcement and belief learning models. On the basis of his results he stated: "There are probably limits to what experimental design and method alone can achieve when models contain lagged dependent variables and subjects are heterogeneous." (Wilcox, 2006, p. 1288). Two other examples are Cabrales & Garcia-Fontes (2000) and Blume et al. (2002). Cabrales & Garcia-Fontes (2000) studied the EWA model of Camerer & Ho (1999) and concluded that only for a very large number of trials (typically 1000 or more), results on the estimation of one of EWAs parameters (δ) were encouraging. Blume et. al came to the same conclusion: accurate or unbiased estimation was poor in data with only a few subjects or number of periods, but improves a lot with an increased sample of subjects and number of periods. The main conclusion in these and other papers is that many different learning models may perform about equally well especially for games with

[1] We refer to re-estimation when the 'data' are simulated using a particular learning model, and the learning model is (re-)estimated using these simulated data.

a shorter number of rounds. Therefore, given experimental data, it may be difficult to identify how players learn while playing the game(s).

As a result of this conclusion, some authors suggested that learning models overfit the data, and therefore suggested to use simpler learning models (Camerer, 2003, Ho et al, 2002, Erev & Roth, 1998, Haruvy & Erev, 2002). As a response to this suggestion, Ho, Camerer, & Chong (2002) developed the functional EWA (fEWA) model, which is a one parameter model constructed from the EWA model that has two variables, in total consisting of 5 parameters. They showed that the fEWA model generally forecasts better than standard belief and reinforcement models, which have more parameters than fEWA. However, the question is whether reducing the number of parameters is not too much a simplification of real behavior in interdepence situations. The parameters of the EWA model represent actual properties of learners. Reducing or combining them to only one parameter therefore leads to a loss in information about how players actually learn. Hence one would not like to reduce the complexity of the model if it has this negative consequence of reducing everything to a rather mystical single parameter.

We therefore consider the Experience Weighted Attraction (EWA) model from Camerer & Ho (1999), and examine whether estimation of (all) its parameters leads to accurate estimates. Some other studies on the estimation of the EWA model have been conducted. Camerer, Ho, & Chong (2002) present an overview of some of the findings. The identification of parameters is quite good in games with many strategies (more than 4) and in various games with a Pareto-optimal strategy (Camerer et al, 2002, p. 138). Despite Salmon's (2001) finding that models are in general poorly recoverable, Salmon found that data generated by $\delta = 0$ (EWA-based reinforcement learning) or $\delta = 1$ (EWA-based belief learning) is estimated well in the sense that the correct restriction is rejected infrequently (20% of the cases) and the wrong restriction is almost always rejected. As stated before, in their work Cabrales & Garcia-Fontes (2000) and Blume et al. (2002) re-estimate the original values of the parameters well, but only using data from play of a very large number of rounds ($1,000$ rounds). For an experimental economist this would be considered an unrealistic number of rounds. However, the previous chapters suggested that qualitative differences between reinforcement and belief learning models can be found after a mere 10 rounds of play. By re-estimation the parameters on the same data, we find out how these two paradoxical findings (distinguish between belief and reinforcement after 10 rounds of play versus one needs a large number of rounds) can be reconciled.

In the previous chapters, we considered the Experience Weighted Attraction (EWA) model from Camerer & Ho (1999) and used theoretical results regarding stable strategy combinations (the strategy combinations that may be played repeatedly) to design games that, at least in theory, can be used to differentiate between different learning models. We derived analytically in Chapter 2 that there are games where two learning models (EWA-based reinforcement learning and belief learning) can be differentiated, even in 2×2 games. In addition, in Chapter 3 we showed that data generated by EWA-based reinforcement learners and by EWA-based belief learners can be discerned by looking at the presence of streaks of play of the same strategy combination. We also showed in Chapter 3 how several other characteristics (how often players change strategy during the game, and how often the different kinds of outcomes occur during the cause of the game, behavior in games with shifted payoffs) can likewise be used to further differentiate the two types of learning. The results of the previous two chapters therefore suggest that accurate estimation of EWA parameters may actually be feasible.

However, there are many types of learning, from which we only considered "pure" belief and "pure" reinforcement learning, and nothing in between. We therefore consider a broader class of learning models in this chapter. The EWA model has a parameter (δ) that specifies different types of learning, with a type of reinforcement ($\delta = 0$) and a type of belief learning ($\delta = 1$) as special and extreme cases. In this chapter, we analyze whether we can correctly re-estimate the value of delta not only from data generated by these types of belief and reinforcement learning, but by learners with $\delta = 0.5$ as well.

Since the EWA model has other parameters than delta corresponding with other properties of players, we also examine if we can accurately (re-)estimate two other parameters: the recall parameter (φ) that specifies the extent to which the history of the game is taken into account by the player, and the sensitivity parameter that specifies how likely a player is to play a favored strategy (λ). The estimation of other and less important EWA parameters is not investigated, for the following reasons. We saw in Chapter 2 that the initial attractions become irrelevant for play in many rounds because $\varphi < 1$, hence we omit this parameter from investigation. In addition, we also omit ρ from further investigation. We justify fixing $\rho = 0$ on several grounds. Empirical results suggest that ρ and φ are fundamentally related, as indicated by the strong correlation (0.88) between their estimates (Camerer & Ho, 1999, p. 867) when fitting EWA to actual play in several games. Since in particular ρ is difficult to recover when estimating the EWA model to data (Camerer & Ho, 1999, p.333), it makes most sense to omit ρ from the model. Finally, whereas the scaling parameter λ and recall parameter φ are commonly used in learning models, many models do not consider the depreciation rate ρ at all (Camerer & Ho, 1999).

To examine whether we can accurately estimate the three EWA parameters, we consider three common but very different types of games, with different parameter combinations for each game. The first game has one pure Nash equilibrium and one (other) strategy combination which is Pareto optimal (the Prisoner's Dilemma game), the second has one pure Nash equilibrium that is also Pareto optimal, and the last type is a mixed equilibrium-only game. For each of these three games we use three different values for the payoffs to create games with only positive, only negative, or mixed payoffs. Play is simulated in relatively short games (10 rounds), long games (150 rounds), and games with 30 rounds. In Section 4.3 we discuss the estimation results, game by game.

As the analyses will show, estimation is difficult. This is mostly the case when we consider those scenarios where many streaks are present. A conclusion that is easiest to grasp through the following intuition: in games without streaks, much information goes back and forth because behavior changes in most rounds, and this helps the re-estimation enormously. As long as streaks of the same choices are being played over and over again, extra rounds do not help much in deriving information about the parameters. However, only one extra change in strategy can give much information. Increasing the number of rounds does result in somewhat better re-estimations in most cases. In general, it also holds that better predictions are available for increasing λ values.

We now first introduce the different games and scenarios with which we are going to simulate data. Then we discuss our methods of re-estimation and describe how these are used on the simulated data to retrieve the original set of parameters. Furthermore, we first discuss general results holding over all games, before discussing the results and conclusions per game. We end with some recommendations for future research on identifying how players learn in games.

4.2 Different games and scenarios

In previous chapters, we showed that we can –on both an analytical level as well as based on simulation results– differentiate between the behavior of reinforcement and belief learners. Both types of learning are extreme cases of the EWA model. In this chapter, we change our approach. Before, we drew conclusions directly from the behavior of players in consecutive rounds. In this chapter, we use these results to estimate parameters. This changes the approach from a more qualitative one (which qualitative differences do we see in the play over rounds) to a more quantitative one: how well can we re-estimate the parameters that were used to model the data.

4.2.1 Choosing EWA parameter values

The EWA model was developed and first applied by Camerer & Ho (1999) and used in a large number of papers on interdependent learning (see Chapter 6 in Camerer, 2003, for an overview). Players are indexed by i $(= 1, 2)$ and the strategy space S_i consists of two discrete choices, **T**op and **B**ottom for player 1 and **L**eft and **R**ight for player 2. Denote the actual strategy chosen by player i in round t by $s_i(t)$ and the strategy chosen by the other player by $s_{-i}(t)$. The scalar-valued payoff function of player i is denoted by $\pi_i(s_i, s_{-i})$. Player i's payoff in a round t equals $\pi_i(s_i(t), s_{-i}(t))$. Finally, $s(t)$ is the strategy combination played in round t. Note that there is an issue that might cause confusion, especially for those familiar with game theory. Here, the term "strategy" and "strategy combination" is used for the choice(s) of a player in a given round, and not for the general set of rules that describes a player's behavior across all possible states of the game (as is common in game theory).

The core of the EWA model consists of two variables that are updated after each round. The first variable is $A_i^j(t)$, which is player i's attraction (sometimes called "propensity", e.g., in Erev & Roth, 1998) for strategy $s_i(t) = j$ *after* round t has taken place:

$$A_i^j(t) = \frac{\varphi N(t-1) A_i^j(t-1) + [\delta + (1-\delta) I(j, s_i(t))] \pi_i(s_i^j, s_{-i}(t))}{N(t)} \tag{4.1}$$

where $\varphi \in [0, 1]$ is a recall parameter (or "discount factor" or "decay rate") that depreciates the attraction value of the given strategy in the previous round. Furthermore, $\delta \in [0, 1]$ is the parameter that determines to what extent the player takes foregone payoffs into account. The $I(x,y)$ is an indicator function equal to 1 when $x = y$ and 0 if not. $A_i^j(0)$ is called $i's$ initial propensity to play j. In the present study it is assumed that all players have the same initial propensities for all strategies. The second core variable of the EWA model, $N(t)$, determines the extent to which previous outcomes play a role in the current choice of behavior. It is updated through

$$N(t) = \rho N(t-1) + 1, \qquad t \geq 1, \tag{4.2}$$

where ρ is a depreciation rate or retrospective discount factor that measures the fractional impact of previous experience, compared to a new round. To guarantee that experience increases over time ($N(t) > N(t-1)$) it is assumed that $N(0) < \frac{1}{1-\rho}$. Another restriction is that $\rho \in [0, \varphi]$.

Another parameter in the EWA model, λ, determines how the strategies' attractions are transformed into probabilities. The probability that player i plays strategy j at time $t+1$ in EWA learning is a logit transformation of the strategies' attractions at time t:

$$P_i^j(t+1) = \frac{\exp^{\lambda A_i^j(t)}}{\exp^{\lambda A_i^j(t)} + \exp^{\lambda A_i^k(t)}} = \frac{1}{1 + \exp^{\lambda(A_i^k(t) - A_i^j(t))}}, \tag{4.3}$$

where $\lambda \geq 0$ is called the sensitivity parameter.

From here on, we consider $\rho = 0$ such that $N(t) = 1$ for all t. Hence, the most recent experience gets equal weight relative to prior experience. We justify not using ρ on two grounds. First, in Chapter 2 we have shown that the effect of ρ on results concerning stability is precisely opposite to that of λ and φ. Camerer & Ho (1999) confirms this notion (p. 870). Moreover, whereas the sensitivity parameter λ and recall parameter φ are commonly used, many learning models do not consider the depreciation rate ρ.

In Chapter 2 and Chapter 3, we saw that if we choose our games and payoffs wisely then differences between learning reinforcement and belief learning models could be found in the simulation data. In this chapter we consider the same games and payoffs. Our first game was the repeated Prisoner's Dilemma game:

Table 4.1 The Prisoner's Dilemma game

	L	R
T	$5+\Delta, 5+\Delta$	$1+\Delta, 7+\Delta$
B	$7+\Delta, 1+\Delta$	$3+\Delta, 3+\Delta$

where $\Delta = 0, -4, -8$. Recall that all strategy combinations were stable under reinforcement learning with positive payoffs, the top-left strategy combination was stable under reinforcement learning with mixed payoffs, and no strategy combinations were stable under reinforcement learning with negative payoffs. For belief learning, the bottom-right strategy combination was stable, regardless of Δ. The second game we considered is the Pareto-optimal Nash equilibrium game:

Table 4.2 The Pareto-optimal Nash equilibrium game

	L	R
T	$7+\Delta, 7+\Delta$	$5+\Delta, 5+\Delta$
B	$5+\Delta, 5+\Delta$	$3+\Delta, 3+\Delta$

where $\Delta = 0, -4, -8$. Recall that all strategy combinations were stable under reinforcement learning with positive payoffs, the top-left, top-right, and bottom-left strategy combinations were stable under reinforcement learning with mixed payoffs, and no strategy combinations were stable under reinforcement learning with negative payoffs. For belief learning, the top-left strategy combination was stable, regardless of Δ. Finally, we considered the game with only a mixed strategy equilibrium:

Table 4.3 The game with only a mixed strategy equilibrium

	L	R
T	$5+\Delta, 2+\Delta$	$1+\Delta, 4+\Delta$
B	$2+\Delta, 5+\Delta$	$4+\Delta, 1+\Delta$

where $\Delta = 0, -3, -6$. Only under reinforcement learning with positive payoffs, we had stability here. In that case there was (potential) stability for all strategy combinations. Note that whether stability actually occurs, also depends on the value of λ and φ.

Next, we consider the parameters of the EWA-model. For δ, we use $\delta = 0$ (reinforcement learning with a fixed reference point), $\delta = .5$, and $\delta = 1$ (belief learning). For the other parameters we use the same values as we used in previous chapters. For λ and φ we use the following six combinations;

$$(\varphi, \lambda) \in \{(0.5, 0.157), (0.5, 1.15), (0.5, 3.45), (0.95, 0.0157), (0.95, 0.115), (0.95, 0.345)\} \tag{4.4}$$

For pairs ($\lambda = 0.0157, \varphi = 0.95$) and ($\lambda = 0.157, \varphi = 0.50$), we have no stable strategy combinations, for pairs ($\lambda = 0.115, \varphi = 0.95$) and ($\lambda = 1.15$, $\varphi = 0.50$), we sometimes have stable strategy combinations, and for pairs ($\lambda = 0.345, \varphi = 0.95$) and ($\lambda = 3.45$, $\varphi = 0.50$) we always have stable strategy combinations (for the details, see Chapter 3). Recall that stable strategy combinations only helped in finding differences for $\delta = 0$ and $\delta = 1$ when they could occur for one, but not for the other. This is, for instance, in the case with negative payoffs, where under reinforcement learning a lot of strategy changes occurred, whereas under belief learning stability may be present and less strategy changes occur in case there is a pure Nash equilibrium.

Furthermore, recall the conclusions from the previous chapter with respect to the three characteristics discussed. First, the strategy that leads to a Nash equilibrium is always an attractor under belief learning, while it is not under reinforcement learning. Hence, the corresponding strategy combinations are played more frequently if the number of rounds increases. Second, the proportion of strategy changes of a player is large under reinforcement learning if all payoffs in the game are negative (at least .5 if all payoffs have the same order of magnitude), whereas this proportion is small under belief learning if there is a pure Nash equilibrium. The onset of streaks does not help us to differentiate between reinforcement and belief learning. Finally, also recall that the sign of the payoffs does not influence belief learners, but it does influence reinforcement learners. In the first case, only the relative difference is of importance. This result is not used in this chapter, because we only re-estimate parameters using the behavior of one game and do not combine multiple games.

With these results in the back of our minds, we can make predictions on how accurate the parameter estimates will be in the conditions. In the previous chapters, we showed that reinforcement and belief learning differ in when they generate streaks. We expect that δ is estimated more accurately in games where reinforcement and belief learning differ in the streaks they may produce. For instance, we expect that in the Prisoner's Dilemma game and the Pareto-optimal Nash equilibrium game, we should be able to accurately estimate δ for negative payoffs; here no streaks are expected for reinforcement learning, but are expected for belief learning for the pure Nash equilibrium. Similarly, more accurate estimation is predicted for PD games with mixed payoffs, since different streaks result from reinforcement learning (mutual cooperation) than for belief learning (mutual defection). Results on stability do not enable us to derive predictions on how accurate estimation of the parameters are in games with no streaks (game with only a mixed strategy equilibrium) or games where the same streaks occur for belief and reinforcement learning (Pareto-optimal Nash equilibrium game with positive payoffs).

Besides the presence of streaks, there are other characteristics that may help us to make predictions on how accurate the parameter estimates will be. For instance, the number of strategy changes during a game. If belief and reinforcement learning differ in the extent to which they lead to strategy changes, estimation probably is more accurate. We know that in games with a pure equilibrium (such as the Prisoner's Dilemma game and the Pareto-optimal Nash equilibrium game) and negative payoffs, the number of strategy changes for reinforcement learning was larger than for belief learning. But also in the game with only a mixed strategy equilibrium with negative payoffs, more strategy changes are expected under reinforcement learning. Hence, we predict that in these games estimation of δ is more accurate. Results on strategy changes do not allow us to derive predictions on how accurate estimation of the parameters are in games where reinforcement learning and belief learning don't differ much in generating strategy changes. This is the case, for instance, in the Pareto-optimal Nash equilibrium game with positive payoffs.

4.2.2 Re-estimating based on maximum likelihood

The parameter values were estimated using the maximum likelihood estimation method. The maximum likelihood estimation (MLE) method is a standard statistical method used to make inferences about parameters of the underlying probability distribution from a given data set.

Loosely speaking, the likelihood of a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. This expression contains the unknown model parameters. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood

Estimators (MLE's).

The likelihood of the data given the parameters, is denoted by

$$L(\underline{\vartheta}) = \prod_{round=1}^{maxround} \prod_{player=1}^{maxplayer} p_{player,k_{player,round}}(round; \underline{\vartheta}). \tag{4.5}$$

The procedure was therefore to first simulate data based on our chosen parameter values, and then to try and see whether maximum likelihood estimation leads to estimates for the parameter values that were close to the originally chosen values.

The ML-estimator is "best asymptotically normal". This implies that when the number of rounds tends to infinity, the distribution of the estimator is a normal distribution with a mean equal to the parameter and a standard deviation equal to the minimum variance bound (the Cramer-Rao limit). The practical importance of this behavior in the limit is that it is often much easier to calculate than the actual variance of a given estimator, and is independent of the choice of estimator (Gourieroux & Monfort, 1995).

We estimated the three parameters for the row player: $\varphi \in [0,1]$, $\lambda \in [0,\infty)$, and $\delta \in [0,1]$. This took approximately 15 seconds per game. For our estimation we used the STATA maximum likelihood module. Iteratively, the numerical method attempts to find the (local) maximum within a given number of iterations. The default number of iterations is 5. We started with 10 iterations. After some first results, however, we increased the number of iterations to 50. The estimation algorithm uses the results from the previous iteration to find a new location to search for the maximum, given a certain tolerance. This tolerance equals $1e-7$ by default in STATA: if the difference between the likelihoods in round t and $t+1$ is smaller than $1e-7$, then the re-estimation is considered converged. In order to increase the number of converged cases, we changed this tolerance to 0.01. In addition, we started out using no initial value for the likelihood to start with, but we also tried to "help" the estimation algorithm finding a decent value by choosing a better starting value of the parameters. This did not change the results for the convergence, however. In case no maximum was found after 50 iterations, we recorded the convergence of the maximum likelihood as failed. Convergence of the estimation procedure is discussed in Gould et al. (2010).

The reason for the failing convergence in our study is hard to pinpoint. A reason could be that the algorithm required more iterations. However, further increasing the number of iterations to 100 did not lead to new cases of convergence. Moreover, using starting values close to the true values did not lead to better results either. Another reason is that play in the games under consideration does not contain enough information to allow accurate estimation of the parameters. We will come back to this issue in the final section.

Next to converged and non-converged estimates, there were also cases where the algorithm was not able to re-estimate parameters at all and failed to find any solution. In this case, we will omit this result (represented by a "." in the tables) from further analysis. Again, these results do not change by changing the tolerance or choosing better starting values; they are the result of flat estimators, which makes it impossible for the STATA module to produce any values (that is, any/no value could fit).

For each type of game, we simulated play for 162 different cases: for positive, mixed, and negative payoffs, for $\delta = 0, 0.5, 1$, for the 6 different λ, φ combinations, and for 10, 30 and 150 rounds. We performed $1,000$ simulations per case. This led to a total of $486,000$ simulations and thereto to the same number of re-estimations.

4.3 The estimation results

On average, the standard error for the parameters δ and φ in the Prisoner's Dilemma game were 0.018, in the Pareto-optimal Nash equilibrium game 0.012 and 0.013, respectively, and in the game with only a mixed strategy equilibrium 0.007 and 0.006. For λ the standard error was high for all types of games (order of 10^{19}). Since this parameter had no upper bound such large values suggest a lack of convergence.[2] We also considered the 90% percentiles of the standard error. In the Prisoner's Dilemma game, in 90% of all cases the standard errors of δ and φ were not larger than 0.03 and 0.04. For the Pareto-optimal Nash equilibrium game, the standard error was not larger than 0.02 for both δ and φ. Finally, for the game with only a mixed strategy equilibrium the standard errors of δ and φ were not larger than 0.01. These standard errors imply that the estimators for these parameters were hardly ever unbiased: the absolute difference of $\hat{\delta} - \delta$ $(\hat{\varphi} - \varphi)$ was most of the times larger than 1.96 times the standard error.

In the tables we report the bias and whether the 98% frequency interval contains the actual value of the parameter. The 98% frequency interval contains all estimated values, except the 1% lowest and 1% highest estimated parameter values. Even when the estimation is biased, we would hope that at least the true parameter value is in the 98% frequency interval.

4.3.1 Convergence

We first consider the Prisoner's Dilemma game and check how many simulations have a converged maximum likelihood estimator (because not all of them do). In addition, we also mention the percentage of simulations where the program could not estimate a parameter. The results can be found in Table 4.4.

[2] We tried imposing different upper bounds for λ and have also tried starting values close to the true value to find out if this could help in estimating λ more accurately, but this did not improve the results.

Table 4.4 The percentage of simulations in which convergence was achieved (between brackets how many times the program could not find a value of the parameters). All table entries concern the Prisoner's Dilemma game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds.

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	11(1)	11(0)	14(0)	13(1)	11(0)	15(0)	13(1)	8(0)	12(0)
	$\frac{0.157}{1-0.50}$	8(15)	17(1)	45(0)	6(12)	17(0)	53(0)	8(8)	8(0)	18(0)
	$\frac{0.115}{1-0.95}$	4(26)	4(26)	5(26)	5(25)	4(26)	6(22)	4(20)	5(19)	4(17)
	$\frac{1.15}{1-0.50}$	4(55)	26(57)	9(56)	9(86)	6(88)	8(88)	0(99)	0(98)	1(97)
	$\frac{0.345}{1-0.95}$	5(56)	17(54)	7(57)	9(65)	7(66)	9(66)	4(72)	3(73)	5(73)
	$\frac{3.45}{1-0.50}$	0(51)	27(49)	2(51)	4(96)	4(95)	4(95)	0(100)	0(100)	0(100)
Mixed	$\frac{0.0157}{1-0.95}$	18(1)	25(0)	26(0)	17(0)	24(0)	32(0)	17(0)	27(0)	31(0)
	$\frac{0.157}{1-0.50}$	16(2)	22(0)	22(0)	18(4)	27(0)	30(0)	15(6)	27(0)	27(0)
	$\frac{0.115}{1-0.95}$	12(7)	16(3)	29(1)	7(12)	10(9)	14(10)	5(23)	7(20)	8(19)
	$\frac{1.15}{1-0.50}$	1(66)	2(39)	16(1)	0(96)	0(96)	1(89)	0(99)	0(99)	2(97)
	$\frac{0.345}{1-0.95}$	1(41)	0(44)	1(38)	0(58)	0(62)	0(59)	0(74)	0(74)	0(72)
	$\frac{3.45}{1-0.50}$	0(98)	1(96)	8(86)	0(100)	0(100)	0(100)	0(100)	0(100)	0(100)
Negative	$\frac{0.0157}{1-0.95}$	10(0)	9(0)	14(0)	9(1)	8(0)	17(0)	10(1)	10(0)	21(0)
	$\frac{0.157}{1-0.50}$	11(0)	14(0)	32(0)	12(1)	20(0)	53(0)	9(8)	21(0)	39(0)
	$\frac{0.115}{1-0.95}$	6(0)	8(0)	20(0)	9(2)	17(0)	21(0)	7(20)	10(18)	10(17)
	$\frac{1.15}{1-0.50}$	14(0)	40(0)	46(0)	6(50)	31(12)	87(0)	0(99)	0(98)	1(98)
	$\frac{0.345}{1-0.95}$	8(0)	12(0)	26(0)	7(34)	9(28)	5(31)	1(74)	1(75)	3(72)
	$\frac{3.45}{1-0.50}$	4(0)	23(0)	46(0)	0(95)	0(94)	3(86)	0(100)	0(100)	0(100)

We first clarify the table. If we for instance look at the estimation results for reinforcement learning, 30 rounds, mixed payoffs, and $\lambda = 0.115$, then the table shows that only 16% of all simulations had a converged maximum likelihood estimator. In addition, in 3% of all simulations, estimation did not yield a parameter value at all (flat likelihood). In the other 81% of all simulations, estimation did result in a value, but the estimation had not yet converged (and given the number of retries was not going to converge any time soon).

Obviously, the first thing to notice from the table is that simulations did not converge often: on average only in 11% of the cases, whereas the total number of times estimation did not result in estimated values was on average 36%. The number of times that convergence was observed tends to increase with the number of rounds, though this did not hold in all cases. In addition, the number of times no parameters were re-estimated decreases over the number of rounds. Apparently, the results support the idea that more rounds provide better results (in line with the results from Cabrales & Garcia-Fontes (2000) and Blume et al. (2002) although "less bad" might be a more appropriate term than "better."

The best convergence results were found for negative payoffs (15%), then for mixed payoffs (10%), and the worst in terms of convergence were the games with positive payoffs (9%). These results are in line with our intuition because for negative payoffs reinforcement and belief learning lead to very different kinds of play.

Convergence occurred most often for $\delta = 0$ (14%), then for $\delta = 0.5$ (12%), and least often for $\delta = 1$ (7%). Comparing the results corresponding to the two ratios $\frac{\lambda}{1-\varphi}$ with the same value (so $\lambda = 0.0157$ and 0.157, $\lambda = 0.115$ and 1.15, and $\lambda = 0.345$ and 3.45), convergence and estimation results were always better for the ratio with the lowest value of λ and the highest value of φ. Hence, the results were better for $\lambda = 0.0157, 0.115, 0.345$ compared to $\lambda = 0.157, 1.15, 3.45$, respectively. Same values for $\frac{\lambda}{1-\varphi}$ signify they satisfy the stability criterion to the same extent. The lower the value of λ in this ratio, the longer it takes to satisfy the stability criterion during the course of the game. Hence these results suggest that the higher the probability to observe streaks soon, the worse the convergence properties and the estimation results are. Indeed, convergence and estimation results are poorest for $\lambda = 3.45$ in combination with $\delta = .05$ and $\delta = 1$; here convergence occurred in at most 4%, and no value could be estimated in at least 86% (!) of the simulations. Convergence and estimation results are best for $\lambda = 0.0157$ and $\lambda = 0.157$. Here, on average 99% of the values could be estimated. However, still only in 19% of these cases convergence occurred .

Table 4.5 The percentage of simulations in which convergence was achieved (between brackets how many times the program could not find a value of the parameters). All table entries concern the Pareto-optimal Nash equilibrium game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds.

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	11(1)	10(0)	15(0)	14(1)	13(0)	17(0)	11(0)	11(0)	12(0)
	$\frac{0.157}{1-0.50}$	9(16)	23(1)	44(0)	8(7)	20(0)	59(0)	8(1)	10(0)	16(0)
	$\frac{0.115}{1-0.95}$	3(24)	7(24)	2(24)	4(15)	3(14)	7(12)	4(5)	3(1)	14(1)
	$\frac{1.15}{1-0.50}$	0(49)	25(50)	2(50)	1(61)	10(62)	10(60)	0(82)	1(64)	6(21)
	$\frac{0.345}{1-0.95}$	0(48)	20(47)	0(51)	2(50)	9(50)	4(50)	1(42)	2(40)	4(40)
	$\frac{3.45}{1-0.50}$	0(50)	25(48)	0(51)	0(54)	25(52)	4(54)	0(100)	0(100)	0(100)
Mixed	$\frac{0.0157}{1-0.95}$	13(0)	16(0)	18(0)	13(0)	17(0)	17(0)	13(1)	17(0)	15(0)
	$\frac{0.157}{1-0.50}$	12(2)	20(0)	25(0)	13(2)	19(0)	29(0)	13(1)	15(0)	20(0)
	$\frac{0.115}{1-0.95}$	9(6)	6(2)	10(3)	8(6)	5(2)	14(1)	7(4)	6(2)	20(1)
	$\frac{1.15}{1-0.50}$	4(57)	6(58)	11(49)	5(78)	6(70)	19(46)	2(84)	4(65)	12(21)
	$\frac{0.345}{1-0.95}$	4(32)	3(35)	3(31)	3(38)	3(37)	5(36)	4(40)	5(38)	4(40)
	$\frac{3.45}{1-0.50}$	1(74)	1(73)	5(70)	3(97)	4(96)	3(96)	0(100)	0(100)	0(100)
Negative	$\frac{0.0157}{1-0.95}$	10(0)	8(0)	14(0)	10(0)	9(0)	16(0)	11(0)	11(0)	17(0)
	$\frac{0.157}{1-0.50}$	12(0)	15(0)	30(0)	11(1)	18(0)	40(0)	12(2)	18(0)	31(0)
	$\frac{0.115}{1-0.95}$	8(0)	10(0)	19(0)	11(1)	20(0)	31(0)	10(4)	17(1)	27(0)
	$\frac{1.15}{1-0.50}$	13(0)	33(0)	44(0)	19(4)	60(0)	98(0)	3(80)	8(66)	49(19)
	$\frac{0.345}{1-0.95}$	7(0)	14(0)	24(0)	13(4)	21(1)	35(0)	5(39)	8(37)	6(39)
	$\frac{3.45}{1-0.50}$	17(0)	5(0)	12(0)	16(44)	19(15)	91(0)	0(100)	0(100)	0(100)

Table 4.5 presents the convergence results for the Pareto-optimal Nash equilibrium game. Again, simulations did not converge often: only in 13% of the cases

(slightly better than the Prisoner's Dilemma game, which only converged in 11%), whereas the total number of times that no parameters could be estimated is 25% (compared to 36% for the Prisoner's Dilemma game). Again, the number of times that convergence was observed mostly increased over the number of rounds, and the number of times no parameters could be estimated decreased over the number of rounds.

As expected, the best numbers for convergence were for negative payoffs (21%), then for mixed and positive payoffs (9%). Convergence occurred most often for $\delta = 0.5$ (17%), then for $\delta = 0$ (13%), and least often for $\delta = 1$ (9%). For positive and mixed payoffs, $\lambda = 0.0157$ and $\lambda = 0.157$ resulted in convergence most often (up to 59%), and the number of times no parameters could be estimated were, in general, also lowest (mostly 0%, with some exceptions). The worst results were found for the highest λ (no results for 66% of the cases).

Table 4.6 The percentage of simulations in which convergence was achieved (between brackets how many times the program could not find a value of the parameters). All table entries concern the game with only a mixed strategy equilibrium with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds.

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	14(0)	16(0)	19(0)	13(0)	17(0)	19(0)	13(0)	16(0)	16(0)
	$\frac{0.157}{1-0.50}$	13(5)	30(0)	45(0)	17(2)	28(0)	65(0)	11(1)	20(0)	33(0)
	$\frac{0.115}{1-0.95}$	10(11)	8(9)	18(9)	12(7)	11(4)	64(0)	9(4)	13(1)	26(0)
	$\frac{1.15}{1-0.50}$	11(51)	10(50)	12(43)	12(30)	56(5)	99(0)	6(2)	32(0)	52(0)
	$\frac{0.345}{1-0.95}$	11(39)	10(36)	11(41)	10(32)	12(27)	30(19)	7(21)	13(7)	24(3)
	$\frac{3.45}{1-0.50}$	10(51)	10(49)	13(51)	5(41)	19(25)	78(1)	2(0)	3(0)	11(0)
Mixed	$\frac{0.0157}{1-0.95}$	30(0)	37(0)	49(0)	30(0)	37(0)	56(0)	29(0)	43(0)	64(0)
	$\frac{0.157}{1-0.50}$	34(1)	48(0)	56(0)	41(1)	58(0)	66(0)	43(1)	66(0)	76(0)
	$\frac{0.115}{1-0.95}$	30(1)	52(0)	91(0)	35(2)	56(0)	96(0)	32(3)	55(0)	94(0)
	$\frac{1.15}{1-0.50}$	45(4)	94(0)	99(0)	38(3)	92(0)	100(0)	33(2)	87(0)	100(0)
	$\frac{0.345}{1-0.95}$	31(8)	61(1)	90(0)	28(15)	59(4)	88(5)	28(18)	55(7)	74(12)
	$\frac{3.45}{1-0.50}$	28(1)	62(0)	100(0)	27(1)	37(0)	87(0)	30(0)	25(0)	58(0)
Negative	$\frac{0.0157}{1-0.95}$	14(0)	14(0)	13(0)	13(0)	15(0)	18(0)	14(0)	14(0)	18(0)
	$\frac{0.157}{1-0.50}$	15(0)	21(0)	35(0)	15(0)	26(0)	56(0)	14(1)	31(0)	42(0)
	$\frac{0.115}{1-0.95}$	14(0)	11(0)	22(0)	13(0)	19(0)	35(0)	11(3)	10(0)	31(0)
	$\frac{1.15}{1-0.50}$	13(0)	35(0)	51(0)	14(0)	70(0)	100(0)	3(3)	27(0)	50(0)
	$\frac{0.345}{1-0.95}$	12(0)	12(0)	20(0)	12(1)	17(0)	42(0)	5(19)	6(8)	19(48)
	$\frac{3.45}{1-0.50}$	13(0)	13(0)	40(0)	6(0)	21(0)	65(0)	0(0)	3(0)	16(0)

Table 4.6 presents the results for the game with only a mixed strategy equilibrium. Estimations converged more often than for the other games: in 32.4% of the cases (compared to 11% and 13% for the Prisoner's Dilemma game and Pareto-optimal Nash equilibrium game, respectively), whereas the total number of times no parameters could be estimated was 5.2% (compared to 36% and 25% for the Prisoner's Dilemma game and Pareto-optimal Nash equilibrium game, respectively).

Convergence increased, and number of times no parameters could be estimated in general decreased in the number of rounds.

Convergence occurred more often for $\delta = 0.5$ (40%) than for $\delta = 0$ and $\delta = 1$ (31% and 30%, respectively). Convergence was best for mixed payoffs (57%), and worse for positive and negative payoffs (21% and 23%, respectively). The best results can be found for $\lambda = 0.0157$ where all parameter could be estimated (finally!). Most converged cases can be found for $\lambda = 1.15$, where 50% of the estimations converged.

4.3.2 The Prisoner's Dilemma game

To get a better understanding where the best results may be expected, we take a closer look at the stability results from the previous chapter. We expect that the values for δ, λ, and φ can be estimated more accurately when the differences in terms of stability are clearest: we expect to be able to distinguish reinforcement and belief learning best for negative and mixed payoffs, expecting better estimations for δ as a result. This can also be seen from Table 4.7. Moreover, since more strategy changes are observed for reinforcement learning than for belief learning in games with negative payoffs and a pure Nash equilibrium, we expect better estimation results for games with negative and mixed payoffs.

Table 4.7 Stability results (with a probability of 99%) for the Prisoner's Dilemma game with different payoffs, and different values of (φ,λ). The Table shows for what parameter combination and for what game, which strategy combination was stable for $\delta = 0$ (reinforcement learning with a fixed reference point), $\delta = 0.5$, and $\delta = 1$ (belief learning).

	positive			mixed			negative		
$\frac{\lambda}{1-\varphi}$	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$
$\frac{0.0157}{1-0.95}$	.	.	.	.	.	.	.	.	.
$\frac{0.157}{1-0.50}$	.	.	.	.	.	.	.	.	.
$\frac{0.115}{1-0.95}$	12	2	2	.	.	2	.	.	2
$\frac{1.15}{1-0.50}$	12	2	2	.	.	2	.	.	2
$\frac{0.345}{1-0.95}$	1234	12	2	1	2	2	.	.	2
$\frac{3.45}{1-0.50}$	1234	12	2	1	2	2	.	.	2

1 = top-left strategy combination; 2 = bottom-right strategy combination; 3 = top-right strategy combination; 4 = bottom-left strategy combination.

The results for the Prisoner's Dilemma game can be found in Table 4.8. When the 98% frequency interval does not contain zero (the true parameter value), then the corresponding entry in the table is in italics. Note that results are only presented for those simulations in which we had an estimate at all, either after convergence was reached or before convergence was reached. Simulations in which no value could be estimated were excluded. We have to keep in mind that by ignoring those simulations without an estimate, we step over the fact that the results are worse than we portray them from now on. Changing the convergence criterium leads to different numbers than presented here, but these results are not improving the estimation value.

Table 4.8 The average difference between the re-estimated δ and δ ($\hat{\delta} - \delta$) for the Prisoner's Dilemma game with positive, mixed, and negative payoffs under $\delta = 0, 0.5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did <u>not</u> have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.63	0.77	0.81	0.17	0.33	0.40	−0.30	−0.10	−0.06
	$\frac{0.157}{1-0.50}$	0.62	0.57	0.28	0.28	0.30	0.15	−0.12	−0.07	−0.07
	$\frac{0.115}{1-0.95}$	0.71	0.62	0.61	0.35	0.29	0.29	−0.11	−0.13	−0.12
	$\frac{1.15}{1-0.50}$	0.03	0.04	0.06	−0.09	−0.07	−0.07	−0.29	−0.22	−0.04
	$\frac{0.345}{1-0.95}$	0.29	0.30	0.29	0.14	0.13	0.14	−0.20	−0.22	−0.20
	$\frac{3.45}{1-0.50}$	0.00	0.00	0.00	−0.04	−0.05	*−0.05*	.	.	.
Mixed	$\frac{0.0157}{1-0.95}$	0.38	0.44	0.53	−0.14	0.00	0.08	−0.62	−0.47	−0.36
	$\frac{0.157}{1-0.50}$	0.43	0.41	0.31	−0.00	0.03	0.01	−0.42	−0.37	−0.29
	$\frac{0.115}{1-0.95}$	0.50	0.47	0.46	0.05	0.02	0.05	−0.45	−0.47	−0.46
	$\frac{1.15}{1-0.50}$	0.08	0.06	0.05	−0.24	−0.21	−0.10	−0.50	−0.36	−0.28
	$\frac{0.345}{1-0.95}$	0.39	0.40	0.40	−0.07	−0.07	−0.11	−0.55	−0.56	−0.55
	$\frac{3.45}{1-0.50}$	0.02	0.05	0.00	.	.	.	.	.	.
Negative	$\frac{0.0157}{1-0.95}$	0.15	0.09	0.11	−0.33	−0.37	−0.39	−0.82	−0.79	−0.69
	$\frac{0.157}{1-0.50}$	0.06	0.04	0.04	−0.34	−0.36	−0.29	−0.73	−0.67	−0.42
	$\frac{0.115}{1-0.95}$	0.11	0.03	0.02	−0.23	−0.06	0.01	−0.60	−0.43	−0.43
	$\frac{1.15}{1-0.50}$	0.05	0.06	0.03	−0.19	−0.10	−0.03	−0.46	−0.46	−0.37
	$\frac{0.345}{1-0.95}$	0.06	0.02	0.01	−0.10	0.02	−0.01	−0.43	−0.43	−0.44
	$\frac{3.45}{1-0.50}$	0.04	0.08	0.04	*−0.15*	−0.10	−0.05	.	*−0.20*	*−0.13*

As the table shows, increasing the number of rounds made the bias (that is, the difference between the estimates and the true values) smaller. In general, we observe that the estimates of δ were less biased if payoffs are positive or negative (on average 0.24) than when they were mixed (on average 0.28). Though the exact results differ per parameter combination we generally see that δ and $\hat{\delta}$ were much more alike for $\delta = 0.5$ (difference of 0.15 on average) than for $\delta = 0$ (on average 0.24) and for $\delta = 1$ (difference of 0.37 on average). Best results were for $\lambda = 3.45$ where the average absolute bias was 0.06, though this is also the case where often no values were estimated at all. For $\lambda = 0.0157$ the bias was largest (average of 0.38). For reinforcement learning, we observe that the estimates of δ were less biased if payoffs are negative (average of 0.06) than when they were positive or mixed (0.37 and 0.30 on average). Recall that for reinforcement learning ($\delta = 0$) and negative payoffs, there were no stable streaks and many strategy changes. These results suggest that the estimates are less biased when there were more strategy changes (as opposed to streaks). However, for belief learning the estimates were less biased in case of positive payoffs (on average 0.15) than when payoffs were negative (on average 0.50).

Table 4.9 The average difference between the re-estimated φ and φ ($\hat{\varphi} - \varphi$) for the Prisoner's Dilemma game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. Numbers in italics were <u>not</u> in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.41	0.24	0.05	0.38	0.22	0.07	0.36	0.18	0.10
	$\frac{0.157}{1-0.50}$	−0.09	−0.02	−0.03	−0.09	−0.04	−0.01	−0.08	−0.02	−0.03
	$\frac{0.115}{1-0.95}$	0.10	−0.01	−0.02	0.12	0.01	−0.01	0.14	0.03	−0.01
	$\frac{1.15}{1-0.50}$	−0.33	−0.25	−0.19	−0.26	−0.29	−0.31	−0.42	−0.24	0.15
	$\frac{0.345}{1-0.95}$	0.01	−0.01	0.01	0.11	0.10	0.06	−0.04	−0.04	*−0.04*
	$\frac{3.45}{1-0.50}$	*−0.50*	*−0.49*	−0.44	−0.45	−0.46	−0.46	.	.	.
Mixed	$\frac{0.0157}{1-0.95}$	0.46	0.30	0.15	0.45	0.24	0.14	0.43	0.23	0.10
	$\frac{0.157}{1-0.50}$	−0.02	−0.03	−0.06	−0.01	−0.03	−0.03	−0.06	0.01	−0.03
	$\frac{0.115}{1-0.95}$	0.29	0.10	0.02	0.18	0.06	0.01	0.14	0.04	−0.00
	$\frac{1.15}{1-0.50}$	−0.15	0.02	0.08	−0.45	−0.26	0.01	−0.40	−0.29	0.07
	$\frac{0.345}{1-0.95}$	0.05	−0.03	−0.03	0.01	*−0.05*	*−0.05*	−0.03	*−0.05*	*−0.05*
	$\frac{3.45}{1-0.50}$	−0.36	−0.10	0.01	.	.	.	.	.	.
Negative	$\frac{0.0157}{1-0.95}$	0.36	0.20	0.12	0.33	0.15	0.02	0.34	0.11	0.11
	$\frac{0.157}{1-0.50}$	−0.07	−0.05	−0.00	−0.16	−0.21	−0.22	−0.23	−0.20	−0.09
	$\frac{0.115}{1-0.95}$	0.20	0.06	0.02	0.13	0.23	0.29	0.11	0.16	0.15
	$\frac{1.15}{1-0.50}$	0.05	0.07	0.04	−0.22	−0.08	−0.03	−0.43	−0.32	−0.17
	$\frac{0.345}{1-0.95}$	0.13	0.04	0.01	0.10	0.21	0.14	0.05	0.04	0.04
	$\frac{3.45}{1-0.50}$	0.25	0.15	0.05	−0.41	−0.18	0.11	.	*0.23*	*0.34*

Table 4.9 shows the results concerning φ. Note that the differences between φ and $\hat{\varphi}$ are smaller than for δ and $\hat{\delta}$. Increasing the number of rounds made the difference smaller as well. For mixed payoffs, the bias was lowest (on average 0.13). The bias in case of negative payoffs was lower (on average 0.15) than for positive payoffs (on average 0.17). Though once more the exact results differ per parameter combination, we tend to see that φ and $\hat{\varphi}$ were more alike for $\delta = 0$ (difference of 0.14 on average) than for $\delta = 1$ (difference of 0.15 on average) and for $\delta = 0.5$ (difference of 0.17 on average). The bias also changes much over the value of λ. The least bias was found for $\lambda = 0.115$ where we find an average of 0.07. The most bias was for $\lambda = 3.45$ where the average value was 0.29.

Table 4.10 The average difference between the re-estimated λ and λ ($\hat{\lambda} - \lambda$) for the Prisoner's Dilemma game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.03	0.02	$> 10^4$	0.04	0.02	$> 10^4$	0.04	0.03
	$\frac{0.157}{1-0.50}$	$> 10^4$	0.05	0.00	$> 10^4$	0.03	0.00	$> 10^4$	0.02	−0.01
	$\frac{0.115}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	144.56	$> 10^4$	$> 10^4$	580.81	−0.38	−0.20
	$\frac{0.345}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	*-1.24*	*-1.26*	−1.00	.	.	.
Mixed	$\frac{0.0157}{1-0.95}$	0.10	0.05	0.03	0.11	0.06	0.04	0.14	0.08	0.04
	$\frac{0.157}{1-0.50}$	0.05	0.00	−0.03	0.11	0.04	0.00	0.13	0.09	0.02
	$\frac{0.115}{1-0.95}$	0.14	0.09	0.05	0.17	0.14	0.11	0.19	0.16	0.13
	$\frac{1.15}{1-0.50}$	−0.38	0.00	0.30	−0.64	−0.44	0.01	*-0.70*	*-0.64*	−0.26
	$\frac{0.345}{1-0.95}$	0.02	−0.04	−0.00	0.07	0.01	0.01	0.05	0.04	0.03
	$\frac{3.45}{1-0.50}$	*-2.70*	*-1.90*	−0.89	.	.	.	.	.	.
Negative	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.06	0.02	$> 10^4$	0.08	0.02	$> 10^4$	0.11	0.07
	$\frac{0.157}{1-0.50}$	$> 10^4$	0.06	0.01	$> 10^4$	0.06	−0.04	$> 10^4$	$> 10^4$	0.03
	$\frac{0.115}{1-0.95}$	$> 10^4$	0.09	0.02	$> 10^4$	$> 10^4$	0.35	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	0.17	$> 10^4$	$> 10^4$	52.48	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{0.345}{1-0.95}$	$> 10^4$	1483.73	0.02	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	760.35	$> 10^4$	$> 10^4$	$> 10^4$	.	*4.63*	*5.02*

Table 4.10 presents the bias of the estimates for λ. All biases were positive and many were dramatic. Many cases were in the order of magnitude of 10^{19}, suggesting a "runaway parameter" that is indicative of poor (or no) real convergence. In the table, instead of reporting the exact value of λ in such cases, we report "$> 10^4$". This basically shows that in these cases the estimates did not converge.

In general, increasing the number of rounds made the bias smaller. We tend to find the least bias for mixed payoffs (no very large λ's at all). For positive and negative payoffs, we find the most bias. There are not many differences in bias between the three types of learning (reinforcement, belief, and mixed learning): apparently the bias differed more across the payoffs than across the values of δ. The best estimates were found for $\lambda = 0.0157$ and $\lambda = 0.157$ and the results worsen for increasing λ.

To summarize, estimation of parameters for the Prisoner's Dilemma game was generally poor; parameter estimates were biased in most conditions. In the game with positive and mixed payoffs, estimation of δ for reinforcement learning has proven to be more difficult than in case of negative payoffs, which was in line with our expectations. Re-estimating φ proves to be difficult, though the results are least biased for $\lambda = .115$, $\lambda = 0.345$, and $\lambda = .157$. The difference between λ and $\hat{\lambda}$ was smallest in case of mixed payoffs. As expected, increasing the number of rounds decreased the bias of the estimates, but even then the biases were still considerable.

4.3.3 The Pareto-optimal Nash equilibrium game

Using the results of the previous chapter, we expect that the bias of estimates of δ, λ, and φ is less when the differences in terms of stability (table 4.11) and strategy changes are clearest: we expect to be able to distinguish reinforcement and belief learning best for negative and mixed payoffs, expecting better estimations for δ as a result.

Table 4.11 Stability results (with a probability of 99%) for the Pareto-optimal Nash equilibrium game with different payoffs, and different values of (φ,λ). The Table shows for what parameter combination and for what game, which strategy combination was stable for $\delta = 0$ (reinforcement learning with a fixed reference point), $\delta = 0.5$, and $\delta = 1$ (belief learning).

	positive			mixed			negative		
	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$
$\frac{0.0157}{1-0.95}$	.	.	.	.	.	.	.	.	.
$\frac{0.157}{1-0.50}$	.	.	.	.	.	.	.	.	.
$\frac{0.115}{1-0.95}$	1234	1	1	1	1	1	.	.	1
$\frac{1.15}{1-0.50}$	1234	1	1	1	1	1	.	.	1
$\frac{0.345}{1-0.95}$	1234	1234	1	134	1	1	.	1	1
$\frac{3.45}{1-0.50}$	1234	1234	1	134	1	1	.	1	1

1 = top-left strategy combination; 2 = bottom-right strategy combination; 3 = top-right strategy combination; 4 = bottom-left strategy combination.

Table 4.12 The average difference between the re-estimated δ and δ ($\hat{\delta} - \delta$) for the Pareto-optimal Nash equilibrium game with positive, mixed, and negative payoffs under $\delta = 0, 0.5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.64	0.70	0.73	0.16	0.26	0.40	−0.32	−0.18	−0.05
	$\frac{0.157}{1-0.50}$	0.52	0.48	0.24	0.22	0.22	0.12	−0.17	−0.07	−0.05
	$\frac{0.115}{1-0.95}$	0.58	0.55	0.51	0.32	0.34	0.33	−0.08	−0.05	−0.04
	$\frac{1.15}{1-0.50}$	0.01	0.01	0.06	−0.21	−0.09	0.08	−0.06	−0.03	−0.02
	$\frac{0.345}{1-0.95}$	0.16	0.19	0.18	0.13	0.12	0.14	−0.07	−0.08	−0.08
	$\frac{3.45}{1-0.50}$	0.00	0.00	0.03	−0.44	−0.37	−0.30	−0.00	.	*-0.27*
Mixed	$\frac{0.0157}{1-0.95}$	0.50	0.58	0.72	0.02	0.13	0.30	−0.44	−0.27	−0.13
	$\frac{0.157}{1-0.50}$	0.55	0.56	0.46	0.14	0.16	0.18	−0.29	−0.21	−0.14
	$\frac{0.115}{1-0.95}$	0.61	0.61	0.55	0.24	0.22	0.25	−0.20	−0.20	−0.15
	$\frac{1.15}{1-0.50}$	0.24	0.25	0.34	0.14	0.19	0.32	−0.13	−0.07	−0.04
	$\frac{0.345}{1-0.95}$	0.55	0.47	0.47	0.21	0.14	0.14	−0.22	−0.21	−0.19
	$\frac{3.45}{1-0.50}$	0.04	0.06	0.14	−0.12	−0.09	−0.09	−0.00	−0.00	*-0.64*
Negative	$\frac{0.0157}{1-0.95}$	0.21	0.10	0.10	−0.31	−0.36	−0.36	−0.80	−0.83	−0.78
	$\frac{0.157}{1-0.50}$	0.14	0.05	0.02	−0.25	−0.15	0.00	−0.62	−0.36	−0.35
	$\frac{0.115}{1-0.95}$	0.12	0.06	0.06	−0.32	−0.38	−0.33	−0.72	−0.71	−0.63
	$\frac{1.15}{1-0.50}$	0.09	0.03	0.01	−0.11	0.04	0.02	−0.45	−0.34	−0.33
	$\frac{0.345}{1-0.95}$	0.08	0.06	0.03	−0.20	−0.13	−0.03	−0.45	−0.41	−0.23
	$\frac{3.45}{1-0.50}$	0.02	0.05	0.08	−0.10	−0.08	−0.02	*-0.22*	*-0.07*	−0.24

The results for the Pareto-optimal Nash equilibrium game can be found in Table 4.12. In general, we observe that the estimates of δ were less biased for positive payoffs (an average difference of 0.21) than for negative payoffs (an average difference of 0.24) and mixed payoffs (an average difference of 0.27). Though the exact results differed per parameter combination, on average, we generally see that bias was lower for $\delta = 0.5$ (difference of 0.20 on average) than for $\delta = 0$ and $\delta = 1$ (difference of 0.27 and 0.26 on average). The lowest average bias was for $\lambda = 3.45$ (0.13) and the highest average bias was for $\lambda = 0.0157$ (0.38). Note that the differences in payoffs depend heavily on the δ (or the other way around). The lowest bias can be found in case of reinforcement learning with negative payoffs: the average bias was only 0.07. This is in accordance with our expectations, since there reinforcement learners keep changing strategy whereas belief learners increase their play of the Nash equilibrium in the course of the game. Another small bias (0.10 on average) can be found for positive payoffs with belief learning.

Table 4.13 The average difference between the re-estimated φ and φ ($\hat{\varphi} - \varphi$) for the Pareto-optimal Nash equilibrium game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.38	0.27	0.07	0.37	0.27	0.10	0.36	0.25	0.14
	$\frac{0.157}{1-0.50}$	−0.07	0.00	−0.02	−0.05	−0.06	−0.10	−0.07	−0.10	−0.13
	$\frac{0.115}{1-0.95}$	0.11	0.03	0.02	0.17	0.03	−0.01	0.23	0.11	0.06
	$\frac{1.15}{1-0.50}$	−0.49	−0.29	0.03	−0.21	0.11	0.29	−0.25	0.05	0.29
	$\frac{0.345}{1-0.95}$	−0.02	0.13	0.32	0.06	0.11	0.14	0.07	0.01	−0.02
	$\frac{3.45}{1-0.50}$	*-0.50*	*-0.31*	−0.03	−0.42	−0.06	0.35	*-0.50*	.	*-0.50*
Mixed	$\frac{0.0157}{1-0.95}$	0.46	0.37	0.17	0.43	0.33	0.16	0.41	0.30	0.15
	$\frac{0.157}{1-0.50}$	−0.01	−0.05	−0.11	−0.03	−0.06	−0.16	−0.03	−0.05	−0.12
	$\frac{0.115}{1-0.95}$	0.36	0.08	0.03	0.25	0.09	0.04	0.27	0.12	0.09
	$\frac{1.15}{1-0.50}$	−0.17	−0.07	−0.01	−0.32	−0.22	0.17	−0.22	−0.00	0.31
	$\frac{0.345}{1-0.95}$	0.10	0.02	0.01	0.07	0.00	0.01	0.07	−0.00	−0.01
	$\frac{3.45}{1-0.50}$	−0.47	−0.44	−0.20	−0.47	*-0.50*	−0.36	*-0.50*	*-0.50*	*-0.50*
Negative	$\frac{0.0157}{1-0.95}$	0.33	0.22	0.12	0.32	0.18	0.06	0.32	0.17	0.05
	$\frac{0.157}{1-0.50}$	−0.10	−0.11	−0.05	−0.19	−0.23	−0.24	−0.17	−0.23	−0.26
	$\frac{0.115}{1-0.95}$	0.19	0.07	0.02	0.15	0.15	0.26	0.18	0.29	0.26
	$\frac{1.15}{1-0.50}$	0.00	0.03	0.01	−0.20	−0.14	−0.04	−0.12	0.00	0.19
	$\frac{0.345}{1-0.95}$	0.13	0.04	0.01	0.18	0.32	0.27	0.15	0.21	0.16
	$\frac{3.45}{1-0.50}$	−0.05	0.01	0.03	−0.12	−0.12	−0.02	*0.27*	*0.38*	*-0.50*

Table 4.13 shows the results concerning φ. First note that increasing the number of rounds decreased bias. Note also that bias for φ was less than for δ. For negative payoffs, the bias was lowest on average (0.16). The bias in case of positive payoffs was lower (on average 0.17) than for mixed payoffs (on average 0.19). Moreover, bias was lower for $\delta = 0$ (difference of 0.14 on average) than for $\delta = 0.5$ (difference of 0.18 on average) and for $\delta = 1$ (difference of 0.20 on average). The lowest average bias was for $\lambda = 0.345$ (0.10) and the highest was for $\lambda = 3.45$ (0.31). The bias was smallest in case of $\delta = 0$ and negative payoffs (average difference of 0.08), which we expected and is in line with the findings on the bias for δ and is therefore.

Table 4.14 The average difference between the re-estimated λ and λ ($\hat{\lambda} - \lambda$) for the Pareto-optimal Nash equilibrium game with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.07	0.03	$> 10^4$	0.07	0.04	$> 10^4$	0.06	0.04
	$\frac{0.157}{1-0.50}$	$> 10^4$	$> 10^4$	0.02	$> 10^4$	0.05	-0.02	$> 10^4$	-0.01	-0.04
	$\frac{0.115}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	0.81	0.13
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	78.05	$> 10^4$	$> 10^4$	0.66
	$\frac{0.345}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> \mathit{10}^{\mathit{4}}$	$> \mathit{10}^{\mathit{4}}$	$> \mathit{10}^{\mathit{4}}$	$> \mathit{10}^{\mathit{4}}$	$> 10^4$	$> \mathit{10}^{\mathit{4}}$	*-3.07*	.	*1.87*
Mixed	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.07	0.03	$> 10^4$	0.08	0.04	$> 10^4$	0.08	0.05
	$\frac{0.157}{1-0.50}$	$> 10^4$	0.01	-0.03	1.10	0.01	-0.04	$> 10^4$	0.03	-0.03
	$\frac{0.115}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	80.08	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	0.59	$> 10^4$	0.69
	$\frac{0.345}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	1.99	1.78	1.82	*-3.01*	*-3.01*	*1.87*
Negative	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.09	0.02	$> 10^4$	0.10	0.03	$> 10^4$	0.13	0.05
	$\frac{0.157}{1-0.50}$	$> 10^4$	0.08	0.00	$> 10^4$	0.08	-0.05	$> 10^4$	0.11	-0.04
	$\frac{0.115}{1-0.95}$	$> 10^4$	0.10	0.02	$> 10^4$	0.34	0.28	$> 10^4$	$> 10^4$	1153.78
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	1.24	$> 10^4$	4732.09	0.05	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{0.345}{1-0.95}$	$> 10^4$	0.17	0.02	$> 10^4$	$> 10^4$	1.03	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	910.32	*-15.73*	4.93	588.44

Table 4.14 shows the results for the estimated λ in the Pareto-optimal Nash equilibrium game. In general, bias decreased in the number of rounds. We tend to find the least bias for negative payoffs (least occurrence of very large λ). For positive and negative payoffs, we find the most bias. There were not many differences in bias between reinforcement, mixed learning or belief learning: apparently the bias differed more across the payoffs than across the value of δ. It appears that the least bias was found for the lower λ's, and that bias increased in λ.

To summarize, estimation of parameters for the Pareto-optimal Nash equilibrium game was generally poor, but better on average than for the Prisoner's Dilemma game; parameter estimates were biased in most conditions. In accordance to our expectations, for the game with positive and mixed payoffs, estimation of δ for reinforcement learning was more difficult than in case of negative payoffs. Estimating φ proved to be difficult as well and varied much over the different parameter combinations, but was also best under reinforcement learning with negative payoffs. The bias for λ was also lowest in case of negative payoffs. Estimating λ, reinforcement learning with negative payoffs performs relatively well, compared to other values of δ under the same conditions. As expected, increasing the number of rounds decreased the bias of the estimates, but even then the biases were still considerable.

4.3.4 The game with only a mixed strategy equilibrium

Table 4.15 Stability results (with a probability of 99%) for the game with only a mixed strategy equilibrium with different payoffs, and different values of (φ,λ). The Table shows for what parameter combination and for what game, which strategy combination was stable for $\delta = 0$ (reinforcement learning with a fixed reference point), $\delta = 0.5$, and $\delta = 1$ (belief learning).

	positive			mixed			negative		
	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$	$\delta = 0$	$\delta = .5$	$\delta = 1$
$\frac{0.0157}{1-0.95}$	.	.	.	.	.	.	.	.	.
$\frac{0.157}{1-0.50}$	.	.	.	.	.	.	.	.	.
$\frac{0.115}{1-0.95}$	14	.	.	.	.	.	.	.	.
$\frac{1.15}{1-0.50}$	14	.	.	.	.	.	.	.	.
$\frac{0.345}{1-0.95}$	1234	.	.	.	.	.	.	.	.
$\frac{3.45}{1-0.50}$	1234	.	.	.	.	.	.	.	.

1 = top-left strategy combination; 2 = bottom-right strategy combination; 3 = top-right strategy combination; 4 = bottom-left strategy combination.

Table 4.16 The average difference between the re-estimated δ and δ ($\hat{\delta} - \delta$) for the game with only a mixed strategy equilibrium with positive, mixed, and negative payoffs under $\delta = 0, 0.5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.59	0.64	0.69	0.12	0.19	0.32	−0.38	−0.26	−0.09
	$\frac{0.157}{1-0.50}$	0.56	0.50	0.28	0.16	0.20	0.14	−0.24	−0.14	−0.09
	$\frac{0.115}{1-0.95}$	0.60	0.50	0.43	0.19	0.21	0.19	−0.20	−0.12	−0.08
	$\frac{1.15}{1-0.50}$	0.18	0.23	0.28	0.22	0.11	0.02	−0.12	−0.21	−0.09
	$\frac{0.345}{1-0.95}$	0.38	0.35	0.37	0.14	0.13	0.12	−0.17	−0.13	−0.09
	$\frac{3.45}{1-0.50}$	0.10	0.10	0.11	0.24	0.17	0.04	−0.08	−0.07	−0.16
Mixed	$\frac{0.0157}{1-0.95}$	0.34	0.38	0.50	−0.16	−0.10	−0.08	−0.66	−0.58	−0.69
	$\frac{0.157}{1-0.50}$	0.35	0.39	0.57	−0.13	−0.08	0.01	−0.64	−0.63	−0.52
	$\frac{0.115}{1-0.95}$	0.38	0.36	0.64	−0.09	−0.08	0.39	−0.58	−0.50	−0.39
	$\frac{1.15}{1-0.50}$	0.40	0.43	0.34	−0.04	0.13	0.12	−0.47	−0.27	−0.15
	$\frac{0.345}{1-0.95}$	0.40	0.63	0.40	−0.05	0.22	0.14	−0.54	−0.30	−0.20
	$\frac{3.45}{1-0.50}$	0.53	*0.76*	*0.93*	0.06	0.24	0.40	−0.47	−0.27	−0.14
Negative	$\frac{0.0157}{1-0.95}$	0.31	0.20	0.08	−0.18	−0.28	−0.37	−0.71	−0.72	−0.72
	$\frac{0.157}{1-0.50}$	0.22	0.12	0.08	−0.20	−0.26	−0.21	−0.58	−0.57	−0.45
	$\frac{0.115}{1-0.95}$	0.21	0.08	0.03	−0.16	−0.17	−0.12	−0.49	−0.30	−0.23
	$\frac{1.15}{1-0.50}$	0.23	0.11	0.03	0.06	−0.03	−0.01	−0.11	−0.17	−0.06
	$\frac{0.345}{1-0.95}$	0.15	0.04	0.01	−0.03	−0.04	−0.05	−0.24	−0.14	−0.14
	$\frac{3.45}{1-0.50}$	0.24	0.26	0.07	0.30	0.13	−0.03	−0.05	−0.02	−0.04

Table 4.15 presents the results concerning stability of the mixed strategy equilibrium. Using results on stability we expect to be able to estimate δ most accurately in case of positive payoffs. Using results on strategy changes, we expect less bias in case of negative payoffs since then more strategy changes are expected under reinforcement learning than under belief learning. The results on δ for the game with only a mixed strategy equilibrium can be found in Table 4.16. In general, we observe that the estimates of δ were less biased if payoffs are negative (on average 0.20) or positive (on average 0.23) than when they were mixed (0.36). Though the exact results differ per parameter combination, we generally see that bias was lowest for $\delta = 0.5$ (on average 0.15) than for $\delta = 0$ and $\delta = 1$ (on average 0.34 and 0.30, respectively). For $\lambda = 0.345$ we have, on average, the least bias (0.17) and for $\lambda = 0.0157$ the most (0.38). The best results were for $\delta = 0$ and negative payoffs, where bias was 0.14, and for $\delta = 1$ and positive payoffs (0.15).

Table 4.17 The average difference between the re-estimated φ and φ ($\hat{\varphi} - \varphi$) for the game with only a mixed strategy equilibrium with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	0.34	0.22	0.08	0.35	0.23	0.06	0.35	0.25	0.07
	$\frac{0.157}{1-0.50}$	−0.07	−0.10	−0.09	−0.05	−0.07	−0.08	−0.04	−0.06	−0.05
	$\frac{0.115}{1-0.95}$	0.19	0.03	−0.01	0.23	0.03	−0.01	0.23	0.05	0.01
	$\frac{1.15}{1-0.50}$	−0.16	−0.01	−0.11	−0.11	0.01	0.02	0.12	0.19	0.07
	$\frac{0.345}{1-0.95}$	0.18	0.15	0.14	0.17	0.06	−0.02	0.21	0.06	0.01
	$\frac{3.45}{1-0.50}$	−0.39	−0.41	−0.31	−0.13	−0.11	0.06	0.10	0.11	0.16
Mixed	$\frac{0.0157}{1-0.95}$	0.40	0.32	0.12	0.40	0.29	0.07	0.39	0.28	0.05
	$\frac{0.157}{1-0.50}$	−0.06	−0.09	−0.16	−0.04	−0.07	−0.09	−0.04	−0.05	−0.04
	$\frac{0.115}{1-0.95}$	0.30	0.08	0.00	0.25	0.04	0.00	0.21	0.01	0.00
	$\frac{1.15}{1-0.50}$	0.00	0.01	0.00	0.01	0.01	0.00	0.00	−0.00	0.00
	$\frac{0.345}{1-0.95}$	0.16	0.01	−0.00	0.14	−0.00	−0.00	0.13	−0.01	−0.01
	$\frac{3.45}{1-0.50}$	0.00	0.00	0.00	0.00	−0.00	0.00	0.01	−0.01	−0.00
Negative	$\frac{0.0157}{1-0.95}$	0.26	0.18	0.08	0.26	0.14	0.03	0.25	0.15	−0.00
	$\frac{0.157}{1-0.50}$	−0.14	−0.16	−0.07	−0.18	−0.20	−0.16	−0.19	−0.22	−0.21
	$\frac{0.115}{1-0.95}$	0.18	0.07	0.02	0.16	0.01	−0.02	0.15	−0.02	−0.03
	$\frac{1.15}{1-0.50}$	0.13	0.09	0.03	0.10	−0.04	−0.01	−0.06	−0.11	−0.04
	$\frac{0.345}{1-0.95}$	0.16	0.04	0.00	0.11	−0.00	−0.02	0.10	−0.03	−0.02
	$\frac{3.45}{1-0.50}$	0.22	0.20	0.06	0.37	0.26	0.08	−0.04	−0.01	−0.03

Table 4.17 shows the results concerning φ. Bias decreased in the number of rounds. Bias was lower for φ than for δ. As opposed to the bias for δ that was highest for mixed payoffs, the bias for φ was lowest on average for mixed payoffs (0.08). The bias in case of negative payoffs was lower on average (0.11) than for positive payoffs (0.13). Though once more the exact results differed per parameter combination, we

tend to see bias was lower for for $\delta = 1$ and $\delta = 0.5$ (0.09 and 0.10, respectively) than for $\delta = 0$ (0.13). The lowest bias was 0.05 for $\lambda = 1.15$ and the highest was 0.21 for $\lambda = 0.0157$.

Table 4.18 The average difference between the re-estimated λ and λ ($\hat{\lambda} - \lambda$) for the game with only a mixed strategy equilibrium with positive and mixed payoffs under $\delta = 0, .5, 1$ and after 10, 30, and 150 rounds. The numbers in italics did not have zero in the 98% frequency interval. In case no estimations were done, the table denotes a ".".

		$\delta = 0$			$\delta = .5$			$\delta = 1$		
Payoffs	$\frac{\lambda}{1-\varphi}$	10	30	150	10	30	150	10	30	150
Positive	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.06	0.01	$> 10^4$	0.05	0.01	$> 10^4$	0.05	0.01
	$\frac{0.157}{1-0.50}$	$> 10^4$	7.04	-0.02	$> 10^4$	0.03	-0.02	$> 10^4$	0.02	-0.01
	$\frac{0.115}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	17.99	$> 10^4$	$> 10^4$	0.02
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	1866.10	$> 10^4$	$> 10^4$	36.89
	$\frac{0.345}{1-0.95}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
Mixed	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.07	0.01	$> 10^4$	0.07	0.02	$> 10^4$	0.08	0.02
	$\frac{0.157}{1-0.50}$	$> 10^4$	-0.01	-0.08	$> 10^4$	0.05	-0.01	$> 10^4$	0.12	0.06
	$\frac{0.115}{1-0.95}$	$> 10^4$	0.03	-0.04	$> 10^4$	0.08	-0.01	$> 10^4$	719.08	0.05
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	-0.26	$> 10^4$	$> 10^4$	-0.04	$> 10^4$	$> 10^4$	0.17
	$\frac{0.345}{1-0.95}$	$> 10^4$	$> 10^4$	-0.06	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	-1.22	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$
Negative	$\frac{0.0157}{1-0.95}$	$> 10^4$	0.09	0.02	$> 10^4$	0.10	0.02	$> 10^4$	0.11	0.02
	$\frac{0.157}{1-0.50}$	$> 10^4$	0.09	0.00	$> 10^4$	0.08	-0.02	$> 10^4$	0.09	-0.02
	$\frac{0.115}{1-0.95}$	$> 10^4$	0.09	0.02	$> 10^4$	0.12	0.03	$> 10^4$	18.20	0.04
	$\frac{1.15}{1-0.50}$	$> 10^4$	$> 10^4$	0.09	$> 10^4$	$> 10^4$	0.15	$> 10^4$	$> 10^4$	0.10
	$\frac{0.345}{1-0.95}$	$> 10^4$	0.12	0.02	$> 10^4$	1.39	0.04	$> 10^4$	$> 10^4$	2264.21
	$\frac{3.45}{1-0.50}$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$

Finally, Table 4.18 presents the results for the estimated λ in the game with only a mixed strategy equilibrium. Again, bias decreased in the number of rounds. The least bias is observed for negative payoffs, whereas bias was larger for positive than for mixed payoffs. There are not many differences in bias between reinforcement, mixed learning or belief learning, but estimation under belief learning appears to be performing slightly worse than under the other two values of δ. In general it holds that the lower the value of λ, the lower the bias.

To summarize, estimation of parameters for the game with only a mixed strategy equilibrium was generally best for all games. However, parameter estimates were still biased in most conditions. The best results were found for reinforcement learning with negative payoffs or for belief learning with positive payoffs. Estimating φ proved to be difficult as well and varied much over the different parameter combinations. Re-estimating φ proves to be difficult and varies much over the different parameter combinations. The difference between λ and $\hat{\lambda}$ was smallest in case of negative payoffs, then for mixed payoffs, and biggest for positive payoffs.

As expected, increasing the number of rounds decreased the bias of the estimates, but even then the biases were still considerable.

4.4 Conclusions and discussion

In the previous chapters we used stable strategy combinations to conclude that we *can* differentiate between EWA-based reinforcement and belief learning in 2 × 2 games. Simulation data generated by EWA-based reinforcement learners and EWA-based belief learners can be discerned by looking at the presence of streaks of the same strategy combination. In addition, other characteristics (how often players change strategy during the game, and how often the different kinds of outcomes occur during the cause of the game) can be used to further differentiate these two types of learning. In addition, we found that EWA-based reinforcement learning and belief learning can be differentiated in an analytical way using stability, even in 2 × 2 games.

Using results on stability and strategy changes we formulated predictions on the conditions in which estimation will be more accurate. Our predictions were that estimation is relatively accurate for mixed payoffs in the PD and in general for negative payoffs, and inaccurate for positive payoffs in the NE game. Although convergence is as expected better for negative payoffs in all games, the predominant conclusion is that estimation is poor in all conditions. Most of the time estimation did not converge, and when estimation converged the estimates were biased most of the time, even in games with 150 rounds. It appears that one of the main problems is that the sensitivity parameter λ could not be estimated, with its estimated value increasing in each iteration leading to ridiculously high values. This occurred often, in all conditions except the PD with mixed payoffs. Apparently the likelihood is almost flat with different parameter combinations leading to approximately the same fit, which makes estimation very difficult and leads to biased parameter estimates.

One important observation is that the concept of stability and the presence of streaks in general does not help the estimation; results on convergence were not so poor for small values of lambda and for ME, conditions in which no streaks are observed. The fact that both players continually end up in the same strategy combination causes the maximum likelihood estimation to be based on not enough (new) information to perform sensible re-estimations. However, we must note that even if no streaks occur, convergence and estimation results were poor.

The results of the present chapter immediately trigger the question why estimation was in general so poor, whereas previous chapters showed that reinforcement and belief learning can be discerned easily in the same games. Playing around with starting values and convergence criteria did not alleviate the problem, revealing that the estimation problems is indeed intrinsic to the games under study. For instance, one approach we tried was to simplify the estimation problem by assigning the real values of λ and φ, and just estimating the single remaining parameter δ using the data of the same games. Even then the bias in the estimates of δ was large, again suggesting that estimation is difficult for these games, even when estimating one parameter.

To further illustrate the estimation problem, consider the following example, which can explain the paradox of on the one hand being perfectly able to easily distinguish reinforcement learning from belief learning, and on the other hand not being able to estimate the learning parameters accurately. Consider the Prisoner's Dilemma game (see Table 4.1), with Δ equal to -4. Results on stability clearly distinguish belief and reinforcement learning here. Only mutual cooperation can be stable under reinforcement learning since it is the only strategy combination giving a positive payoff (equal to 1) to both players. And only mutual defection can be stable under belief learning, since it is the only pure Nash equilibrium. Now consider the EWA model with an unknown value of δ between zero and one. Mutual cooperation can be a stable strategy for $\delta < 1/3$ and mutual defection can be a stable strategy for $\delta > 1/3$. For $\delta < 1/3$, the attraction of cooperation increases more (1) than the attraction of defection ($\delta * 3 < 1$), after mutual cooperation. And, for $\delta > 1/3$, the attraction of defection increases ($-1 + 3 * \delta > 0$) whereas the attraction of cooperation decreases ($-3 + \delta < 0$). Consequently, if streaks of mutual defection are observed in this game, these can arise not only from belief learning, but also from a wide range of learning models with $\delta > 1/3$. A similar argument can be constructed for observing a large number of strategy changes; a range of parameter values can predict this phenomenon. Hence, when estimating the best fitting EWA parameters for a game, one therefore may not easily find a unique and best fitting value, but instead a range of values that fit more or less equally well. As the example illustrates, this can happen even when the two extreme forms of learning (belief and reinforcement learning) predict completely opposite behavior in that same game.

What further complicates the problem is a very practical question: in these last three chapters, we use the different learning models precisely to emphasize differences, while in literature it is shown that subjects are more heterogeneous and do not follow one type of model precisely (e.g. Cox et al. (2001) and Wilcox (2006). A first answer to this is that in most situations where players follow one model, it already proves to be difficult to estimate the parameters. If we allow also that players can mix learning models, it becomes even harder.

Note also that the main goal of our research is not to say: people who are behaving like reinforcement learners always play the following game in such a way, resulting in these strategy combinations played. What we try to do is develop tools for, e.g., experimental economists that allow them to differentiate EWA-based reinforcement from belief learning given experimental data. A type of tool for instance is choosing a PD game with positive payoffs. Then, when an experimental economist find that mutual cooperation is played repeatedly, or as we call it: when streaks occur, then we conclude that the type of learner is a reinforcement learner (or at any rate, not a belief learner). Does this mean that this player is a 100% reinforcement learner and could not be a 10% fictitious player or any other type of learner? Unfortunately not, but it does show that the player prefers the obtained payoff in that strategy combination (like a reinforcement learner would do), without being tempted to switch to the strategy "defect" which has a higher payoff (as a belief learner would do). As a first step to address this concern, we choose in this chapter to also add the possibility of $\delta = 0.5$, allowing so-called mixed learners and see whether we can distinguish their behavior from strict reinforcement and strict belief learners.

It is very hard to take into account all linear combinations of the two types of learning (using for instance the EWA-parameter δ). This also complicates the study because you enter a grey area where both types of behavior are possible. This leads to the situation where you can explain any behavior by different combinations of different learning models. This study is about emphasizing differences between general types of learning models, not discussing all possible combinations.

In conclusion, estimation of parameters in the 2x2 games we studied is poor, even after 150 rounds of play. However, the data provide enough (mostly qualitative) information on stability and strategy changes to distinguish the EWA-based reinforcement learning and belief learning in many games, even after only 10 rounds of play. We therefore conclude that if one would want to find out whether play in a set of experimental data originates from either (pure) belief or reinforcement learners, one should use the characteristics used in the previous chapter. Finding exact parameter values will always be difficult in the 2x2 games studied.

Is accurate estimation of EWA parameters doomed, or can we create situations in which estimation is considerably improved? We can think of several possible suggestions. A first suggestion is to use games in which players are less inclined to play streaks and more often change strategies. For instance, in games without a Nash equilibrium or games with mixed and negative payoffs. However, the results of the present chapter show that although estimation problems (no convergence and biased estimates) are less severe, they still exist. A second suggestion is to use fewer parameters in our re-estimation. As stated before, the results from the functional EWA model (which uses only one parameter) have proven to be reasonably good at forecasting behavior. Given that there is too little information available to re-estimate three parameters, re-estimating one parameter might prove to be feasible. However, our example of estimating only δ shows that this need not necessarily be the case. Moreover, this approach only makes sense when one is satisfied with trying to forecast the type of learning based on a single parameter only. However, in doing so a lot of the information present in the original EWA model is lost, and hence misses out on a lot of the intuitive understanding of the parameters of the model. The second suggestion may therefore not be generally viable.

A third suggestion is to switch to games with more players and more strategies, which is what Salmon (2001) has also suggested. This could lead to more detailed information and therefore less flat likelihoods, particularly so if players do not end up playing the same strategies over and over (otherwise, the extra strategies and players still do not give much additional information). After a thorough analysis, Erev & Barron (2005) suggest to use games with limited feedback on play and with outcomes with low probability. A fifth suggestion is to estimate the parameters using data of players playing multiple games (e.g., Erev & Haruvy, 2005). This is also the approach in many papers (Roth & Erev, 1995, Erev & Roth, 1998, Bereby-Meyer & Erev, 1998). This appears a good suggestion, since we know that one of the distinguishing features between reinforcement and belief learning is that a shift in payoffs does not affect belief learning, but it does reinforcement learning. Although this fifth approach seems to be most viable, there is evidence that people behave differently in different games (Salmon, 2004, Ho et al, 2002, Stahl, 2003). This, of course, complicates matters further. A final suggestion is suggested by Ichimura

& Bracht (2001): they examine whether the EWA parameters are technically even identified and finds that they generally are not. This could clarify the problems we find in our re-estimation. However, we must acknowledge that their analyses has some limitations. First of all, they only consider the game "Matching Pennies" and they use zero payoffs in their games, something we discourage in the last chapters. The re-parameterizations they use are, however, worth investigation in future research.

To conclude, a logical next step would be to investigate the above recommendations to find out which is the most promising. Start out with a one-parameter model (for instance, but not restricted to, the fEWA model) and try to re-estimate this parameter in a diverse set of games. For instance, one could use data from a game with only a mixed strategy equilibrium with many players and many strategies per player. In case this leads to good predictions, one can try to use the same data to re-estimate more parameters from more complex learning models. Step-by-step, these analyses should lead to quantitative good re-estimations for learning models in different games.

5 The effects of social preferences on 2×2 games with only a mixed strategy equilibrium

We study the effect of a player's social preferences on his own payoff in 2×2 games with only a mixed strategy equilibrium, under the assumption that the other player has no social preferences. We model social preferences with the Fehr-Schmidt inequity aversion model, which contains parameters for "envy" and "spite". Eighteen different mixed equilibrium games are identified that can be classified into Regret games, Risk games, and RiskRegret games, with six games in each class. The effects of envy and spite in these games are studied in five different status scenarios in which the player with social preferences receives much higher ("Much Higher Status"), mostly higher ("Higher Status"), about equal ("Equal Status"), mostly lower ("Lower Status"), or much lower payoffs ("Much Lower Status"). The theoretical and simulation results reveal that the effects of social preferences are variable across scenarios and games, even within scenario-game combinations. However, we can conclude that the effects of envy and spite are analogous, on average beneficial to the player with the social preferences, and most positive in the "Equal Status" and in Risk games.

5.1 Introduction

Someone is said to have social preferences if his well-being is not only affected by his own outcomes but is a function of other persons' outcomes as well. Social preferences is a general term that encompasses fairness, altruism (e.g., Becker, 1974), inequality aversion (e.g., Fehr & Gintis, 2007, Bolton & Ockenfels, 2000), reciprocity (e.g., Rabin, 1993), concern for efficiency (e.g., Charness & Rabin, 2002), equity (e.g., Homans, 1961), and inequity aversion (Konow, 2000). Empirical evidence that humans have social preferences is abundant (e.g., Camerer, 2003, Rabin, 2006).

Given that apparently social preferences are abundant the question arises why we have them. Possible answers are that in the evolution of human kind, social preferences gave an advantage to humans having social preferences, and that social preferences benefit those who have them in social interactions. In the literature, this first answer is addressed by the "indirect evolutionary approach" according to which preferences induce behavior, behavior determines fitness, and fitness regulates the evolution of the preferences. This approach has been applied by, for instance, Güth & Yaari (1992) to explain the evolution of reciprocity, and Koçkesen, Ok & Sethi (2000) to explain the evolution of social preferences in a class of aggregative games including common recourse pool and public good games (see also Dekel, 2007, and Samuelson, 2001, for a discussion and application of the approach to the evolution of social preferences). These papers generally agree that social preferences can give an advantage. The second answer addresses if introducing social preferences for one agent, while others remain self-interested payoff maximizers, will benefit the agent with the social preferences. The latter approach is for instance applied by Engelmann & Steiner (2007) to examine the effect of risk preferences on play and outcomes in 2 × 2 games with only a mixed strategy equilibrium. In the present paper we apply the second approach to examine if social preferences are beneficial in 2 × 2 games with a mixed strategy equilibrium.

Poulsen & Poulsen (2002) show that social preferences benefit players in games with Pareto-inefficient equilibria such as the Prisoner's Dilemma game. More precisely: when players have information about opponentŠs preferences, reciprocity will always evolve. Hence, cooperation will survive. This leads to their conclusion that there is an evolutionary foundation for the experimentally observed fact that many individuals have social preferences that differ from the materialistic preferences that are normally assumed. While this is rather straightforward to show in games like the Prisoner's Dilemma game, it is not so obvious in games with only a mixed strategy equilibrium. Table 5.1 presents the most simple mixed-strategy equilibrium game, the 2 × 2 game.

Table 5.1: A 2 × 2 game .

	L	R
T	π_1^{TL}, π_2^{TL}	π_1^{TR}, π_2^{TR}
B	π_1^{BL}, π_2^{BL}	π_1^{BR}, π_2^{BR}

The mixed-strategy equilibrium is (p_1, p_2) where p_1 is the probability that player 1, the row player, plays strategy T with

$$p_1 = \frac{\pi_2^{BR} - \pi_2^{BL}}{\pi_2^{BR} - \pi_2^{BL} + \pi_2^{TL} - \pi_2^{TR}}, \tag{5.1}$$

where π_2^{XY} denotes player 2's payoff after play of X by player 1 and Y by player 2. Note how p_1 is unaffected by player 1's payoffs, and is solely determined by player 2's payoffs. If player 1 plays T with another probability $p \neq p_1$, then player 2 changes p_2 resulting in a lower payoff for 1 and a higher payoff for 2 than they obtain in the equilibrium. However, if player 1 has social preferences, then p_2 can be affected, and subsequently the equilibrium and player 1's outcomes can be affected as well. In that case, p_1 is unaffected and p_2, the probability to play strategy L equals

$$p_2 = \frac{U_1^{BR} - U_1^{TR}}{U_1^{BR} - U_1^{TR} + U_1^{TL} - U_1^{BL}}, \tag{5.2}$$

where U_i^{XY} is player 1's utility for the outcomes π_1^{XY} and π_2^{XY}.

In our analysis of the effect of player 1's social preferences on his expected payoff we make a number of assumptions. First, we assume player 2 is a self-interested payoff maximizer. Second, we assume that preferences are observable (which reduces the question of whether social preferences might be beneficial to the question whether social preferences might be beneficial given that they can be signaled credibly). That is, player 2 "computes" the "actual outcomes" of player 1 based on player 1's actual social preferences as applied to the monetary payoffs. Because of player 1's social preferences the transformed game with player 1's actual outcomes may no longer be a mixed-strategy equilibrium game. These games are included in the analysis, however, since the game in monetary payoffs is a mixed-strategy equilibrium game. Our third assumption is that both players play a one-shot game and choose their strategy simultaneously. Fourth and finally, we consider one type of social preferences based on the Fehr-Schmidt inequity aversion model (2007).

In the subsequent section, this chapter starts by listing and discussing the basic ingredients of our analysis. The Fehr-Schmidt inequity aversion model is explained and reasons are provided why we assumed this model in our analysis. The Fehr-Schmidt model models social preferences with envy and spite, hence we study if it is beneficial for player 1 if he is envious or spiteful. Then we derive all eighteen different mixed-strategy equilibrium 2 × 2 games. These can be divided into six "Regret games", six "Risk games", and six "RiskRegret games" that share properties of the first two types. Five different scenarios are introduced to examine whether the effect of envy and spite is dependent on how high player 1's monetary outcomes are relative to those of player 2. Identifying the relative outcomes by status, the scenarios studied are "Much Higher Status" (MHS, 1's outcomes are all larger than those of 2), "Higher Status" (HS, most of 1's outcomes are larger than those of 2), "Equal Status" (ES, same outcome distribution for both players), and "Lower Status" (LS, most of 1's outcomes are smaller than those of 2) and "Much Lower Status" (MLS, 1's outcomes are all smaller than those of 2).

The effects of envy and spite on the expected value of player 1's outcome are derived for all eighteen games, by introducing either a little envy or spite. To derive a numerical estimate of these effects as well, we ran simulations for all 90 game by scenario combinations, assuming uniform outcome distributions for each player. The theoretical and simulation results reveal social preferences affect the expected payoff of the gameṪhe effects of social preferences are shown to be variable across scenarios and games, and even within scenario-game combinations, depending on the players' payoffs. However, we can conclude from our theoretical and simulation results that the effects of envy and spite are analogous, on average beneficial to the player with the social preferences, and most positive in the "Equal Status" and in Risk games. These and other results are discussed in the final section of the paper.

5.2 Basic ingredients

5.2.1 Social Preferences: the Fehr-Schmidt inequity aversion model

To examine the effect of a player's social preferences on the mixed strategy equilibrium and this player's payoffs, a choice must be made on how to model these social preferences. Three models for social preferences have received more than average attention in the literature; the Fehr-Schmidt inequity aversion model (Fehr & Gintis, 2007), the equity, reciprocity, and competition model (ERC) of Bolton and Ockenfels (2000), and the fairness model of Rabin (1993).

We choose to model social preferences using the Fehr-Schmidt inequity aversion model for four reasons. First, the Fehr-Schmidt inequity aversion model has in general been able to explain results from experimental data fairly well. Diekmann & Voss (2003), for instance, show that their result —that rational actors are able to enforce social norms with sanctions— can be obtained by using the Fehr-Schmidt inequity aversion model. In addition, the Fehr-Schmidt model has appeared to be a better predictor than the other models in the taxation game, which is considered a neutral playground for the comparison of this model with others. This result is driven by the fact that the Fehr-Schmidt model is in line with maximin preferences, as opposed to the other models (Engelmann & Strobel, 2004). Secondly, of all social preference models, the Fehr-Schmidt inequity aversion model is most often applied. Third, both the Rabin's model and the Bolton & Ockenfels' ERC model include a quadratic term, which makes a theoretical analysis more difficult. Finally, our analysis serves the purpose of illustrating the effects of social preferences, and not of a detailed analysis of all possible social preference preferences like altruism, fairness, reciprocity, inequity aversion, equity, or a combination of them. Because of our choice for the Fehr-Schmidt inequity aversion model over the Rabin's model, we assume that players' utilities depend on players' actions only, and not on their beliefs.

In the Fehr-Schmidt inequity aversion model, player i's utility $U_i(XY)$ after players i and k play strategies X and Y, respectively, in a two-player game is

$$\begin{aligned} U_i(XY) = \pi_i^{XY} &- \alpha_i \max[\pi_k^{XY} - \pi_i^{XY}, 0] \\ &- \beta_i \max[\pi_i^{XY} - \pi_k^{XY}, 0], \qquad k \neq i. \end{aligned} \tag{5.3}$$

The parameters α_i and β_i characterize player i's inequity aversion. Parameter α_i represents the extent to which player i feels envy if player k earns a larger material payoff than player i, whereas β_i represents the extent to which player i feels spite if he earns a larger material payoff than player k.

5.2.2 Characterization of 2×2 games with only a mixed strategy equilibrium

Equation (5.2) shows that if player 1 has social preferences, p_2 is affected by player 1's social preferences, as expressed in the utilities. The utilities in p_2 are affected by the simultaneous ordering of all payoffs of both players. Denote the four material payoffs of the row player by $x_0 < x_l < x_h < x_1$.[1] First, we consider all possible configurations of these four payoffs. For the row player 1 there are 24 (= 4!) possible configurations of the payoffs in the 2 × 2 game, of which 12 do not result in a dominant strategy. Without any loss of generality we place x_1 in the top-left cell, corresponding to π_1^{TL}, which reduces the number of different payoff configurations for player 1 to three. These three configurations are shown in Table 5.2, Table 5.3, and Table 5.4.

Table 5.2. The Regret game.

	L	R
T	$x_1, \ldots$	$x_l, \ldots$
B	$x_0, \ldots$	$x_h, \ldots$

Table 5.3. The RiskRegret game.

	L	R
T	$x_1, \ldots$	$x_0, \ldots$
B	$x_l, \ldots$	$x_h, \ldots$

Table 5.4. The Risk game.

	L	R
T	$x_1, \ldots$	$x_0, \ldots$
B	$x_h, \ldots$	$x_l, \ldots$

The game in Table 5.2 is called a "Regret game" for the row player since the difference between the obtained and the foregone payoff is highest for one of player 2's strategies, and small for the other. The game in Table 5.4 is labeled "Risk game" because player 1's strategy T has a higher risk or payoff variance than strategy B. The game labeled "RiskRegret" in Table 5.3 is similar to the Risk game but has larger differences between the obtained and the foregone payoff for the row player, similar to the "Regret game".

Turning to the column player, also 12 out of 24 configurations of player 2's payoffs do not result in a dominant strategy in the 2 × 2 game. Combining these 12 configurations with each of the games in Table 5.2 until Table 5.4 results in 18 games with one mixed strategy equilibrium and 18 games with at least one pure strategy Nash equilibrium. Hence, we end up with 18 different 2 × 2 games with only a mixed strategy equilibrium (or differently put: all 2 × 2 games with only a mixed strategy equilibrium are one of these 18 games). All 18 games and their names can be found in Figure 5.1. The games are labeled "A-B-number", with "A" and "B" being the type of game (Regret, RiskRegret, Risk) for the row and the column player, respectively, with 'number' running from 1 to 18. Numbers $1-6$, $7-12$, $13-18$ correspond to a Regret, RiskRegret, Risk game for the row player, respectively. Below we will refer to these games by their numbers.

[1] Actually, we have $x_0 \leq x_l < x_h \leq x_1$ or $x_0 < x_l \leq x_h < x_1$. Note that in the case of all payoffs being equal, we have a (weak) pure strategy Nash equilibrium. Without loss of generalization, we assume that all the payoffs differ.

Regret-Regret-1

	L	R
T	x_1, x_0^*	x_l, x_1^*
B	x_0, x_h^*	x_h, x_l^*

Regret-Regret-2

	L	R
T	x_1, x_l^*	x_l, x_h^*
B	x_0, x_1^*	x_h, x_0^*

Regret-RiskRegret-3

	L	R
T	x_1, x_l^*	x_l, x_1^*
B	x_0, x_h^*	x_h, x_0^*

Regret-RiskRegret-4

	L	R
T	x_1, x_0^*	x_l, x_h^*
B	x_0, x_1^*	x_h, x_l^*

Regret-Risk-5

	L	R
T	x_1, x_h^*	x_l, x_1^*
B	x_0, x_l^*	x_h, x_0^*

Regret-Risk-6

	L	R
T	x_1, x_0^*	x_l, x_l^*
B	x_0, x_1^*	x_h, x_h^*

RiskRegret-Regret-7

	L	R
T	x_1, x_0^*	x_0, x_1^*
B	x_l, x_h^*	x_h, x_l^*

RiskRegret-Regret-8

	L	R
T	x_1, x_l^*	x_0, x_h^*
B	x_l, x_1^*	x_h, x_0^*

RiskRegret-RiskRegret-9

	L	R
T	x_1, x_l^*	x_0, x_1^*
B	x_l, x_h^*	x_h, x_0^*

RiskRegret-RiskRegret-10

	L	R
T	x_1, x_0^*	x_0, x_h^*
B	x_l, x_1^*	x_h, x_l^*

RiskRegret-Risk-11

	L	R
T	x_1, x_h^*	x_0, x_1^*
B	x_l, x_l^*	x_h, x_0^*

RiskRegret-Risk-12

	L	R
T	x_1, x_0^*	x_0, x_l^*
B	x_l, x_1^*	x_h, x_h^*

Risk-Regret-13

	L	R
T	x_1, x_0^*	x_0, x_1^*
B	x_h, x_h^*	x_l, x_l^*

Risk-Regret-14

	L	R
T	x_1, x_l^*	x_0, x_h^*
B	x_h, x_1^*	x_l, x_0^*

Risk-RiskRegret-15

	L	R
T	x_1, x_l^*	x_0, x_1^*
B	x_h, x_h^*	x_l, x_0^*

Risk-RiskRegret-16

	L	R
T	x_1, x_0^*	x_0, x_h^*
B	x_h, x_1^*	x_l, x_l^*

Risk-Risk-17

	L	R
T	x_1, x_h^*	x_0, x_1^*
B	x_h, x_l^*	x_l, x_0^*

Risk-Risk-18

	L	R
T	x_1, x_0^*	x_0, x_l^*
B	x_h, x_1^*	x_l, x_h^*

Figure 5.1. The 18 different mixed-strategy Nash equilibrium only 2 × 2 games

The utilities in equation (5.2) for p_2 depend on the simultaneous ordering of the payoffs of both players. In each cell of one game, the payoff for the row player can be higher or lower than the payoff for the column player, which results in a maximum of $2^4 = 16$ different possibilities. The number of possibilities differs per game. A more detailed analysis reveals that there are in total 122 different 2 × 2 games with only a mixed strategy equilibrium depending on the order of all 8 payoffs in all games.[2]

Instead of analyzing the effect of social preferences on the outcome of the row player in each of the 122 different games, we continue to examine the eighteen games for two reasons. First, restricting ourselves to eighteen games keeps the analysis orderly. Second, it is not clear how the results on the 122 games should be summarized. To summarize one must specify the relative weight of the results for each of the 122 cases. A simple specification is to weigh all 122 cases equally. However, one would like the weights to correspond to the likelihood that each case occurs. These likelihoods are unknown and depend on the distributions of the payoffs in the game. We therefore analyze the effect of social preferences on the row player's outcome for each of the eighteen games using five different scenarios, each assuming different outcome distributions of both players. The scenarios correspond to sensible interpretations in terms of the status difference of the two players.

5.2.3 Five different scenarios

The scenarios we distinguish on the players' payoff distributions are (1) "Much Higher Status" (MHS), (2) "Higher Status" (HS), (3) "Equal Status" (ES), (4) "Lower Status" (LS), (5) "Much Lower Status" (MLS).

(i) *Much Higher Status*: The distribution of the row player's payoffs solely contains higher payoffs than the column player's distribution. Later on, this is simulated by drawing the row player's payoffs from $U(\frac{2}{3}, 1)$, whereas the column player's payoffs are drawn from $U(0, \frac{1}{3})$, where $U(a, b)$ is the uniform density function with minimum a and maximum b. This scenario corresponds to asymmetric relations, such as boss/employee and parent/child relationships.

(ii) *Higher Status*: Half of the distribution of the row player's payoffs solely contains higher payoffs than the column player's distribution. This scenario is simulated by drawing the row player's payoffs from $U(\frac{1}{3}, 1)$, whereas the column player's payoffs are drawn from $U(0, \frac{2}{3})$. Consequently, the probability is .875 that a randomly drawn the row player's payoff is larger than a randomly drawn the column player's payoff. This scenario where the row player has a predominantly higher status can be found in older brother/younger brother or in a co-operating boss/employee relationship.

(iii) *Equal Status*: Both players' outcome distributions are the same. This is simulated by drawing the row player's and the column player's payoffs from $U(0, 1)$. This scenario can be found among friends and colleagues. Hence there is a probability of .5 that a randomly drawn outcome of the row player is higher than a randomly drawn outcome of the column player.

[2] More precisely, games 4 and 7 have 5 possibilities, games 1, 2, 6, 9, 10 and 13 have 6 possibilities, games 3, 8, 11, 12, 15, and 16 have 7, games 5 and 14 have 8 possibilities, and games 17 and 18 have 9 possibilities.

(iv) *Lower Status* and

(v) *Much Lower Status*: In these scenarios, the roles of the column and the row player are reversed compared to their roles in 1 and 2.

5.2.4 Method

The effect of social preferences is examined by a combination of theoretical derivation and simulation analysis. We derive the effect of the row player's envy on his expected payoffs in each scenario, using the uniform distributions as described earlier. We also show that the results for spite are analogous to the results for envy. We add simulations to assess the size and variability of the effect of social preferences.

In the simulations the effect of social preferences is examined by analyzing the partial derivative of the row player's expected payoff of the game in α and β at $\alpha = 0$ and $\beta = 0$, that is, at the point where there are no social preferences. If the partial derivative is positive, social preferences increase the player's expected payoff of the game. The partial derivatives are numerically approximated by computing the expected payoff of the game for (α,β) equal to (0,0) and (.001,0), and for (0,0) and (0,.001). The expected payoff of the game G of the row player is calculated as

$$\mathbb{E}[G] = p_1 p_2 \pi_1^{TL} + p_1(1-p_2)\pi_1^{TR} + (1-p_1)p_2\pi_1^{BL} + (1-p_1)(1-p_2)\pi_1^{BR}, \quad (5.4)$$

with p_1 and p_2 as defined in (5.1) and (5.2), respectively. The utilities in p_2 are calculated by applying the Fehr-Schmidt model to the randomly generated outcomes in each cell of the game.

For each game (18) × scenario (5) combination $10,000$ simulations were run, which amounts to $900,000$ simulations in total. We report the average increase in expected payoff of the game and the proportion of the simulations where the expected payoff of the game increased after introducing social preferences. The results on the average increase can be different from the results on the proportion if the distribution of the increase is asymmetric at 0.

5.3 Results

5.3.1 Theoretical results

Denote the difference between the expected payoff of the game after introducing social preferences (game G′) and the expected payoff of the game without social preferences (game G) as

$$\mathbb{E}(G') - \mathbb{E}(G) = \Delta p_2(\alpha, \beta) C(\text{game}), \quad (5.5)$$

where $\Delta p_2(\alpha, \beta)$, further abbreviated by Δp_2, is the change in the mixed-equilibrium due to the social preferences, and where

$$C(\text{game}) = p_1(\pi_1^{TL} - \pi_1^{TR}) + (1-p_1)(\pi_1^{BL} - \pi_1^{BR}). \quad (5.6)$$

Equation (5.5) signifies that the expected payoff of the game increases when either $\Delta p_2 > 0$ and $C(\text{game}) > 0$ or when both are negative. For each of the eighteen games we analyze how both are affected by the game and social preferences, starting with the effect on $C(\text{game})$. $C(\text{game})$ is unaffected by social preferences since it is not dependent on p_2. Hence $C(\text{game})$ is only affected by game.

Let δ denote the expected difference between the highest and the single highest payoff. Then, assuming a uniform distribution of payoffs, the player's expected payoffs $x'_1, \ldots, x'_4$ can be written as:[3]

$$x'_1 = x'_h + \delta = x'_l + 2\delta = x'_0 + 3\delta, \tag{5.7}$$

Substitution of these expected payoffs into equation (5.6) provides insight into whether $C(\text{game})$ is mostly positive or negative.

For Regret games, the two expected payoff differences in equation (5.6) are 2δ and -2δ, respectively. Hence, whether $C(\text{game})$ is mostly positive or negative depends on the order of magnitude of p_1. This depends on the type of Regret game. Substituting equation (5.7) into equation (5.1) yields $p_1 \approx 1/4$ for game 1, $p_1 \approx 3/4$ for game 2, and $p_1 \approx 1/2$ for games 3 to 6. Hence we expect $C(\text{game})$ to be mostly negative in game 1 and mostly positive in game 2. These results are summarized in the third column of Table 5.5; '+' and '-' indicate that $C(\text{game})$ is mostly positive and mostly negative, respectively. Omission of '+' and '-' signifies that $C(\text{game})$ is about equally often positive and negative.

In the RiskRegret games, the first expected payoff difference in equation (5.6) is 3δ and the second is δ. Hence $C(\text{game})$ is positive if $p_1 > \frac{1}{4}$. This is true for all RiskRegret games, except game 7, where $p_1 \approx 1/4$. In game 8 we particularly expect $C(\text{game})$ to be positive since $p_1 \approx 3/4$. Finally, for all Risk games $C(\text{game}) > 0$, because $\pi_1^{TL} > \pi_1^{TR}$ and $\pi_1^{BL} > \pi_1^{BR}$. All these results are also summarized in the third column of Table 5.5.

[3] The expected difference is in fact equal to $\frac{b-a}{5}$, with a and b equal to the lower and upper value of the uniform density function.

Table 5.5. Expectations of the sign of $C(\text{g[ame]})$, Δp_2 $(p_2(\alpha))$,and change in expected payoff of the game (ΔE) with respect to envy (α), for each game.

			$\Delta p_2(\alpha)$				ΔE			
Gametype	Game	$C(g)$	*MLS*	*LS*	*ES*	*HS*	*MLS*	*LS*	*ES*	*HS*
Regret	1	−	+	+	+	+	−	−	−	−
	2	+	+	+	+	−	+	+	+	−
	3		+	+	+	+				
	4			+		−				
	5		+	+	+	+				
	6		−	−	−	−				
RiskRegret	7			+	+	+				
	8	+								
	9	+	+	+	+	+	+	+	+	+
	10	+	−	−	−		−	−	−	
	11	+	+	+	+	+	+	+	+	+
	12	+	−	−	−	−	−	−	−	−
Risk	13	+		+	+	+		+	+	+
	14	+		+	+	+		+	+	+
	15	+	+	+	+	+	+	+	+	+
	16	+	−	−	+	+	−	−	+	+
	17	+	+	+	+	+	+	+	+	+
	18	+	−	−	−	−	−	−	−	−

Consider now p_2, the only part of equation (5.6) that is determined by the social preferences. A payoff's utility decreases in spite and envy. Let the decrease in a payoff's utility be equal to Δ. Then Δp_2 is positive when

$$p_2(G') = \frac{\pi^{BR} - \Delta^{BR} - \pi^{TR} + \Delta^{TR}}{\pi^{BR} - \Delta^{BR} - \pi^{TR} + \Delta^{TR} + \pi^{TL} - \Delta^{TL} - \pi^{BL} + \Delta^{BL}} \tag{5.8}$$

$$> \frac{\pi_1^{BR} - \pi_1^{TR}}{\pi_1^{BR} - \pi_1^{TR} + \pi_1^{TL} - \pi_1^{BL}} = p_2(G). \tag{5.9}$$

This means that if

$$(\Delta_1^{TR} - \Delta_1^{BR})(\pi_1^{TL} - \pi_1^{BL}) > (\Delta_1^{BL} - \Delta_1^{TL})(\pi_1^{BR} - \pi_1^{TR}), \tag{5.10}$$

then $\Delta p_2 > 0$. For the different types of games, this implies that $\Delta p_2 > 0$ when:

$$3{\cdot}(\Delta_1^{TR} - \Delta_1^{BR}) > (\Delta_1^{BL} - \Delta_1^{TL}), \qquad \text{for Regret games} \tag{5.11}$$

$$(\Delta_1^{TR} - \Delta_1^{BR}) > (\Delta_1^{BL} - \Delta_1^{TL}), \qquad \text{for RiskRegret and Risk games.} \tag{5.12}$$

Equations (5.11) and (5.12) are analyzed by verifying which Δ is larger than others for each game × scenario combination, which is determined by the difference between the row and the column player's payoffs in the same cell of the 2×2 game. The results of the analysis are summarized in columns four to seven of Table 5.5, for envy. Rather than explicating all results in detail, we illustrate the analysis by working out the results for all scenarios in game 16.

In scenario “Much Higher Status” the row player’s payoff is strictly larger than the column player’s payoff. Hence no effect of envy is present in this scenario, and is therefore omitted from Table 5.5. Consider now scenario "Higher Status". To evaluate if $(\Delta_1^{TR} - \Delta_1^{BR}) > (\Delta_1^{BL} - \Delta_1^{TL})$ we first calculate Δ_1^{TR}. For convenience and without loss of generality, we assume that $\alpha = 1$. The expected value of x_0 then equals $1/5 \times 2/3 + 1/3 = 7/15$, whereas the expected value of x_h^* equals $3/5 \times 2/3 = 6/15$.[4] These values indicate that on average $x_0 > x_h^*$, but with some probability $x_h^* > x_0$, resulting in envy. Simulation shows that the expected value of Δ_1^{TR} equals .039.[5] The expected value of Δ_1^{BR} equals .002. Hence the expected value of $\Delta_1^{TR} - \Delta_1^{BR}$ equals .037. The expected value of $\Delta_1^{BL} - \Delta_1^{TL}$ can be shown to be equal to $.010 - .000 = .010$. To conclude the tedious calculations, $\Delta_1^{TR} - \Delta_1^{BR} = .037 > \Delta_1^{BL} - \Delta_1^{TL} = .010$, hence envy in scenario "Higher Status" in game 16 results in $\Delta p_2 > 0$. This conclusion is presented as a ’+’ in the fourth column of Table 5.5.

Similarly, the expected value of Δp_2 is also larger than 0 in scenario "Equal Status" in game 16; $\Delta_1^{TR} - \Delta_1^{BR} = .197 > \Delta_1^{BL} - \Delta_1^{TL} = .155$. But the expected value of the difference is negative in scenarios "Lower Status" (.264 - .563) and "Much Lower Status (.400 - .800). Hence the expected value of $\Delta p_2 < 0$ for these two scenarios, which is presented as ’-’ in columns six and seven of Table 5.5.

The results on Δp_2 for all game × scenario combinations are shown in columns four to seven of Table 5.5. A cell is empty if it’s expected value equals 0. Inspection of these columns reveals that the sign of the effect of envy on Δp_2 can be different for the scenarios, even in the same game. As a result, the effect of envy on the row player’s payoff can also be different across scenarios. Columns eight to eleven present the effects of envy on the expected payoff of the game (ΔE) for each game × scenario combination. The result in a cell is obtained by multiplying the sign of the corresponding Δp_2 (four columns to the left) with the sign of C(game) (column three).

The latter four columns of Table 5.5 suggest that there is no generally negative of positive effect of envy on the row player’s payoffs in regret games across scenarios, and across games in the "Much Lower Status" scenario. However, the effect of envy seems positive across games in the "Lower", "Equal", and "Higher" scenarios, and across scenarios in the Risk games. Averaged across all game × scenario combinations the effect appears positive, since the table contains more ’+’ than ’-’ signs.

The results of spite on the row player’s payoffs are analogous to the effect of envy. That is, the effect of envy in each game × scenario combination corresponds to an equal effect of spite in another game × scenario combination. The analogy is that the effect of envy in a "Much Higher Status", "Higher Status", "Equal Status", "Lower Status", "Much Lower Status" scenario is equal to the effect of spite in a

[4] The expected value of the lowest of four random draws of a uniform distribution on interval (a,b) equals $1/5 \times (b - a) + a$.

[5] It is not straightforward to calculate the expected value of the absolute difference of two order statistics. Hence we simulated the distributions of all combinations of order statistics for all game × scenario combinations. To compute the expected value of the difference we used 1,000,000 runs. This results in highly accurate estimated expected values, that is, with very low standard error.

"Much Lower Status", "Lower Status", "Equal Status", Higher Status", "Much Higher Status" scenario, respectively. It turns out that the effects of envy and spite are equal in games 3-8,13,14. The effects of envy in games 1, 2, 9, 10, 11, 12, 15, 16, 17, 18 are equal to the effects of spite in games 2, 1, 10, 9, 12, 11, 16, 15, 18, 17, respectively. For example, consider the effect of envy in game 1 and the effect of spite in game 2. The payoff differences in the cells of these games are exact mirror images of each other. Consequently, expected experienced envy in game 1 equals expected experienced spite in game 2, in all scenarios.[6]

The analogy of the results of envy and spite has four important implications. First, their effects averaged over all game × scenario combinations are equal. That is, if envy is on average beneficial for a player, then so is spite. Second, because the effect of envy in a specific type of game (regret, risk/regret, risk) is analogous to the effect of spite in the same type of game, the main effects of envy and spite in each type of game are equal. Alternatively, if envy on average has a beneficial effect in risk games, then so has spite. Third, because the effect of envy in a scenario is equivalent to the effect of spite in the mirror-image of that scenario, the main effects of envy and spite in these corresponding scenarios are equal. That is, for example, if the effect of envy is beneficial for the player in scenario "Lower Status", then the effect of spite is equally beneficial for that player in scenario "Higher Status". Fourth, since the results of envy and spite are analogous, we below only need to present the results of envy.

Using the analogy of the results of envy and spite, we preliminarily conclude that the combined effect of envy and spite is:

(i) On average beneficial to the player;

(ii) Not or hardly beneficial in regret games;

(iii) Not or hardly beneficial in the "Much Lower Status" and "Much Higher Status" scenarios;

(iv) On average beneficial in risk-regret, and more beneficial in risk games;

(v) On average beneficial in scenarios "Lower Status", "Equal Status", "Higher Status".

These conclusions, and the size and variability of the effects of envy and spite are addressed when discussing the simulation results.

5.3.2 Simulation results

Table 5.6 presents the effect of the introduction of envy on the (slope of the) expected payoff for the row player with envy. Since no envy is experienced in Scenario "Much Higher Status", the results for that scenario are omitted from the table. The numbers

[6] More precisely, the analogy is the following. Cell top-right in game 1 presents the difference between the 3rd and 1st order statistic, cell bottom-left in game 2 between the 2nd and 4th order statistic. Note that the difference in order statistics is the same. For cells bottom-right, bottom-left, top-left in game 1 a difference is obtained between order statistics 2 and 3, 4 and 2, 1 and 4, respectively. The same differences in order statistics are obtained for game 2; 3 and 2, 1 and 3, 4 and 1, for corresponding cells top-right, top-left, bottom-left, respectively.

in the cells and marginals of the table correspond exactly with our theoretical results of Table 5.5 and formulated at the end of previous section. Hence we do not reiterate these conclusions here again. Interestingly, the results in Table 5.6 suggest that the positive effect of envy is on average substantially larger (six times) in "Equal Status" than in the "Lower Status" and "Higher Status" scenarios. Additionally, the positive effect of envy in RiskRegret games is much smaller (eleven times) than in Risk games.

Table 5.6. The mean slope of the expected payoff of the game in α for each of the games under each of the scenarios.

Gametype	Game	HS	ES	LS	MLS	Total	
Regret	1	−0.0021	−0.0373	−0.0400	−0.0184	−0.0196	
	2	−0.0034	0.0163	0.0331	0.0177	0.0127	
	3	0.0024	0.0097	0.0010	−0.0008	0.0025	
	4	0.0002	0.0010	0.0004	0.0001	0.0003	
	5	0.0028	0.0082	0.0003	−0.0015	0.0020	
	6	−0.0003	−0.0038	−0.0011	−0.0005	−0.0011	
	Total	0.0001	−0.0010	−0.0011	−0.0006		−0.0005
RiskRegret	7	0.0001	0.0009	0.0003	0.0001	0.0003	
	8	0.0026	0.0105	0.0013	−0.0008	0.0027	
	9	0.0126	0.0491	0.0355	0.0177	0.0230	
	10	0.0002	−0.0044	−0.0237	−0.0184	−0.0093	
	11	0.0135	0.0667	0.0668	0.0351	0.0364	
	12	−0.0033	−0.0404	−0.0597	−0.0371	−0.0281	
	Total	0.0043	0.0137	0.0034	−0.0006		0.0042
Risk	13	0.0602	0.2256	0.0591	−0.0029	0.0684	
	14	0.0302	0.1751	0.0091	−0.0127	0.0403	
	15	0.0768	0.3589	0.2123	0.0929	0.1482	
	16	0.0212	0.0625	−0.1369	−0.1086	−0.0324	
	17	0.0773	0.4116	0.3798	0.1961	0.2130	
	18	−0.0108	−0.1810	−0.3376	−0.2092	−0.1477	
	Total	0.0425	0.1755	0.0310	−0.0074		0.0483
Total		0.0156	0.0627	0.0111	−0.0028	0.0173	

To examine the variability of the effect of envy, we also examined the proportion of increases in expected payoff of the game after introducing envy for the row player. These results are presented in Table 5.7. First, the results on the proportion of payoff increases mimic the results on expected payoff of the game increase. In 52.70% of all the games, the row player's expected payoff of the game increases, suggesting that envy helps the row player. The row player's expected payoff of the game increases most often in "Equal Status" (57.77%), and increases but less in “Higher Status” (52.30%) and “Lower Status” (51.25%). Among the three game types, the row player's expected payoff of the game increases most often in Risk games, increases slightly more often than decreases in RiskRegret games (51.08%), and even decreases most often in regret games (46.00%).

Table 5.7. The proportion of an increase of player 1's expected payoff of the game as a function of scenario and game type, after introducing envy.

Gametype	Game	HS	ES	LS	MLS	Total	
Regret	1	41.38	17.88	16.98	21.02	24.31	
	2	25.26	49.82	74.78	77.78	56.91	
	3	49.44	51.85	49.47	49.37	50.03	
	4	48.35	49.88	49.47	49.15	49.21	
	5	44.97	48.08	49.12	49.03	47.80	
	6	45.25	47.90	48.93	48.87	47.74	
	Total	42.44	44.24	48.13	49.20		46.00
RegretRisk	7	43.04	47.89	48.59	49.15	47.17	
	8	43.63	52.60	50.03	49.37	48.91	
	9	71.01	85.26	79.36	77.78	78.35	
	10	40.63	47.19	28.33	21.02	34.29	
	11	62.88	75.60	76.84	76.84	73.04	
	12	27.80	25.71	22.86	22.47	24.71	
	Total	48.17	55.71	51.00	49.44		51.08
Risk	13	81.41	92.13	62.65	50.13	71.58	
	14	58.88	72.42	53.16	49.41	58.47	
	15	81.43	94.73	81.54	77.14	83.71	
	16	58.77	66.34	29.36	22.10	44.14	
	17	81.49	98.47	100.00	100.00	94.99	
	18	35.74	16.13	0.94	0.00	13.20	
	Total	66.29	73.37	54.61	49.80		61.02
Total		52.30	57.77	51.25	49.48	52.70	

At least two interesting observations can be made from the results presented in Table 5.7. First, and most importantly, the effect of envy on the row player's payoff is in general very variable. Except for the effect of envy in two scenarios in Risk game 17, the effect of envy is neither uniformly negative nor positive in a game, and consequently, also not in a scenario or game type. Second, interestingly, in 16 cells of Table 5.6, the sign of the average effect does not correspond with the results on the proportions in Table 5.7. For example, the mean slope of game 3 in the "Higher Status" scenario is slightly positive (0.0024), whereas the proportion of increase is slightly less than 50% (49.44). In 15 out of 16 cases the slope was positive and the percentage less than 50%. This indicates that the distribution of the slope is skewed to the right. From these 15 cases, 10 occurred in the Regret game, which explains why the average slope almost equals zero, whereas the average proportion of 46% is considerably less than 50%.

5.4 Conclusions and discussion

This paper examines the effect of one player's social preferences on his outcomes in 2 × 2 games with only a mixed strategy equilibrium. Before, we stated that even though it is easy to show that social preferences benefit players in games with Pareto-inefficient equilibria such as the Prisoner's Dilemma game, the case of 2 × 2 games with only a mixed strategy equilibrium is particularly interesting and challenging; it is not at all obvious if and how a player's social preferences can benefit him. Following Fehr-Schmidt inequity aversion model, we assume two types of social preferences, envy and spite. Moreover, we assume the other player has no social preferences. The effect of envy and spite is assessed in a population of mixed strategy equilibrium games with uniformly distributed outcomes, in five different scenarios. The scenarios differ with respect to the relative position of the players' outcomes, with the row player's outcomes being strictly lower ("Much Lower Status"), mostly lower ("Lower Status"), about equal ("Equal Status"), mostly higher ("Higher Status"), and always higher ("Much Higher Status").

The general conclusions from our theoretical and simulation results are that the effects of envy and spite on outcomes in 2 × 2 games with only a mixed strategy equilibriumare variable, even for a given scenario in one particular game. However, we can conclude that the effects of envy and spite are (i) analogous, (ii) beneficial to the player with the social preferences (averaged across games and scenarios), (iii) variable across games, and consequently across game types and scenarios as well, (iv) most positive in the "Equal Status", less positive in the "Lower Status" and "Higher Status", and neither negative nor positive in the "Much Lower Status" and "Much Higher Status" scenarios, and finally (v) most positive in Risk games, less positive in RiskRegret games, and neither negative nor positive in Regret games.

Note that the first result that the effects of envy and spite are analogous is at least partly based on the assumption of uniform payoff distributions. That is, if we had assumed asymmetric payoff distributions, the effects of envy and spite would not have been symmetric. However, we suspect that choosing other common continuous distributions of outcomes (e.g., normal or exponential distribution) will not affect our other general conclusions; it will only affect the size and variability of the effects.

The meaningfulness of our second result that the player's social preferences are on average beneficial to the player who has them can be disputed. Following from the third result, since the benefits clearly depend on game type and scenario, it is definitely more meaningful to consider the effects of social preferences in certain game types and scenarios. Put differently, if a player is consistently confronted with games in one particular game and scenario, the effects of social preferences could be very different from its effects in another context.

The fourth result implies that in an "Equal Status" context the effects of envy and spite are most beneficial. That is, among players with equal status such as friends and colleagues, it is most beneficial to have social preferences in interdependent situations without a pure Nash equilibrium. The effects are also slightly beneficial if the other player has predominantly more status or predominantly less status, as in older sibling/younger sibling or co-operating boss/employee relationships. No positive effects exist in strictly asymmetric relationships, such as boss/employee and parent/child relationships. Similarly, result five implies that particularly in interdependence relations with one risky and one less risky alternative, envy and spite can be beneficial.

What are the implications of our results for the development of social preferences in the evolution of humans? Obviously, our results concern only a limited number of interdependence situations. However, they do show that not only in straightforward interdependence situations such as Prisoner's Dilemmas, but also in interdependence situations without a pure Nash equilibrium, social preferences can be beneficial. Moreover, they are beneficial when only one player has social preferences where the other does not. That is, our results suggest that if humans recognized these interdependence situations as well as their partners perceiving envy and spite, then their outcomes increased, providing an evolutionary advantage. Note that with outcomes increased, we do not just mean that they are more content with their outcome, but we also mean that their outcome is higher in a strict monetary sense. This holds in particular when interacting with equal status actors, or with actors with slightly lower or higher status. A consequence of our results is also that a population of individuals without social preferences is not evolutionary stable, if they would be consistently interacting in 2×2 games with only a mixed strategy equilibrium. Evolution would result in a population where at least some envy and spite is present.

Samenvatting

Leermodellen in interdependente situaties

Veel modellen die gebruikt worden om leergedrag in spellen te beschrijven, vallen in één van de volgende twee klassen: “reinforcement” en “belief” leermodellen. Het reinforcement leermodel gaat uit van de theorie dat succesvolle acties in het verleden een hogere kans hebben om opnieuw gespeeld te worden. Het belief leermodel neemt aan dat een speler een bepaalde verwachting heeft wat de tegenstander gaat doen en past zijn strategie vervolgens zo aan dat hij de maximale beloning zal krijgen, gegeven zijn verwachting van wat de tegenspeler doet. Het reinforcement en belief leermodel zijn speciale gevallen van het hybride leermodel “Experience Weighted Attraction” (EWA). Enkele studies laten expliciet zien dat het moeilijk is om te bepalen welk leermodel (reinforcement, belief of een ander) verantwoordelijk is voor welke data in spellen. Dit leidt tot de volgende hoofdvraag van dit proefschrift: *Kunnen we onderscheid maken tussen de verschillende typen van op EWA gebaseerd leren, met reinforcement leren en belief leren als speciale gevallen, in* 2×2 *spellen?*

In Hoofdstuk 2 leiden we voorspellingen af van gedrag in drie typen spellen gebruik makend van het EWA leermodel, waarbij we gebruik maken van het concept van stabiliteit: er is een grote kans dat alle spelers dezelfde keuze maken in ronde $t + 1$ als in ronde t. Hiermee, concluderen we dat belief en reinforcement leren *wel* onderscheiden kan worden, zelfs in 2×2 spellen. Het duidelijkste verschil in gedrag veroorzaakt door belief of reinforcement leren wordt gevonden in spellen met een puur Nash evenwicht met negatieve beloningen en tenminste één strategie combinatie met alleen maar positieve beloningen. Onze resultaten helpen onderzoekers om spellen te identificeren waar we belief en reinforcement leren kunnen onderscheiden.

Onze theoretische resultaten impliceren dat leermodellen onderscheiden kunnen worden *na voldoende rondes gespeeld te hebben*, maar het is niet duidelijk hoe groot dat getal moet zijn. Het is ook niet duidelijk hoe waarschijnlijk het is dat stabiliteit eigenlijk optreedt in spellen. Daarom beschouwen we in Hoofdstuk 3 nu ook de hoofdvraag door data te simuleren van verschillende leermodellen. We gebruiken dezelfde drie typen 2×2 spellen als in het vorige hoofdstuk en onderzoeken of we onderscheid kunnen maken tussen reinforcement en belief leren in een experimentele setting. Onze conclusie is dat dit ook mogelijk is, zeker in spellen met positieve beloningen en in het herhaalde “Prisoner’s Dilemma spel”, zelfs als het herhaalde spel relatief weinig rondes is. We laten ook zien dat andere karakteristieken van het gedrag van een speler, zoals het aantal keer dat een speler van strategie verandert en het aantal gespeelde strategie combinaties, helpen met het onderscheid maken van de twee leermodellen.

Tot dusver, beschouwden we enkel het "pure" belief en het pure "reinforcement" leren en niets daar tussenin. In Hoofdstuk 4, beschouwen we daarom een algemenere klasse van leermodellen en we proberen uit te zoeken in welke gevallen, we drie parameters uit het EWA leermodel kunnen terugschatten uit simulatie data, gegenereerd voor verschillende spellen en scenario's. De resultaten laten lage convergentie percentages van het schattingsalgoritme zien en zelfs als het algoritme convergeert, dan zien we veel onzuivere schatters van de parameters. Dus, we moeten concluderen dat het terugschatten van de exacte parameters op een kwantitatieve manier moeilijk is in de meeste experimentele opzetten. Maar, op een kwalitatieve manier kunnen we wel patronen vinden die laten zien of we in de richting van belief leermodellen of reinforcement leermodellen moeten kijken.

Ten slotte, in het laatste hoofdstuk, bestuderen we het effect van sociale preferenties van een speler op zijn eigen beloning in 2×2 gemixte evenwicht spellen, onder de aanname dat de andere speler geen sociale preferenties heeft. We modelleren sociale preferenties met het model van Fehr & Schmidt, die parameters bevat voor "afgunst" en "medelijden". Achttien verschillende gemixte evenwicht spellen worden er afgeleid die gesorteerd kunnen worden in Spijt spellen, Risico spellen en Risico-Spijt spellen, met zes spellen in elke klasse. De effecten van afgunst en medelijden in deze spellen worden bestudeerd in vijf verschillende scenario's waar de speler met sociale preferenties veel hogere, overwegend hogere, ongeveer dezelfde, overwegend lagere en veel lagere beloningen krijgt. De theoretische en simulatie resultaten laten zien dat het effect van sociale preferenties variëren over de verschillende scenario's en spellen, en interacties daartussen. Echter, we kunnen concluderen dat de effecten van afgunst en medelijden vergelijkbaar zijn, gemiddeld gesproken gunstig voor de speler met sociale preferenties en het gunstigst als de beloningen ongeveer gelijk zijn en in Risico spellen.

Abstract

Learning models in interdependence situations

Many approaches to learning in games fall into one of two broad classes: reinforcement and belief learning models. Reinforcement learning assumes that successful past actions have a higher probability to be played in the future. Belief learning assumes that players have beliefs about which action the opponent(s) will choose and that players determine their own choice of action by finding the action with the highest payoff given the beliefs about the actions of others. Belief learning and (a specific type of) reinforcement learning are special cases of a hybrid learning model called Experience Weighted Attraction (EWA). Some previous studies explicitly state that it is difficult to determine the underlying process (either reinforcement learning, belief learning, or something else) that generated the data for several games. This leads to the main question of this thesis: *Can we distinguish between different types of EWA-based learning, with reinforcement and belief learning as special cases, in repeated* 2×2 *games?*

In Chapter 2 we derive predictions for behavior in three types of games using the EWA learning model using the concept of stability: there is a large probability that all players will make the same choice in round $t + 1$ as in t. Herewith, we conclude that belief and reinforcement learning *can* be distinguished, even in 2×2 games. Maximum differentiation in behavior resulting from either belief or reinforcement learning is obtained in games with pure Nash equilibria with negative payoffs and at least one other strategy combination with only positive payoffs. Our results help researchers to identify games in which belief and reinforcement learning can be discerned easily.

Our theoretical results imply that the learning models can be distinguished *after a sufficient number of rounds have been played*, but it is not clear how large that number needs to be. It is also not clear how likely it is that stability actually occurs in game play. Thereto, we also examine the main question by simulating data from learning models in Chapter 3. We use the same three types of 2×2 games as before and investigate whether we can discern between reinforcement and belief learning in an experimental setup. Our conclusion is that this is also possible, especially in games with positive payoffs and in the repeated Prisoner's Dilemma game, even when the repeated game has a relatively small number of rounds. We also show that other characteristics of the players' behavior, such as the number of times a player changes strategy and the number of strategy combinations the player uses, can help differentiate between the two learning models.

So far, we only considered "pure" belief and "pure" reinforcement learning, and nothing in between. For Chapter 4, we therefore consider a broader class of learning models and we try to find under which conditions, we can re-estimate three parameters of EWA learning model from simulated data, generated for different games and scenarios. The results show low rates of convergence of the estimation algorithm, and even if the algorithm converges then biased estimates of the parameters are obtained most of the time. Hence, we must conclude that re-estimating the exact parameters in a quantitative manner is difficult in most experimental setups. However, qualitatively we can find patterns that pinpoint in the direction of either belief or reinforcement learning.

Finally, in the last chapter, we study the effect of a player's social preferences on his own payoff in 2×2 games with only a mixed strategy equilibrium, under the assumption that the other player has no social preferences. We model social preferences with the Fehr-Schmidt inequity aversion model, which contains parameters for "envy" and "spite". Eighteen different mixed equilibrium games are identified that can be classified into Regret games, Risk games, and RiskRegret games, with six games in each class. The effects of envy and spite in these games are studied in five different status scenarios in which the player with social preferences receives much higher, mostly higher, about equal, mostly lower, or much lower payoffs. The theoretical and simulation results reveal that the effects of social preferences are variable across scenarios and games, even within scenario-game combinations. However, we can conclude that the effects of envy and spite are analogous, on average beneficial to the player with the social preferences, and most positive when the payoffs are about equal and in Risk games.

Acknowledgements

Though most people view the PhD position as a lonely one, I do not know where to begin thanking people.

I will start with my two supervisors without whom this thesis would be unreadable, less precise, and considered by many as the rampaging art of a mathematician. Thank you very much Chris and Marcel. In spite of your busy schedules, you were available or arranged to be available or reachable during the last years. Besides your many, many comments and feedback, you also took time to teach me principle methods and theories in science: psychology, sociology, economy, and even mathematics. Even after heated discussions or major disagreements, you always showed professionalism and humor to make sure we came to a mutual understanding. Moreover, you were always pleasant to work with, to have dinner with, to attend conferences with, and to play games with. Men, I will improve my game of chess.

Furthermore, I would like to thank my committee for all the work they have done to improve my manuscript.

Next, I would like to thank all my colleagues in Eindhoven and Tilburg for the time I worked with them. Not only have I greatly enjoyed lunches and discussions we had, it was always inspiring to discuss mathematical or social science related problems. I would particularly like to thank Marianne Jonker for always having her door open in Eindhoven. Next, I thank Jeroen Weesie and Isabel Canette for their help over-and-over-again with respect to understanding STATA. I owe them greatly for their help in continually increasing the performance of my re-estimation programs. Furthermore, I thank Dan Roozemond greatly for many helpful suggestions in all kinds of areas: mathematics, LaTeX, programming, and many more topics. Dan, you are so multi-talented that you are not only a good friend, but an enormous help for people in your surroundings. Finally, I would like to thank Uwe Matzat for a good cooperation in writing a paper (totally unrelated to this thesis) together and presenting it at conferences.

On a more personal level, I would like to thank Joke for keeping me focussed on my work when I needed it. I would like to thank my parents, my brothers, their partners, and children for allowing me to have such a great family throughout the year. Special thanks for my parents for having an unlimited supply of trust and faith in me and my work. Thanks to my brother Coen for making the booklet and front page of this thesis. Thanks in advance to Rob and Jacco for being my paranimphs and helping me through my defence. A particular thanks goes out to my nephew Bert Hoeks for all his interest in my research and the many suggestions

he gave during the last four years. It is a luxury to have someone in my vicinity with your expertise and interest to tell my research to.

Finally, I like to thank my wife Lieke, for innumerable things. Even though you did not understand much of my thesis, by explaining things to you, I always got my story straightened out and my mind clear again. You helped me through every stage, though you will not recognize half of what you have done to help my research and make this thesis.

Wouter van der Horst
Best, February 2011

Curriculum Vitae

Wouter van der Horst was born on July 7, 1982 in Oss, the Netherlands. He did his pre-university education at the Maasland College in Oss where he received his diploma in 1999.

After obtaining his first year certificates for Applied Mathematics and Computer Science at the Eindhoven University of Technology, he continued to study Applied Mathematics. In the spring of 2004, we wrote his Bachelor's thesis "*The (im)possibility of predicting the weather in the long run*" under supervision of prof. dr. Mattheij. That summer, he co-wrote a paper "*Increasing the production from horizontal oil wells by heating*" during the 18th ECMI modeling week in Lappeenranta. In the fall of 2004, he carried out his internship at TNO-TPD in Delft, where he worked on software that modeled the sound levels of acoustic liners on planes, a combined project with the faculty of Applied Physics. His master thesis continued this work by modeling and attempting to reduce these same sound levels under the supervision of Sjoerd Rienstra (Mathematics) and Mico Hirschberg (Physics). He graduated in August, 2005, after writing his Master's thesis "*Reduction of noise generated by the wings of an airplane*", receiving his Master's of Science degree in Industrial and Applied Mathematics, speciality scientific computing.

Less than a year before graduating, he dedicated himself to discovering how to apply mathematical skills on social science problems. After a first great meeting with dr. Spier in Amsterdam, he came in contact with prof. dr. Snijders and dr. van Assen and landed on a PhD position entitled "Development and application of learning models to behavior in interdependence situations". During the course of 2005-2009 he worked on this multi-disciplinary thesis both at the Eindhoven University of Technology as well as on the University of Tilburg attending several congresses throughout the world, writing papers related to this thesis and non-related, and presenting his work to fellow colleagues. His research focused on learning models in all kinds of interdependence situations. Apart from this work, Wouter likes to create, and perform in theater and –though contradictory sometimes– spends as much time as possible with his family, friends, and above all his wife.

After his work on his PhD thesis, Wouter started a new job at NSpyre, working as a software engineer at the Metrology department of ASML.

Bibliography

Agell, J. & Lundborg, P. (1995). Theories of Pay and Unemployment: Survey Evidence from Swedish Manufacturing Firms. *Scandinavian Journal of Economics*, XCVII, pp. 295-308.

Akerlof, G.A. (1982). Labor contracts as partial gift exchange. *The quarterly journal of economics*, XCVII, november 1982, No.4.

Ashworth, T. (1980). Trench Warfare, 1914-1918: the live and let live system. New York, Holmes & Meier

Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books.

Becker, G.S. (1974). A Theory of Social Interactions. Journal of Political Economy, 82(6), pp. 1063-1093

Becker, G.S. (1976). The Economic Approach to Human Behavior. University of Chicago Press.

Bereby-Meyer, Y. & Erev, I. (1998). On learning to become a successful loser: a comparison of alternative abstractions of learning processes in the loss domain. *Journal of Mathematical Psychology*, 42, pp. 266-286.

Bewley, T.F. (1995). A Depressed Labor Market as Explained by Participants. *The American Economic Review*, Vol. 85, No. 2.

Binmore, K. (1992). Fun and games. D. C. Heath and Company, Lexington, Massachusetts

Bisin, A. & Verdier, T. (2001). The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory*, 97, pp. 298-319.

Blinder, A.S. & Choi, D.H. (1990). A Shred of Evidence on Theories of Wage Stickness. *Quarterly Journal of Economics*, CV, pp. 1003-1016.

Blume, A., Dejong, D.V., Neumann, G.R., & Savin, N.E. (2002). Learning and communication in sender-receiver games: an econometric investigation. *Journal of applied econometrics*, 17, pp. 225-247

Bolton, G.E., Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, 2000

Borgers, T. & Sarin., R. (2000). Naive reinforcement learning with endogenous aspirations. *Internatiol Economic Review*, 41-4, November 2000.

Bosch-Domenech, A., Montalvo, J.G., Nagel, R., & Satorra, A. (2002). One, Two, (Three), Infinity, ... : Newspaper and Lab Beauty-Contest Experiments. *The American Economic Review*, Vol. 92, No. 5, pp. 1687-1701

Bush, R.R. & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, Vol. 58, pp. 313-323.

Cabrales, A. & Garcia-Fontes, W. (2000). Risk dominance selects the leader: An experimental analysis. *International Journal of Industrial Organization*, Vol. 18-1, January 2000, pp. 137-162

Cachon, G.P. & Camerer, C.F. (1996). Loss-Avoidance and Forward Induction in Experimental Coordination Games. *The Quarterly Journal of Economics*, Vol. 111-1 (Feb., 1996), pp. 165-194.

Camerer, C.F. (2003). Behavioral game theory: experiments in strategic interaction. Russell Sage Foundation: Princeton University Press.

Camerer, C.F. & Ho, T.H. (1999). Experience-weighted attraction learning in normal-form games. *Econometrica*, Vol. 67, pp. 827-874.

Camerer, C.F., Ho, T.H., & Chong, J.K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, Vol. 104, Issue 1, May 2002, pp. 137-188.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, August 2002

Cheung, Y-W. & Friedman, D. (1997). Individual learning in normal form games: some laboratory results. *Games and Economic Behavior*, Vol. 19, pp. 46-76.

Cox, J.C., Shachat, J., & Walker, M. (2001). An Experiment to Evaluate Bayesian Learning of Nash Equilibrium Play. *Games and Economic Behavior*, Vol. 34, Issue 1, pp. 11-33

Dekel, E. (2007). Evolution of preferences. *Review of economic studies*, 74, pp. 685-704.

Diekmann, A. & Voss, T. (2003). Social Norms and Reciprocity. Presented during conference Sektion Modellbildung und Simulation, Leipzig, October 2002

Engelmann, D. & Strobel, M. (2004). Inequality Aversion, Efficienty, and Maximin Preferences in Simple Distribution Experiments. *The American Economic Review*, Vol. 94, No. 4. (Sep., 2004), pp. 857-869.

Engelmann, D. & Steiner, J. (2007). The effects of risk preferences in mixed-strategy equilibria of 2×2 games. *Games and Economic Behavior*, Volume 60, Issue 2, August 2007, pp. 381-388

Erev, I. & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitivite strategies. *Psychology Review*, Vol. 112, pp. 912-931.

Erev, I. & Haruvy, E. (2005). Generality, repetition, and the role of descriptive learning models. *Journal of Mathematical Psychology*, Vol. 49, pp. 357-371.

Erev, I. & Roth, A.E. (1998). Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, Vol. 88, pp. 848-881.

Erev, I., Bereby-Meyer, Y., Roth, A.E. (1999). The effect of adding a constant to all payoffs: experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior & Organization*, Vol. 39 (1999), pp. 111-128.

Erev, I., Roth, A.E., Slonim, R.L., & Barron, G. (2007). Learning and equilibrium as useful approximations: Accuracy of prediction on randomly selected constant sum games. *Economic Theory*, Vol. 33, pp. 29-51.

Fehr, E. & Gintis, H. (2007). Human motivation and social cooperation: experimental and analytical foundations. *Annual Review of Sociology*, 33, pp. 43-64.

Fehr, E, Schmidt, K.M. (1999). A Theory of Fairness, Competition and Cooperation. *The Quarterly Journal of Economics*, August, 1999

Feltovich, N. (2000). Reinforcement-Based vs. Beliefs-Based Learning in Experimental Asymmetric-Information Games. *Econometrica*, Vol. 68, No. 3 (May, 2000), pp. 605-641.

Feltovich, N., A. Iwasaki, and H.S. Oda (2008). Payoff levels, equilibrium selection, and learning: an experimental study of the stag hunt. Working paper, University of Aberdeen.

Flache, A. & Macy, M.W. (2001). Stochastic collusion and the power law of learning: a general reinforcement model of cooperation. *Journal of Conflict Resolution*, 46, pp. 629-653.

Fudenberg, D. & Kreps, D. (1995). Learning in Extensive Games, I: Self-Confirming Equilibrium. *Games and Economic Behavior*, 8, pp. 20-55.

Fudenberg, D. & Levine, D.K. (1998). The theory of learning in games. Cambridge, MIT Press

Geanakoplos, J., Pearce, D. & Stacchetti, E., (1989). Psychological Games and Sequential Rationality. *Games and Economic Behavior*, 1, pp. 60-79.

Gould, W., Pitblado, J., & Poi, B. (2010). Maximum Likelihood Estimation with Stata, Fourth Edition. Stata Press.

Gourieroux, C., Monfort, A. (1995), *Statistics and Econometric Models, volume one*, Cambridge University Press, Cambridge.

Güth, W. & Tietz, R. (1990). Ultimatum Bargaining Behavior - A Survey and Comparison of Experimental Results. *Journal of Economic Psychology*, XI, pp. 417-449.

Güth, W. & Yaari, M. (1992). An evolutionary approach to explain reciprocal behavior in a simple strategic game. *Explaining Process and Change – Approaches to Evolutionary Economics*, Univ. of Michigan Press, Ann Arbor, pp. 23-34

Güth, W. (1995). An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives. *International Journal of Game Theory*, 24, pp. 323-344.

Haruvy, E., & Erev, I. (2002). Interpreting parameters in learning models. In R. Zwick, & A. Rapoport (Eds.), Experimental business research (pp. 285-300). Dordrecht: Kluwer Academic Publishers.

Ho, T., Camerer, C., & Chong, K. (2002). Functional EWA: a one-parameter theory of learning in games. Caltech Working Paper.

Homans, G. (1961). Social Behaviour: Its Elementary Forms. London, Routledge and Kegan Paul.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, vol. 70, pp. 2141-2166.

Ichimura, H. & Bracht, J. (working paper). Estimation of Learning Models on Experimental Game Data. *working paper.*

Kahneman, D., Knetsch, J.L., & Thaler, R. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review*, LXXVI, pp. 728-741.

Kockesen, L., Ok, E.A., & Sethi, R. (2000). Evolution of Interdependent Preferences in Aggregative Games. *Games and Economic Behavior*, vol. 31(2), pp. 303-310, May.

Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *American Economic Review*, vol. 90, issue 4, pp. 1072-1091

Macy, M. & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, Vol. 99, pp. 7229-7236. Poundstone (1992)

Macy, M. W. & Flache, A. (1995). Beyond Rationality in Models of Choice. *Annual Review of Sociology*, 21, pp. 73-91.

Macy, M. W. & Flache, A. (2009). Social dynamics from the bottom up. Agent-based models of social interaction. In P. Hedström & P. Bearman (Eds.), The Oxford Handbook of Analytical Sociology (pp. 245-268). Oxford: Oxford University Press.

Mookerjee, D. & Sopher B. (1994). Learning behavior in an experimental matching pennies game. *Games and Economic Behavior*, Vol. 7, pp. pp. 62-91

Mookerjee D. & Sopher B. (1997). Learning and decision costs in experimental constant sum games. *Games and Economic Behavior*, Vol. 19, pp. 97-132.

Okun, A. (1981). Prices and Quantities: A Macroeconomic Analysis. *Washington: The Brookings Institution*, 1981.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, Vol. 10, No. 3, pp. 137-158.

Poulsen, A., Poulsen, O. (2002). Social Preferences in the Prisoner's Dilemma Game: an Evolutionary Analysis. *Aarhus School of Business.*

Poundstone, W. (1992). Prisoner's Dilemma. Doubleday, NY NY.

Rabin, M. (1993), Incorporating Fairness into Game Theory and Economics. *American Economic Review*, LXXXIII, pp. 1281Ű1302.

Rabin, M. (2006), The experimental study of social preferences. *Social research*, 73, pp. 405-428.

Roth, A.E. & Erev, I. (1995). Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior*, Vol. 8, No. 1, pp. 164-212.

Roth, A.E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *The American Economic Review*, Vol. 81, No. 5, pp. 1068-1095.

Rustichini, A. (1999). Optimal Properties of Stimulus Response Learning Models. *Games and Economic Behavior*, 29, pp. 244-273.

Rydval, O. & Ortmann, A. (2005). Loss avoidance as selection principle: Evidence from simple stag-hunt games. *Economics Letters*, Volume 88, Issue 1, July 2005, pp. 101-107.

Sally, D. (1995). Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992. *Rationality and Society*, 1995, 7, pp. 58.

Salmon, T. (2001). An evaluation of econometric models of adaptive learning. *Econometrica*, Vol. 69, pp. 1597-1628.

Salmon, T. (2004). Evidence for learning to learn behavior in normal form games. *Theory and Decision*, Vol. 56, pp. 367-404.

Samuelson, L. (2001), Introduction to the Evolution of Preferences. *Journal of Economic Theory*, Vol. 97, pp. 225-230

Sarin, R. and F. Vahid (1999). A Payoff assessments without probabilities: A simple Dynamic Model of Choice. *Games and Economic Behavior*, 28, pp. 294-309.

Schwartz, B. & Reisberg, D. (1991). Learning and Memory. New York and London: W.W. Norton & Co.

Skinner, B.F. (1938). Behavior of organisms. New York: Appleton-Century-Crofts.

Snijders, C. & Raub, W. (1996). Does 'The motivating power of loss' exist? An experimental test of the effect of losses on cooperation. In Wim B.G. Liebrand and David M. Messick (eds.), Frontiers in Social Dilemmas Research, Berlin: Springer Verlag, pp. 205-214.

Stahl, D.O. (2003). Action-reinforcement learning versus rule learning. Mimco, University of Texas

Thaler, R.H., & Johnson, E.J. (1990). Gambling with the house money and trying to break even: he effects of prior outcomes on risky choices. *Management Science*, 36, pp. 643-660.

Train, K. E., McFadden, D. L. and Goett, A. A. (1987). Consumer Attitudes and Voluntary Rate Schedules for Public Utilities. *Review of Economics and Statistics*, 64, pp. 383-91.

Tsebelis, G., (1989). The Abuse of Probability In Political Analysis: The Robinson Crusoe Fallacy. *The American Political Science Review*, Vol. 83, No. 1. pp. 77-91.

van Assen, M.A.L.M., Snijders, C.C.P., Weesie, J. (2006). Behavior in repeated prisoner's dilemma games with shifted outcomes analyzed with a statistical learning model. *Journal of Mathematical Sociology*, 30 (2), pp. 159-180.

Wilcox, N.T. (2006). Theories of learning in games and heterogeneity bias. *Econometrica*, Vol. 74, No. 5 (September, 2006), pp. 1271-1292.