The matching law and melioration learning

From individual decision-making to social interactions

von der Fakultät für Sozialwissenschaften und Philosophie
der Universität Leipzig

genehmigte

DISSERTATION

zur Erlangung des akademischen Grades

doctor rerum politicarum

(Dr. rer. pol.),

vorgelegt

von Johannes Zschache, M.A.

Gutachter: Prof. Dr. Thomas Voss, Prof. Dr. Andreas Flache

Tag der Verleihung: 20. Dezember 2016

Contents

List of Tables

List of Figures

Foreword

This thesis is the result of my doctoral studies in sociology during the years 2011-2016 at Leipzig University. The topic has evolved continuously during these years. The initial proposal was entitled "The structure and dynamics of reciprocal exchange networks." In particular, I planned to base the behavioural model of reciprocal exchange on the principles of operant conditioning, as it was intended by the pioneers of exchange theory George C. Homans and Richard M. Emerson.

Multiple obstacles formed during my first attempts to model the evolution of reciprocal exchange networks. Without the help of my supervisor, my colleagues, and the participants of multiple workshops and conferences, this thesis would have never been completed or look very different.

It is a pleasure to thank those who made this work possible. Especially important was my supervisor Thomas Voss, who suggested the matching law as a foundation of the model of reciprocal exchange. His comments on and trust in my work provided the support and freedom that I needed to pursue different lines of thought and to finish this thesis with the present topic.

I am also grateful for the advice and comments of Roger Berger, Andreas Tutić, and the participants of our colloquium, who listened patiently to my talks about various aspects of the matching law.

The computer simulations were run on machines of the University Computer Centre and on a 16-cores workstation, which was purchased especially for this purpose. Furthermore, this thesis would not have been possible without diverse software, such as Java, Scala, Akka, NetLogo, LaTeX, GNU R, and ggplot2.

I would also like to show my gratitude to "the giants on which shoulders I stand" (see the list of references).

Chapter 1

Introduction and overview

When adopting the explanatory framework of methodological individualism, assumptions about the actors' decision-making are required (Coleman, 1990, p. 11). Even though there are different versions of methodological individualism (Udehn, 2002), the actors' properties and decisions are always regular in at least some relevant aspects. These regularities allow to formulate rules or axioms of behaviour and, therefore, promote the explanation of social phenomena.

Different rules of individual behaviour are applicable in sociological theories that build on methodological individualism. For example, the assumption of rational behaviour has been borrowed from economic theories and used repeatedly in sociology (e.g. Coleman, 1990; Gintis, 2009; Braun and Gautschi, 2011). In many situations, this assumption facilitates the explanation of social phenomena because solution concepts that are based on rational behaviour can be readily transferred from economics to sociology.

Next to economic theory, also behavioural psychology has provided a basis for assumptions about individual behaviour. Well-known sociologists such as George C. Homans (1961), Richard M. Emerson (1972a), Karl-Dieter Opp (1972), John H. Kunkel (1975), and many others (Burgess and Bushell, 1969; Hamblin and Kunkel, 1977) employed principles from behavioural psychology in order to study social phenomena. Influences from behaviourism are still visible in contemporary theories of social exchange (Molm, 2006, p. 29) or backward-looking rationality (Macy and Flache, 2009, pp. 250-251).

In behavioural psychology, two main forms of learning account for regularities in behaviour. One of them is called *classical conditioning*. It describes the emergence of a natural reflex in connection with a previously neutral stimulus. For example, in a famous experiment with dogs, the salivation reflex was conditioned to emerge in connection with the sound of a bell (Pavlov, 1927). In 1937, B. F. Skinner marked out a different learning process, named *operant conditioning*. In contrast to a conditioned reflex, any kind of behaviour can be acquired by operant conditioning. Learning takes place as soon as "a reinforcer [..] follows upon the organism's own behaviour" (Skinner, 1953, p. 65).

The theory of operant conditioning is in opposition to cognitive theories of rational choice. An action is fully explained by its previous reinforcements, and there is no need to consider thoughts, emotions, or any other cognitive states (Rachlin and Laibson, 1997, p. 7). Similar to the process of natural selection in the evolution of species, "the causal processes producing the behavior [..] are instances of *selection by consequences*" (Ringen, 1999, p. 168, italics added). Instead of understanding an observed action as the result of cognitive processes and anticipation, it is seen as being controlled by past events.

Richard Herrnstein, who was a student of Skinner, refined the ideas of operant conditioning. He argued that behaviour depends not only on the occurrence of a timely proximate reinforcer but, more specifically, on the general rate of reinforcements (Herrnstein, 1969). In support of this hypothesis is a widely observed empirical regularity, which has been called the **matching law** (Herrnstein, 1997). According to this law, the relative frequency of a particular action equals the relative frequency of its reinforcements.

Since the matching law describes a regularity in decision-making, it is, similar to the assumption of rationality, a potential foundation of individual behaviour in sociological theories.

> Hence, the topic of this thesis is the application of the matching law as micro-level assumption in the explanation of social phenomena.

Although the matching law was repeatedly shown to hold in social situations (e.g. Conger and Killeen, 1974; Hamblin, 1977, 1979; Sunahara and Pierce, 1982; McDowell, 1988; Borrero et al., 2007), it has not been used to derive hypotheses about social phenomena yet. Two exceptions are the work of Louis N. Gray and colleagues (Gray and von Broembsen, 1976; Gray et al., 1982), which is reviewed in section 2.2.4, and a paper on learning in games by Brenner and Witt (2003), which is discussed in section 4.4.

A possible reason for the limited sociological interest in the matching law is the lack of a formal framework that allows the theoretical derivation of hypotheses. In case of the rationality assumption, such a framework was provided by economic theories, for example by general equilibrium theory or game theory. It is demonstrated in the following chapters that some parts of these frameworks can be employed with the matching law as well. But additional concepts must be introduced, and restricting assumptions have to be made.

Figure 1.1 illustrates the undertaking of this thesis in reference to Coleman's (1990, p. 646) micro-macro scheme.

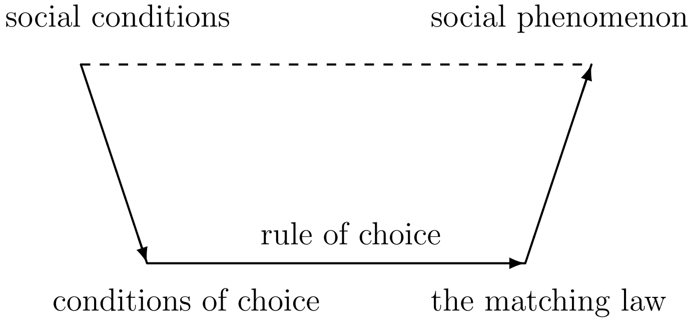


Figure 1.1: The matching law in Coleman's micro-macro scheme

The matching law takes the place of the outcome of individual decision-making. The transition from individual outcomes to a social phenomenon is usually complex. A particular problem in case of the matching law is its reference to relative frequencies and, thus, to an aggregated measure of a sequence of decisions. This point is clarified by figures 1.2 and 1.3.

In the diagram of figure 1.2, an actor is assumed to choose repeatedly one of several alternatives by following a fixed rule. The circles stand for points in time. A horizontal arrow from one circle to another displays a choice. As indicated by the

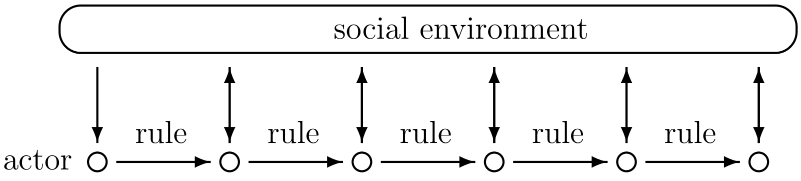


Figure 1.2: The repeated interaction of an actor with the social environment

vertical arrows, the choice alternatives and the result of a decision are conditioned by the social environment. At the same time, the actions may have an effect on the social environment. It cannot be assumed that decisions are made independently. In contrast, the choice at one point may affect a later decision directly, for example by a learning process, or indirectly via the social environment.

Figure 1.3 narrows the previous diagram down to the interaction between multiple actors, which is seen as part of the social environment. The actors influence each other in their decisions. It is indicated that the matching law (ML) refers to relative frequencies of a sequence of decisions (marked by the horizontally stretched ovals). For instance, the matching law may state that actor 1 chooses a particular alternative in 2 out of 5 decisions. But it is not known, which alternative is chosen at a particular point in time. Neither does the matching law imply a stochastic model that allows the specification of probabilities of choice.

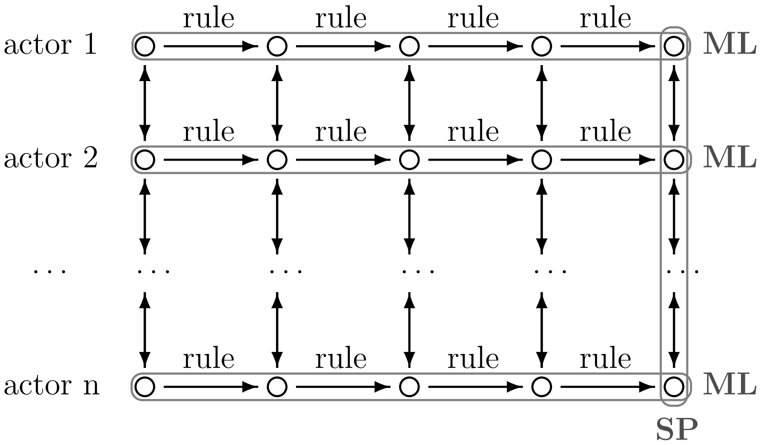


Figure 1.3: The repeated interaction of multiple actors

A social phenomenon (SP) commonly refers to a cross-sectional or longitudinal measurement of individual properties or decisions. Therefore, it corresponds to one or several ovals that vertically stretch over the circles of the diagram. It is apparent that a vertically recorded social phenomenon is not compatible with the horizontally spreading matching law. No macro-level derivation can be made from the matching law because nothing is known about the decisions of the actors at a particular point in time.

A solution to this problem is the establishment of a rule of decision-making (see figure 1.3) that results in the matching law on the individual level and can be used to derive outcomes on the social level. Multiple rules exist that meet these requirements. With **melioration learning**, a relatively simple one is presented and analysed in this thesis.

The next chapter introduces and illustrates the matching law. An experiment that led to the formulation of this law is summarised, and various extensions are discussed: the generalised matching law, the incorporation of delay, the law of response strength, and social matching. Chapter 2 also gives a short overview of empirical studies that detected the matching law in human behaviour.

In previous research, the matching law was empirically justified by fitting its generalised version to data from experiments. Because the parameters of this model were not specified beforehand, the matching law was not tested statistically. Instead, only the "goodness of fit" was evaluated after the parameters had been estimated from the data.

In order to use the matching law as micro-level assumption, its parameters must be derived from situational properties. This requires a new theoretical perspective. In chapter 3, such a perspective is presented by integrating the matching law into economic consumer theory. Given an adequate definition of a situation, this approach allows to theoretically derive the parameters, to predict outcomes of individual decisions, and to specify empirically testable and falsifiable hypotheses.

The integration into economic consumer theory also facilitates the comparison of the matching law to standard predictions of behaviour. More specifically, it is shown in chapter 4 that optimal behaviour corresponds to the matching law in a certain class of situations. But generally, the predictions of the matching law are not optimal. It can be regarded as an alternative explanation of social behaviour if standard economic solutions fail.

In chapter 5, a rule of decision-making is introduced that is supposed to result in the matching law. Following earlier work on this subject (Herrnstein and Vaughan, 1980), this rule is called melioration learning. Basically, it is assumed that individuals always choose one of the alternatives with the currently highest average value. The average value of an alternative is obtained from previous experiences. It is shown that this learning process converges to the matching law if the situation is sufficiently stationary.

However, if multiple actors interact with each other, the situation is generally non-stationary, and the behaviour of actors who learn by melioration may never converge. Therefore, computer simulations are used to analyse the long-term dynamics of interactive melioration learning. Various two-person situations that are known from game theory are considered in chapter 6. It is demonstrated that the actors learn to play a dominant strategy or one of the pure Nash equilibria. If no pure equilibrium exists, the relative frequencies of choice converge to the mixed Nash equilibrium in some of the games.

Situations with more than two actors are examined in chapter 7. First, everyone interacts simultaneously with several partners in a coordination game. This model allows to explore the evolution of social conventions and institutions (Young, 1998). The simulations reveal that the network structure of interactions affects a group's ability to coordinate its members' choices. Additionally, it is shown that the actors can learn to volunteer in the volunteer's dilemma and to cooperate in a multi-person prisoner's dilemma. In the latter case, the option to punish defectors or to abstain from the interaction further mitigates the dilemma.

The last chapter presents an evolutionary justification of the matching law. Since behaviour that conforms to this law is not necessarily optimal, it is disputable that such a fundamental behavioural regularity has evolved by natural selection. Rules of choice that guarantee optimal results should have replaced rules that lead to the matching law. Nevertheless, melioration learning is shown to be evolutionary advantageous in uncertain competitive environments.

Chapter 2

The matching law

This chapter contains an incomplete summary of the extensive research on the matching law. In section 2.1, its first version, which will be called the *strict matching law*, is introduced as an empirical regularity of individual behaviour. Because systematic deviations from the equation of the strict matching law were discovered experimentally, extensions to larger families of equations have been established. Some of these extensions are described in section 2.2. For example, the *generalised matching law* uses free parameters to fit the equation of the strict matching law to observed behaviour. Further generalisations regard the delay of reinforcement, the absolute frequency of choice, and social matching.

A short overview of empirical studies is given in section 2.3. While most experiments were conducted with animals, such as pigeons or rats, this overview concentrates on evidence of the matching law in human behaviour.

2.1 The strict matching law

One of the first experiments that led to the formulation of the matching law was reported by Richard Herrnstein (1961). In this experiment, pigeons were placed in a box with two response-keys, denoted as key 1 and key 2. Food was released occassionally as a reinforcement of pecking on one of the response-keys. During the experiment, the absolute frequencies of pecks k_1 and k_2 and the corresponding absolute frequencies of reinforcements s_1 and s_2 were recorded for each key. After plotting these values for different schedules of reinforcement, an *approximate matching* of the relative frequencies of choice $\frac{k_1}{k_1+k_2}$ to the relative frequencies of reinforcement $\frac{s_1}{s_1+s_2}$ was observed:

$$\frac{k_1}{k_1+k_2} = \frac{s_1}{s_1+s_2}. \tag{2.1}$$

Equation (2.1) is called the **strict matching law**. Alternatively, the following condition of the strict matching law is commonly found in the literature:

$$\frac{s_1}{k_1} = \frac{s_2}{k_2}. \tag{2.2}$$

Condition (2.2) is equivalent to condition (2.1) if $k_1, k_2, s_1, s_2 > 0$.

The strict matching law can be illustrated by the penalty kick situation of (European) football games. The kicker is assumed to choose between the left and the right side of the goal. The reinforcement of either choice depends, among other things, on the action of the goal-keeper. The player scores by kicking the ball into the goal and fails if the ball misses the goal, hits the posts, or is blocked by the keeper. Let us assume one particular player who was engaged in 50 penalty kicks. A hypothetical distribution of choices and reinforcements is shown in table 2.1.

Table 2.1: A sample distribution of choices in penalty kick situations

choice of kicker	left (k_1)		right (k_2)	
	40		10	
reinforcement	success (s_1)	failure	success (s_2)	failure
	24	16	6	4

According to equation (2.1), the strict matching law holds in this example because

$$\frac{k_1}{k_1+k_2} = \frac{40}{50} = 0.8 = \frac{24}{30} = \frac{s_1}{s_1+s_2}.$$

Also equation (2.2) holds:

$$\frac{s_1}{k_1} = \frac{24}{40} = 0.6 = \frac{6}{10} = \frac{s_2}{k_2}.$$

2.2 Extensions of the strict matching law

The strength of the strict matching law is its simplicity and wide applicability. It is supposed to describe behaviour in any situation in which an individual chooses repeatedly between two *similar* alternatives. There has been extensive research on the empirical validity of the matching law. Some of this research is summarised in section 2.3. Further overviews are given in Herrnstein (1997), Mazur (2001, ch. 14), Poling et al. (2011), or Reed and Kaplan (2011). Reviews of empirical studies are, for example, written by de Villiers and Herrnstein (1976), Baum (1979), Hamblin (1979), Pierce and Epling (1983), and McDowell (2005).

In spite of several empirical confirmations of the strict matching law, the researchers agree that it often fails to describe behaviour. As summarised by Baum (1974, 1979), most of the deviations originate from the presence of *asymmetries* in choice alternatives or reinforcements. For instance, if the choice of one alternative requires additional effort or if the reinforcements differ in their amount, quality, or deprivation rate, the strict matching law fails.

The detection of systematic deviations led to the extension of the strict matching law to a set of equations, which has been called the *generalised matching law.* As indicated in section 2.2.1, this generalised version is able to account for most of the deviations. Section 2.2.2 presents the incorporation of another widely observed empirical phenomenon, which is known as sensitivity to delay. An extension of the matching law that considers absolute frequencies of choice is discussed in section 2.2.3. Finally, a previous attempt to extend the individual matching law to a social matching law is described in section 2.2.4.

2.2.1 The generalised matching law

According to condition (2.2), the strict version of the matching law requires equality between the rates of reinforcement. If $s_2 > 0$, this is equivalent to

$$\frac{k_1}{k_2} = \frac{s_1}{s_2}.$$

The **generalised matching law** (e.g. Baum, 1974; McDowell, 2013a) accounts for systematic deviations from this condition by adding two parameters $\alpha, \beta \in (0, \infty)$:

$$\frac{k_1}{k_2} = \beta \cdot \left(\frac{s_1}{s_2}\right)^{\alpha}. \tag{2.3}$$

It is said that the generalised matching law holds if there exist $\alpha, \beta \in (0, \infty)$ such that condition (2.3) is true. This means that the generalised matching law corresponds to a set of indefinitely many equations.

Figure 2.1 illustrates the effect of different values of β and α on the relationship between the relative frequencies of choice and reinforcement (equation (2.1)):

$$\frac{k_1}{k_1 + k_2} = \frac{\beta \cdot s_1^{\alpha}}{\beta \cdot s_1^{\alpha} + s_2^{\alpha}}.$$

The curves show the function $y := f\left(\frac{s_1}{s_1+s_2}\right) = \frac{\beta \cdot s_1^{\alpha}}{\beta \cdot s_1^{\alpha} + s_2^{\alpha}}$. The diagonals indicate this relation in case of strict matching ($\beta = \alpha = 1$). The left-sided graph contains the generalised matching law with $\alpha = 1$ and different values of β. In the right-sided graph, β is set to 1, and α is varied between 0.1 and 10.

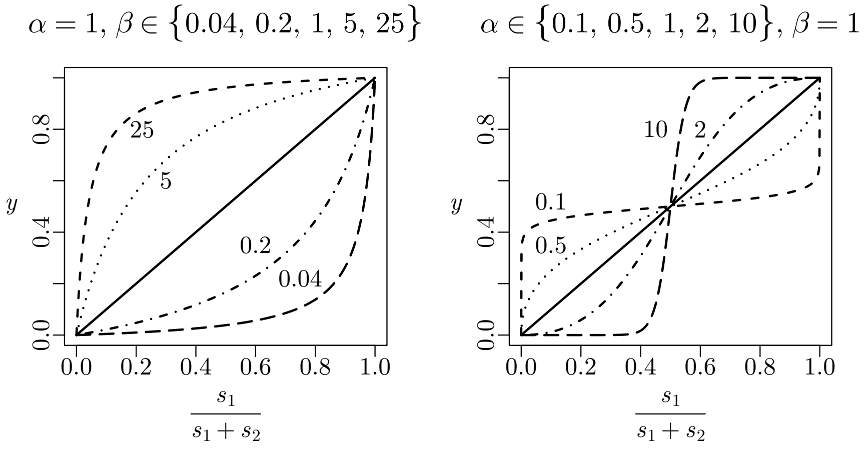


Figure 2.1: The generalised matching law with different values of β and α

By setting $\beta \neq 1$, the generalised matching law captures *bias*, which is a systematic preference for one of the alternatives (Baum, 1974). As pictured by the left-sided graph of figure 2.1, the first alternative is chosen more often than predicted by the strict matching law if $\beta > 1$. If $\beta < 1$, the second alternative appears more frequently. Baum (1974) mentioned different sources of bias. First, an individual may prefer one alternative because of additional effort that comes with the other one. For example, if the individual is left-handed, a left-sided response key is more easily accessible than a right-sided key. Second, the resources that are received as reinforcements may differ in their amount or quality. In the experiments of Herrnstein (1961), the pigeons received always the same amount of the same kind of grain. But, if pecking on one key results in a more valuable resource or in a greater amount of grain than pecking on the other key, the former is likely to be pecked more often than predicted by the relative frequency of reinforcement.

Also without the parameter β, differences in amount of reinforcement were taken into account in the past. On the one hand, the variables s_1 and s_2 of equation (2.1) were interpreted as the aggregated amount of all reinforcers instead of the absolute frequency of reinforcement (e.g. de Villiers and Herrnstein, 1976). On the other hand, differences in amount were explicitly modelled and added to the equation (Rachlin, 1971). Let $b_1, b_2 \in [0, \infty)$ indicate the average amounts of the resources that are consumed during a reinforcement of alternative 1 and 2, respectively. An extended version of the strict matching law is given by

$$\frac{k_1}{k_1 + k_2} = \frac{b_1 \cdot s_1}{b_1 \cdot s_1 + b_2 \cdot s_2}, \tag{2.4}$$

or

$$\frac{k_1}{k_2} = \frac{b_1 \cdot s_1}{b_2 \cdot s_2}. \tag{2.5}$$

In comparison to the generalised matching law of equation (2.3), this approach suggests that bias is captured by setting $\beta = \frac{b_1}{b_2}$ or $\beta = \left(\frac{b_1}{b_2}\right)^{\alpha}$.

The difficulty of evaluating this explanation of bias is the measurement of actually consumed resources. Consider, for example, the experiment of Fantino et al. (1972), in which pigeons chose between two response-keys. In contrast to the experiment of Herrnstein (1961), the reinforcements differed in their maximally available amount of resources (6 vs. 1.5 seconds access to food). The authors included these differences by multiplying “the number of reinforcements observed on that schedule [..] by the duration of each reinforcer” (Fantino et al., 1972, p. 40). In respect to the notations above, it was assumed that $b_1 = 6$ and $b_2 = 1.5$, even though b_1 and b_2 stand for the average amounts of consumed grain and not the average amounts of grain that could be maximally consumed. If a pigeon's actual consumption differed from the maximally possible consumption, the authors may have rejected this explanation of bias by mistake.

The costs of choosing an alternative is another supposed cause of bias and can be included similarly. Gray and Tallman (1984) suggested that the total costs $c_i \in (0, \infty)$ that come with choosing an alternative $i \in \{1, 2\}$ should be added inversely proportionally to equation (2.5):

$$\frac{k_1}{k_2} = \frac{c_2}{c_1} \cdot \frac{b_1 \cdot s_1}{b_2 \cdot s_2}. \tag{2.6}$$

In a series of studies (Gray and Tallman, 1984; Stafford et al., 1986; Gray et al., 1991; Judson and Duran-Aydintug, 1991), equation (2.6) was shown to fit experimental data more accurately than alternative behavioural models that include the effects of costs or punishment. It should be noted that, in contrast to the interpretation of b_i, the term c_i denotes the total costs of choosing $i \in \{1, 2\}$, and $\frac{c_i}{k_i}$ stands for the average costs per choice.

The second parameter α of equation (2.3) accounts for two systematic deviations that have been called *undermatching* and *overmatching* (Baum, 1979). In case of overmatching, alternatives with high relative frequencies of reinforcement are chosen more often than predicted by the strict matching law. In the right-sided graph of figure 2.1, this is modelled by $\alpha > 1$ and depicted by curves that are close to $y = 0.0$ if $\frac{s_1}{s_1+s_2} < 0.5$ and close to $y = 1.0$ if $\frac{s_1}{s_1+s_2} > 0.5$. In contrast, undermatching stems from a systematic preference for the less reinforced alternative and is modelled by $\alpha < 1$. This is shown by curves that are close to $y = 0.5$.

In experimental studies, undermatching was observed more often than overmatching. For example, undermatching occurred if a frequent switching between the alternatives was possible (Baum, 1979). In most experiments, a changeover delay (COD) was included that punished a switch between the response-keys by a delay in the next reinforcement. Herrnstein (1970) referred to an experiment of Shull and Pliskoff (1967) when stating that the strict matching law was observed if the COD was neither too small nor too large. Otherwise, undermatching (small COD) or overmatching (large COD) was found.

Another reason of undermatching is the presence of interrelated deprivation rates (Baum and Nevin, 1981; Green and Freed, 1993). If the consumption of one resource increases the demand for another resource (e.g. the consumption of food may increase the demand for water), an increase in reinforcement of one alternative raises the frequency of choosing the other one. This results in undermatching or even in an inversion of the matching relation ($\alpha < 0$), which means that the less reinforced alternative is chosen more often than the highly reinforced alternative. Interrelated deprivation rates are further discussed in section 4.2.

Moreover, since undermatching implies that low relative frequencies of choice are adjusted to somewhat higher levels, this behaviour might be interpreted as *experimenting*. Theoretical support for this conjecture was given in the work of McDowell (2013b), who modelled individual decision-making as an evolutionary process (see also section 5.3.4). It was shown that the undermatching parameter α decreases if the amount of experimental behaviour (implemented as mutation rate) increases (McDowell and Caron, 2007, p. 102). Although this effect is not linear (McDowell and Popa, 2010, p. 251) and interacts with other components of the model, empirical support for this explanation of undermatching is found in some of the studies that are mentioned in section 2.3.

2.2.2 Delay, hyperbolic discounting, and self-control

Another extension of the matching law, which is often discussed in the literature, includes time delay. If the reinforcers of an alternative $i \in \{1, 2\}$ have a total amount of $a_i := b_i \cdot s_i$, but every reinforcement is delayed by a time period $d_i \in [0, \infty)$, this factor can be included by **hyperbolic discounting**. According to Mazur and Herrnstein (1988), this means that the subjective value v_i of alternative $i \in \{1, 2\}$ is given by

$$v_i := \frac{a_i}{1 + \delta \cdot d_i}, \tag{2.7}$$

with a scale factor $\delta \in [0, 1]$. Subsequently, the strict matching law is extended to

$$\frac{k_1}{k_1 + k_2} = \frac{v_1}{v_1 + v_2}. \tag{2.8}$$

There are several implications of the assumption of equation (2.7). First, time delay has a *discounting effect* on the value of a reinforcer. The left-sided plot of figure 2.2 illustrates this effect for the choice between two alternatives 1 and 2 with $a_1 = 10$ and $a_2 = 20$. Alternative 2 (solid line) is preferred to alternative 1 as long as there is no difference in delay ($d_1 = d_2$). But this preference relation changes if the delay of the reinforcement of alternative 2 increases. The two straight lines indicate that, if $d_1 = 2$, a delay of $d_2 > 5$ sets the value of alternative 2 below the value of alternative 1.

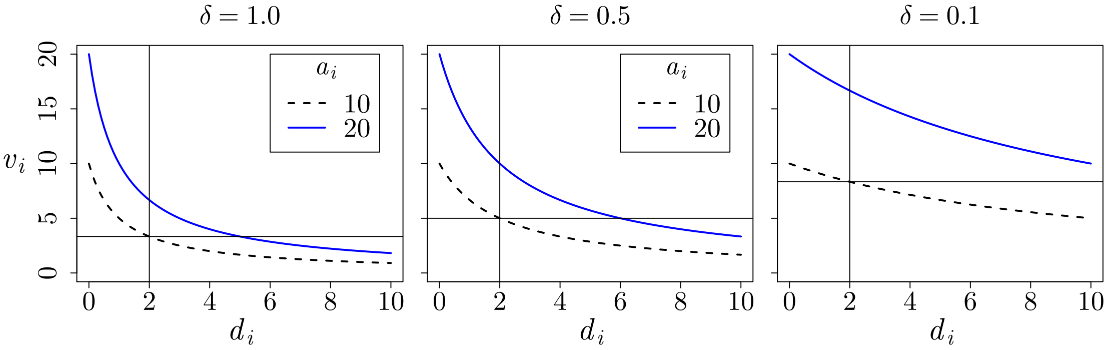


Figure 2.2: The value v_i in dependence of delay d_i and sensitivity to delay δ

A second implication concerns the effect of δ, which is interpreted as an actor's *sensitivity to delay* or *impulsiveness*. The left-sided plot of figure 2.2 shows an actor who is maximally sensitive to delay ($\delta = 1.0$). The actor's preferences depend on the differences in delay. If the sensitivity δ approaches zero, the actor becomes more tolerant in regard to delay. In fact, if $\delta = 0.0$, the two curves are horizontal lines with $v_2 > v_1$ for all values of d_1 and d_2. Herrnstein (1997, p. 141) stated that the impulsiveness parameter "can vary across species, individuals, and situations". It may be subject to training and education, for a decrease during life is often observed. Children are usually more impulsive than adults. But also adults lack some form of self-control in many situations (Rachlin, 2000).

Third, the sensitivity to delay may lead to a *change in preferences* if the moment of reinforcement comes closer in time. As explicated by Rachlin (2000, pp. 30-41), this distinguishes hyperbolic discounting from exponential discounting. If the values are discounted exponentially, the following form is assumed: $v_i = \delta^{d_i} \cdot a_i$,

$i \in \{1,2\}$. Given that $v_1 = \delta^{d_1} \cdot a_1 > \delta^{d_2} \cdot a_2 = v_2$, this relationship holds for any constant $x \in \mathbb{R}$ added to d_1 and d_2: $\delta^{d_1+x} \cdot a_1 > \delta^{d_2+x} \cdot a_2$. In contrast, the preference relation can change if equation (2.7) is used. For example, it holds that $\frac{10}{1+\delta \cdot 0} > \frac{20}{1+\delta \cdot 2}$ in case of $\delta = 1.0$. If adding $x = 100$ to each time delay d_1 and d_2, the opposite relation arises: $\frac{10}{1+\delta \cdot 100} \approx 0.1 < 0.2 \approx \frac{20}{1+\delta \cdot 102}$. From a large distance in time, the differences in amount are correctly perceived. But the perception of the amounts are distorted when being temporarily close to a reinforcement.

According to the Ainsle-Rachlin theory, the commonly observed failure of self-control is explained by hyperbolic discounting (see e.g. Mazur, 2001, ch. 14.3, or Rachlin, 2000, ch. 2). A failure of self-control is observed if an individual is aware of the high benefit of one alternative but still chooses another less beneficial alternative because of its immediate reinforcement. There is numerous experimental and everyday evidence for a predominant lack of self-control in humans (Herrnstein, 1990b; Mazur, 2001). For instance, many people know the benefit of having cereals instead of cake for breakfast but still choose the latter. The benefit of cake is immediate, but the value of eating healthy is delayed. People who show this kind of behaviour may be characterised as lacking self-control. Their behaviour is explained by equation (2.7) and a high value of δ. Another example, which is explained by hyperbolic discounting, is the commonly observed regret of addicts who recognise the predominance of long-term costs of addictive behaviour but cannot resist in the moment of choice.

Hyperbolic discounting also explains the affinity towards a variable-interval reinforcement schedule in contrast to a fixed-interval schedule (e.g. Bacotti, 1977). Both schedules may reinforce an action after the same amount of time *on average*. But, while the time is constant in case of the fixed-interval schedule, it varies in case of the variable-interval schedule. This means that the reinforcements on the variable-interval schedule usually take place later or earlier than the reinforcements on the fixed-interval schedule. According to hyperbolic discounting, the immediate reinforcements have a much greater value than the late reinforcements. This makes the variable-interval schedule more valuable than the fixed-interval schedule. Consequently, variable-interval reinforcement schedules are more effective in teaching desired behaviour than fixed-interval schedules (Mazur, 2001). In a similar manner, compulsive gambling is explained.

2.2.3 The law of response strength

While equation (2.1) requires the choice between two alternatives, also the repeated selection of a single alternative can be examined with respect to the matching law. This situation is referred to as the "take or leave it"-choice, which means that an actor obtains the opportunity to take a particular action or to do nothing. In this case, Herrnstein (1970) suggested the following equation:

$$k_1 = (k_1 + k_0) \cdot \frac{s_1}{s_1 + s_0}. \tag{2.9}$$

In equation (2.9), $k := k_1 + k_0$ denotes the total number of occasions in which a single alternative is available for choice. Similar to the previous sections, k_1 is the actual absolute frequency of choice, and s_1 gives the number of choices that were reinforced. Furthermore, s_0 equals the number of occasions that were reinforced without choosing the alternative (thus: $s_0 \leq k_0$). It is hypothesized that an actor becomes distracted by various aspects of the environment and that these distractions are sometimes experienced as reinforcement. For example, after a longer period of pecking, pigeons may appreciate some rest, or they get distracted by an itch and spend the next couple of seconds scratching.

Equation (2.9) adds two aspects to the study of the matching law. First, the number of choice opportunities in which no alternative was chosen (k_0), and, second, the number of reinforcements that are not connected to any alternative (s_0). Both factors are required if only one alternative is considered. In the case of two alternatives, these variables can be included as well. This leads to the following version of the matching law, which has been called the **law of response strength** (Herrnstein, 1970; de Villiers and Herrnstein, 1976):

$$k_1 = k \cdot \frac{s_1}{s_0 + s_1 + s_2}, \tag{2.10}$$

with $k := k_0 + k_1 + k_2$. In comparison to the strict matching law, the interpretation of k_0 and s_0 diverges from the interpretation of k_1 or k_2 and s_1 or s_2. Another difference is the emphasis on the absolute frequency k_1 instead of the relative frequency $\frac{k_1}{k}$. But formally, the law of response strength is a mere extension of equation (2.1) to a larger set of alternatives.

2.2.4 Social matching

In general, the interpretation of the matching law presumes that the variables describe the behaviour of a single actor. Some authors (Gray and von Broembsen, 1976; Gray et al., 1982) argued that, due to social comparison processes, the matching law also holds on the social level. For example, two actors, who are denoted by the letters x and y, may choose repeatedly among several alternatives. Similar to the definition of the strict matching law, the frequencies of choosing a particular alternative are denoted by k_x and k_y for each actor, and the frequencies of reinforcement are given by s_x and s_y. The authors stated that

$$\frac{k_x}{k_x + k_y} = \frac{s_x}{s_x + s_y}. \tag{2.11}$$

In contrast to the strict matching law, the sums of the denominators run over the set of actors and not over the set of alternatives.

Gray and von Broembsen (1976) applied this formula of **social matching** to different data and demonstrated its applicability to various social settings, such as communication and power structures. Gray et al. (1982) analysed this equation theoretically in the context of social exchange. The authors argued that differences in exchange outcomes are explained by matching processes on the social level.

As mentioned by Gray and von Broembsen (1976), equation (2.11) is derived from the assumptions of social comparison processes and perfect information about the other actor's choices and reinforcements. In case that each of the two actors observes the choices and reinforcements of the other actor, social comparison might mean that both actors adjust their choice distributions until their success rates match:

$$\frac{s_x}{k_x} = \frac{s_y}{k_y}. \tag{2.12}$$

Equation (2.12) is the "social counterpart" of condition (2.2), which implies that equation (2.11) results from equation (2.12):

$$\frac{k_x}{k_x + k_y} = \frac{k_x}{s_x \cdot \frac{k_x}{s_x} + s_y \cdot \frac{k_x}{s_x}} = \frac{s_x}{s_x + s_y}.$$

Therefore, the social matching law derives from social comparison processes.

2.3 Empirical evidence

All of the previously mentioned versions of the matching law were analysed empirically. In experimental settings, the frequencies of choice (k_1, k_2) and the frequencies of reinforcement (s_1, s_2) can be measured accurately. But the free parameters α, β, k_0, and s_0 are usually not known. Instead of *testing* the generalised matching law or the law of response strength, most authors estimated the free parameters by fitting equation (2.3) or (2.10) to observed data.

De Villiers and Herrnstein (1976) summarised various experiments with pigeons, rats, monkeys, and humans. The law of response strength (equation (2.10)) accounted for a high percentage of variance in the data if the parameters k_0 and s_0 are fitted properly. In a more recent review, McDowell (2013a, p. 9) disagreed and stated that, in most cases, the law of response strength performs badly. Only if additional parameters were added to equation (2.10) (similar to equation (2.3)), the law of response strength fit the observations.

A similar conclusion was usually drawn from the large number of studies that applied equations (2.1) and (2.3) to experimental data. Reviewers of these studies (e.g. Baum, 1979; Pierce and Epling, 1983; McDowell, 2005; Reed and Kaplan, 2011; Poling et al., 2011) recapped that the generalised matching law, but not the strict version, accurately described the choice behaviour under a wide variety of conditions. This finding held for animals as well as humans. For instance, in an experiment of Sunahara and Pierce (1982), human participants exchanged monetary points with two different partners by pressing buttons on an interaction panel. The results confirmed the generalised matching law. Due to inequalities in the reinforcements, the bias parameter β was different from one.

An experiment that more adequately resembled real situations was conducted by Conger and Killeen (1974). The authors observed the behaviour of members of a discussion group. The matching law correctly described the relationship between the rate of speaking to one of the discussion partners and the rate of positive responses from this partner. However, the authors fit the matching law to the pooled data of all (five) subjects (Conger and Killeen, 1974, p. 412). The results might be misleading because matching on a population level does not necessary imply matching on the individual level (see also Caron, 2013b,a).

Further experiments and field studies with humans were summarised by Hamblin (1977) or Pierce and Epling (1983). The situations included verbal interactions, gambling, group discussions, and even the publication of encyclopedia articles. McDowell (1988, pp. 103-104) reported a series of studies that confirmed the matching law in natural human environments (outside of the laboratory). In these studies, undesired behaviour, such as self-injurious scratching or disruptive behaviour in school, was either negatively reinforced or confronted with positive reinforcement of alternative behaviour. The decrease of undesired behaviour in dependence of the relative rate of reinforcement corresponded to the predictions of the matching law. Additionally, desired behaviour, such as good performance in school, diminished because of random background reinforcement. The rate of change was in accordance with the law of response strength.

A list of more recent studies is found in McDowell (2013a, p. 2). For example, Borrero et al. (2007) applied the generalised matching law to subjects who participated in a discussion about juvenile delinquency. In contrast to the earlier study of Conger and Killeen (1974), Borrero et al. (2007) used individual data to fit the matching law equation and obtained partly affirmative results.

In another study, Vollmer and Bourret (2000) analysed the success of two- and three-point shots during college basketball games. In case of experienced players, the relative frequencies of attempted three-point shots matched the relative frequencies of scored three-points shots. This result was confirmed in further studies (Alferink et al., 2009). Moreover, when comparing the players of different divisions, the levels of undermatching varied. Players of highly ranked teams showed less undermatching then players of low ranks. This finding supports the conjecture that undermatching results from experimenting. Lowly ranked players had maybe insufficient experience and, therefore, tried three-point shots even though they were less successful than two-point shots.

A comparable result was obtained in an experiment in which participants played a simulated rock-paper-scissors game (Kangas et al., 2009). The computer opponent chose its answer by a schedule of fixed probabilities. In this case, the only outcome that is in line with the strict matching law is the exclusive choice of the alternative with the highest probability of success (e.g. Herrnstein, 1982, p. 78). In a first treatment, the subjects had no information about the probabilities. Their choices deviated from the strict matching law, as did the choices of the less experienced basketball players. In another treatment, information about the reinforcement probabilities were revealed. Subsequently, the choices of almost all subjects were in agreement with the strict matching law. Further analyses indicated a development from undermatching to strict matching during the trials of the first treatment. These findings suggest that undermatching is the consequence of insufficient experience with the underlying reinforcement procedures.

2.4 Conclusion

One might say that the matching law is a widely studied and often observed empirical phenomenon. But the term "matching law" was not used consistently in the past. Initially, this law was formulated in light of a small number of experiments. With the analysis of a broader range of situations, systematic deviations from this first version were discovered, and classes of equations with several free parameters were introduced. The success of the "matching law" results from the connection of this term to these classes of equations and from the fact that almost every behaviour can be described by them.

Chapter 3

Integration into consumer theory

Although the generalised matching law describes behaviour more accurately than the strict version, it has the disadvantage of containing a set of free parameters. In most studies, these parameters were arbitrarily chosen in order to fit the model to the data. This approach reduces the falsifiability of the matching law and its applicability as micro-level assumption in social situations.

Another problem of the generalised matching law is its sensitivity to situational constraints. Caron (2015) demonstrated that, even if there is no effect from reinforcements (s_1, s_2) to choices (k_1, k_2), fitting the generalised matching law to simulated data produces spurious correlations. These correlations are due to the constraints $s_1 \leq k_1$ and $s_2 \leq k_2$. The level of explained variance was, with $41-49\%$, lower than in behavioural experiments. Nevertheless, the finding stresses the necessity of a “more comprehensive and exhaustive description of environmental constraints” (Caron, 2015, p. 232).

In this chapter, a formalisation of the matching law that is able to account for environmental constraints is advanced. Similar to the generalised matching law, a greater empirical validity is achieved by taking some causes of bias and undermatching into account. But instead of introducing free parameters, the frequencies of reinforcement (s_1 and s_2) are substituted by more adequate measures of an actor's preferences and situational constraints. This procedure is in line with the interpretations of the matching law by Herrnstein (1997), Rachlin et al. (1980), and Gray and Tallman (1984).

More specifically, the integration of the matching law into economic consumer theory is pursued. In some form, many of the following results are found in previous work (e.g. Rachlin et al., 1980; Herrnstein, 1997). The main contribution of this chapter is the unification of different lines of work and the generalisation of earlier results. One goal is the presentation of an empirically testable matching law. Additionally, it is aimed for the consideration of the matching law as an alternative to standard solutions of economic theory.

A connection between the matching law and economic theory has already been worked out by Howard Rachlin and colleagues (Rachlin et al., 1976, 1980, 1981). The authors attempted to establish general aspects of microeconomic theory in behavioural psychology. They pointed to the advantages of modelling the experimental situation by utility and budget functions and compared the predictions of utility maximisation to experimental observations. While the main focus of these papers and some of its critics (e.g. Baum and Nevin, 1981; Herrnstein, 1981) was the attempt to introduce the maximisation assumption in behavioural psychology, this chapter deals with another point that was made by the authors. In Rachlin et al. (1980), it was argued that the generalised matching law can be applied to the microeconomic framework. It was also indicated that free parameters must be derived from a theory of value (Rachlin et al., 1981, p. 409).

Following Rachlin et al. (1976), a situation of repeated choice, such as a behavioural experiment or a real-world setting, is modelled by the consumer problem of microeconomic theory. In order to apply the matching law, restricting assumptions about the consumer problem have to be made. The resulting class of situations is called the *problem of distributed choice* (section 3.1).

While most of the preceding work on the matching law was limited to the choice among two alternatives, the problem of distributed choice allows any finite number of alternatives (as suggested by Herrnstein, 1970, 1971, 1974). Furthermore, a previously disregarded connection to economic utility theory, i.e. the theory of additive conjoint measurement, is explicated. This enables the theoretical derivation of the model parameters from assumptions about the actor's preferences and a formal description of the situation.

In section 3.2, the matching law is introduced as solution to the problem of distributed choice. Similar approaches can be found in past theoretical research:

first, a general class of *problems* is defined, and, subsequently, a *solution* to these problems is presented. For example, the Nash equilibrium was suggested as a solution to any non-cooperative game with two or more players (Nash, 1951). Comparably, several theoretical studies in behavioural psychology focused on the formal modelling of an experimental situation and compared different theoretical predictions to each other and to observed behaviour (Houston and McNamara, 1981; Staddon, 2001; Belinsky et al., 2004, 2005).

In contrast to earlier work on the generalised matching law, the new formalisation requires the specification of the situation ex ante and allows to test the matching law instead of fitting it to empirical data. Hence, a more accurate test of the matching law is made possible.

Moreover, the next chapter emphasises the usage of the matching law as an alternative to standard solutions of economic theory, which are usually built on utility maximisation. Despite the large amount of experimental research by psychologists, the matching law has been largely ignored by economists and sociologists. By restricting the consumer problem to the problem of distributed choice, the matching law can be applied as micro-level assumption of individual behaviour. Similar to the usage of other micro-level assumptions, this provides an opportunity to derive social phenomena from the matching law.

3.1 The problem of distributed choice

Since the matching law cannot be applied to any instance of the consumer problem, restricting assumptions about the *situation* and the *preferences* of an actor are required. Subsequently, it is possible to substitute the frequencies of reinforcement (s_1 and s_2) by more adequate measures of reinforcement. In the following, sufficient conditions for the application of the matching law are given.

3.1.1 Assumptions about the situation

For any $m \in \mathbb{N}$, a set $E = \{e_1, e_2, \ldots, e_m\}$ is regarded as set of choice alternatives. An actor is assumed to repeatedly choose one of the alternatives of E. In order to be in line with previous research, it is distinguished between *behaviour* and *outcomes.* The **behaviour** of an actor is modelled by the set of choice distributions

$$\mathcal{P} := \left\{ \boldsymbol{p} = (p_{e_1}, p_{e_2}, \dots, p_{e_m}) \in [0,1]^E \mid \sum_{j \in E} p_j = 1 \right\}.$$

Each element of $\mathcal{P}$ denotes a frequency distribution over the alternatives E.

The set of **outcomes** is restricted to $\mathcal{X} := [0, \infty)^E$. If $\boldsymbol{x} = (x_{e_1}, x_{e_2}, \dots, x_{e_m}) \in \mathcal{X}$, the elements $x_{e_1}, x_{e_2}, \dots, x_{e_m}$ should be seen as the *relative amounts of reinforcement* that are obtained after choosing the corresponding alternatives. An outcome depends on the behaviour and the situation. This relationship is commonly specified by a **budget function** $g : \mathcal{P} \to \mathcal{X}$.

The budget function of a behavioural experiment depends on the experimental procedure. For example, in case of two alternatives ($E = \{1, 2\}$) and a concurrent fixed-ratio schedule, a reinforcement is released with a fixed probability $b_e \in (0, 1)$ after choosing an alternative $e \in E$. If there are no further differences in reinforcement, the budget function is defined by

$$g(\boldsymbol{p}) := (b_1 \cdot p_1, b_2 \cdot p_2), \text{ for all } \boldsymbol{p} = (p_1, p_2) \in \mathcal{P}. \tag{3.1}$$

If the first alternative is chosen 30% of the time ($p_1 = 0.3$) and this alternative is reinforced with probability $b_1 = 0.5$, the relative amount of reinforcement is $x_1 = 0.15$. During 100 rounds of the experiment, the subject obtains reinforcements in approximately 15 of the 30 rounds in which the first alternative was chosen.

According to Rachlin et al. (1980), the budget function of a general behavioural experiment with m alternatives ($E = \{1, 2, \dots, m\}$) is given by

$$g(\boldsymbol{p}) := (b_1 \cdot p_1^{r_1}, b_2 \cdot p_2^{r_2}, \dots, b_m \cdot p_m^{r_m}), \text{ for all } \boldsymbol{p} = (p_1, p_2, \dots, p_m) \in \mathcal{P}. \tag{3.2}$$

The parameters $(r_1, \dots, r_m) \in (0, 1]^m$ and $(b_1, \dots, b_m) \in [0, \infty)^m$ specify the reinforcement schedules. If $r_i = 1$, the alternative $i \in E$ is reinforced by a ratio schedule. An interval schedule is characterised by $0 < r_i < 1$ (Rachlin et al., 1980, p. 362). The second set of parameters $(b_1, b_2, \dots, b_m)$ allows to include differences in probabilities or amounts of reinforcement. Following the arguments of Baum (1974), this should account for some of the systematic deviations from the strict matching law that were captured by the free parameters α and β of equation (2.3).

In the budget function of equation (3.2), the outcome $x_i = b_i \cdot p_i^{r_i}$ of an alternative $i \in E$ depends on the frequency p_i of this alternative. Generally, also the frequency of another alternative can influence this outcome. This requires at least three alternatives and is usually avoided in laboratory experiments. In real-world situations, complex correlations are possible. For example, a person may choose between three routes to work each morning: $E = \{A, B, C\}$. The selection of a route is regarded as successful if the person is not caught up in a traffic jam. It pays off not to take one route exclusively because of seasonal differences, such as school holidays or weather conditions. In other words, the outcome x_A of choosing route A depends on its frequency p_A. However, x_A may also depend on the frequency p_B of choosing route B. For instance, if other persons choose between route A and C every morning, the first person should avoid route C most of the time such that the others keep clear of route A. Hence, the outcome x_A is improved by choosing route B more often than route C.

While it is difficult to specify the budget function of real-world situations, it is usually straightforward to do this for laboratory experiments. A greater challenge is the derivation of a subject's preferences in regard to the outcomes. As demonstrated by Rachlin et al. (1981), economic utility theory can be used to derive preferences from behavioural experiments. But additional assumptions about the preferences of an actor have to be made if the matching law is applied.

3.1.2 Assumptions about the preferences of an actor

One fundamental requirement is that the actor's preferences over the set $\mathcal{X}$ can be described by a **non-negative** and **additive utility function**. More specifically, let $\succsim$ be a binary relation on $\mathcal{X}$ that describes the actor's preferences. It is required that there exist functions $u_{e_1}, u_{e_2} \ldots, u_{e_m} : [0, \infty) \to [0, \infty)$ such that for every $\boldsymbol{x} = (x_{e_1}, x_{e_2}, \ldots, x_{e_m}) \in \mathcal{X}, \boldsymbol{y} = (y_{e_1}, y_{e_2}, \ldots, y_{e_m}) \in \mathcal{X}$:

$$\boldsymbol{x} \succsim \boldsymbol{y} \Leftrightarrow \sum_{j \in E} u_j\left(x_j\right) \geq \sum_{j \in E} u_j\left(y_j\right). \tag{3.3}$$

Since the functions u_i depend only on x_i instead of the whole vector $(x_{e_1}, \ldots, x_{e_m})$, a connection to the theory of *additive conjoint measurement* is made.

The theory of additive conjoint measurement characterises preferences that allow an additive decomposition of the utility function (Fishburn, 1970; Krantz et al., 1971). Necessary conditions for the existence of an additive utility function are *weak ordering* (completeness and transitivity), *independence* among the alternatives, and the *Archimedean axiom* (Krantz et al., 1971). If, additionally, at least three alternatives are *essential* (they actively affect the preference relation, Krantz et al., 1971, p. 256) and if *restricted solvability* holds (Krantz et al., 1971, p. 301), these conditions are sufficient for the existence of an additive utility function. The last axiom may be replaced by other assumptions (Jaffray, 1974; Nakamura, 2002). But if $u_{e_1}, u_{e_2} \ldots, u_{e_m}$ are continuous on $[0, \infty)$, restricted solvability holds.

The following definition summarises the assumptions. It describes a class of situations to which the matching law can be applied. In line with previous research on this subject, this situation is labelled the *problem of distributed choice* (Herrnstein and Prelec, 1991).

Definition 3.4. *Let* $m \in \mathbb{N}$*,* $E = \{e_1, e_2, \ldots, e_m\}$*,* $\mathcal{X} := [0, \infty)^E$*, and* $\mathcal{P} := \Big\{(p_{e_1}, p_{e_2}, \ldots, p_{e_m}) \in [0,1]^E \mid \sum_{j \in E} p_j = 1\Big\}$*. A* ***problem of distributed choice*** *is given by a triple* (E, g, u) *if*

- $g : \mathcal{P} \to \mathcal{X}$ *is a* ***budget function*** *with* $p_j = 0 \Rightarrow x_j = 0$*,* $\forall j \in E$*, and if*
- $u : \mathcal{X} \to [0, \infty)$ *is a* ***utility function*** *such that there exist functions* $u_{e_1}, u_{e_2} \ldots, u_{e_m} : [0, \infty) \to [0, \infty)$ *with* $u_j(0) = 0$*,* $\forall j \in E$*, and*

$$u(x_{e_1}, \ldots, x_{e_m}) = \sum_{j \in E} u_j(x_j) \text{ for every } (x_j)_{j \in E} \in \mathcal{X}.$$

It is required that $p_e = 0 \Rightarrow x_e = 0$ and $x_e = 0 \Rightarrow u_e(x_e) = 0$ for every $e \in E$. On the one hand, this is in accordance with the general understanding that, if an alternative $e \in E$ is not chosen, then neither an outcome nor any utility is linked to this alternative.

On the other hand, this assumption enables the application of the matching law because it implies that the measurement of utility must be done on ratio scales. As stated above, certain axioms are required for the utility function to be additively decomposable in non-negative functions $u_{e_1}, \ldots, u_{e_m}$. The same axioms guarantee that these functions are unique up to positive linear transformations (Krantz et al., 1971, p. 302, theorem 13). Since $u_e(0) = 0$ for all $e \in E$, the intercept of any linear transformation of $u_{e_1}, \ldots, u_{e_m}$ must be zero, and the functions are unique up to scale. This is a property of the ratio scale.

Definition 3.4 distinguishes between budget and utility function. It is commonly accepted that the utility function is attached to an actor and does not change with the environment. The budget function, on the contrary, is defined by situational properties, such as the mechanisms of reinforcement. For the purpose of an easier presentation, the two functions are concatenated to $v := u \circ g$ in the following definition. This concatenation also stresses that, instead of the outcome $\boldsymbol{x} \in \mathcal{X}$, which is usually the focus of economic analysis, the behaviour $\boldsymbol{p} \in \mathcal{P}$ is the variable of interest.

Definition 3.5. *Let (E, g, u) be a problem of distributed choice. The function $v : \mathcal{P} \to [0, \infty)$ with $v := u \circ g$ is called **value function**. A problem of distributed choice is also denoted by the pair (E, v). Furthermore, let $\pi_j : [0, \infty)^E \to [0, \infty)$, $j \in E$, be the projection function that maps $(x_{e_1}, x_{e_2}, \ldots, x_{e_m})$ to x_j. The function $v_j : \mathcal{P} \to [0, \infty)$ with $v_j := u_j \circ \pi_j \circ g$ is the **component value function** of $j \in E$. If $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ and $p_j > 0$, $j \in E$, then*

$$\overline{v}_j(\boldsymbol{p}) := \frac{v_j(\boldsymbol{p})}{p_j}$$

*is called the **average value** of j.*

The component value functions $\{v_j\}_{j \in E}$ assign non-negative real numbers to every element of $\mathcal{P}$. Because of the additive structure of the utility function, it holds that

$$v(\boldsymbol{p}) = \sum_{j \in E} v_j(\boldsymbol{p}) \text{ for every } \boldsymbol{p} \in \mathcal{P}.$$

Consequently, for a given $\boldsymbol{p} \in \mathcal{P}$, $v_j(\boldsymbol{p})$ returns the share of the total value $v(\boldsymbol{p})$ that is associated with alternative $j \in E$.

Two examples of a problem of distributed choice with $E = \{1, 2\}$ are shown in table 3.1. In the first example, the linear utility function $u(\boldsymbol{x}) = 8x_1 + 6x_2$, for all $\boldsymbol{x} = (x_1, x_2) \in \mathcal{X}$, is considered. This function can be additively decomposed into

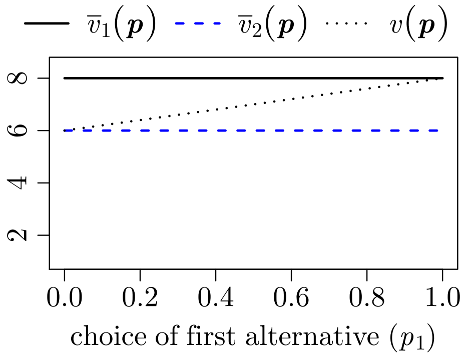
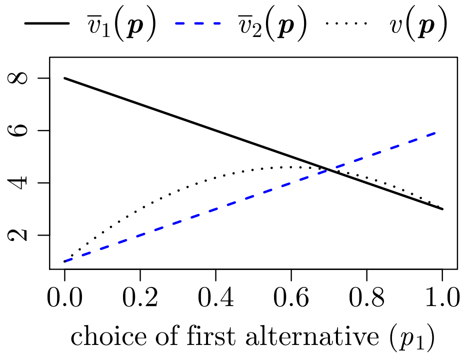
example 1	**example 2**
$v(\boldsymbol{p}) = 8p_1 + 6p_2$	$v(\boldsymbol{p}) = 8p_1 - 5p_1^2 + 6p_2 - 5p_2^2$
$\overline{v}_1(\boldsymbol{p}) = 8$, $\overline{v}_2(\boldsymbol{p}) = 6$	$\overline{v}_1(\boldsymbol{p}) = 8 - 5p_1$, $\overline{v}_2(\boldsymbol{p}) = 6 - 5p_2$
	

Table 3.1: Examples of a problem of distributed choice with $E = \{1, 2\}$

$u_1(x_1) = 8x_1$ and $u_2(x_2) = 6x_2$. If a reinforcement follows a choice with certainty, the budget function is given by $g(\boldsymbol{p}) = \boldsymbol{p}$. This implies the value functions and average values that are listed in the left-sided column.

In the second example of table 3.1, the average values $\overline{v}_1$ and $\overline{v}_2$ decrease with the relative frequencies of choosing the respective alternatives. This may, for instance, result from a utility function that accounts for the effects of satiation or from a budget function which outcome depends on the frequency of previous decisions.

3.2 The matching law solution

In order to introduce the matching law as a solution to the problem of distributed choice, the experiment of Herrnstein (1961) (see section 2.1) and the strict matching law are reconsidered. The experiment consisted of the repeated choice between two alternatives ($E = \{1, 2\}$). If an actor distributes $n \in \mathbb{N}$ decisions over E, the observed behaviour is any pair (p_1, p_2) with

$$(p_1, p_2) \in \left\{(0,1), \left(\tfrac{1}{n}, \tfrac{n-1}{n}\right), \left(\tfrac{2}{n}, \tfrac{n-2}{n}\right), \ldots, \left(\tfrac{n-1}{n}, \tfrac{1}{n}\right), (1,0)\right\} \subset \mathcal{P}.$$

In respect to the notation of section 2.1, this means that

$$p_1 = \tfrac{k_1}{k_1+k_2} \text{ and } p_2 = \tfrac{k_2}{k_1+k_2},$$

with $n = k_1 + k_2$ and k_i denoting the absolute frequency of choosing alternative $i \in E$. Since Herrnstein (1961) used a set of concurrent variable-interval schedules, the budget function is given by

$$g(\boldsymbol{p}) := (b_1 \cdot p_1^{r_1}, b_2 \cdot p_2^{r_2}), \text{ for all } \boldsymbol{p} = (p_1, p_2) \in \mathcal{P},$$

with $r_1, r_2 \in (0,1)$ and $b_1, b_2 \in (0, \infty)$ specifying the schedule. It follows that the absolute frequencies of reinforcement $s_1, s_2 \in \mathbb{N}$ approximate $n \cdot b_1 \cdot p_1^{r_1}$ and $n \cdot b_2 \cdot p_2^{r_2}$, respectively, if n is large.

Because there were no differences in outcomes (same amounts of the same kind of grain), the preferences are represented by $u(x_1, x_2) = x_1 + x_2$, which is clearly additive with $u_1(x_1) = x_1$ and $u_2(x_2) = x_2$. Accordingly, the component value functions are given by $v_1(\boldsymbol{p}) = b_1 \cdot p_1^{r_1}$ and $v_2(\boldsymbol{p}) = b_2 \cdot p_2^{r_2}$, for all $\boldsymbol{p} = (p_1, p_2) \in \mathcal{P}$. With regard to equation (2.1), the strict matching law holds if

$$p_1 = \frac{k_1}{k_1 + k_2} = \frac{s_1}{s_1 + s_2} \approx \frac{n \cdot b_1 \cdot p_1^{r_1}}{n \cdot b_1 \cdot p_1^{r_1} + n \cdot b_2 \cdot p_2^{r_2}} = \frac{v_1(\boldsymbol{p})}{v_1(\boldsymbol{p}) + v_2(\boldsymbol{p})}.$$

This observation motivates the following general definition of the matching law.

Definition 3.6. *Given a problem of distributed choice (E, g, u) and the value functions $v, \{v_j\}_{j \in E}$ as specified by definition 3.5, the* ***matching law*** *holds for a given $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ if $v(\boldsymbol{p}) > 0$ and*

$$p_i = \frac{v_i(\boldsymbol{p})}{\sum_{j \in E} v_j(\boldsymbol{p})}, \textit{ for all } i \in E. \tag{3.7}$$

The set of all $\boldsymbol{p} \in \mathcal{P}$ that satisfy condition (3.7) is called ***matching law solution*** *and denoted by $M(E, g, u)$ or $M(E, v)$.*

For the examples of table 3.1, the matching law solutions are $\{(0,1), (1,0)\}$ and $\{(0,1), (1,0), (0.7, 0.3)\}$, respectively. It can be easily checked that condition (3.7) holds for the elements of both sets. In order to make sure that a matching law solution is complete, the following characterisations should be used.

First, a matching law solution is never empty. Let $e \in E$ be an alternative with $p_e = 1$, then $p_j = 0$, for all $j \neq e$. Because it is required by definition 3.4 that $v_j(\boldsymbol{p}) = 0$ if $p_j = 0$, the conditions of equation (3.7) reduce to $1 = p_e = \frac{v_e(\boldsymbol{p})}{v_e(\boldsymbol{p})} = 1$ and $0 = p_j = \frac{0}{v_e(\boldsymbol{p})}$, for all $j \neq e$.

Observation 3.8. *Given a problem of distributed choice (E, v):*

$$\{(p_j)_{j\in E} \in \mathcal{P} : p_j \in \{0,1\}, \textit{ for all } j \in E\} \subseteq M(E, v).$$

A choice distribution of the set $\{(p_j)_{j\in E} \in \mathcal{P} : p_j \in \{0,1\}$, for all $j \in E\}$ consists of zeros and a single one at an index $e \in E$. These distributions can be interpreted as "choosing only alternative e".

Second, given a solution $(p_j)_{j\in E} \in M(E, v)$ and any $e \in E$ with $p_e > 0$, it must hold that $\sum_{j\in E} v_j(\boldsymbol{p}) = \frac{v_e(\boldsymbol{p})}{p_e}$. This is expressed by the next observation.

Observation 3.9. *For any element $\boldsymbol{p} = (p_j)_{i\in E} \in M(E, v)$ and $e \in E$, it holds that*

$$p_e > 0 \Rightarrow \overline{v}_e(\boldsymbol{p}) = v(\boldsymbol{p}).$$

Since the total value $v(\boldsymbol{p}) = \sum_{j\in E} v_j(\boldsymbol{p})$ is constant for a given choice distribution $\boldsymbol{p} = (p_j)_{j\in E} \in \mathcal{P}$, the average values $\overline{v}_j(\boldsymbol{p})$ of all $j \in E$ with $p_j > 0$ are equal. This is a sufficient and necessary condition of the matching law:

Proposition 3.10. *Given a problem of distributed choice (E, v) and any $\boldsymbol{p} = (p_j)_{i\in E} \in \mathcal{P}$: $\boldsymbol{p} \in M(E, v)$ if and only if, for all $i, j \in E$ with $p_i, p_j > 0$,*

$$\overline{v}_i(\boldsymbol{p}) = \overline{v}_j(\boldsymbol{p}). \tag{3.11}$$

Proof. (i) $\Rightarrow$: Let $i, j \in E$ with $p_i, p_j > 0$. Because of observation 3.9, it follows that $\overline{v}_i(\boldsymbol{p}) = v(\boldsymbol{p}) = \overline{v}_j(\boldsymbol{p})$. (ii) $\Leftarrow$: Let $i \in E$. a) If $p_i = 0$, then $p_i = 0 = \frac{v_i(\boldsymbol{p})}{\sum_{j\in E} v_j(\boldsymbol{p})}$. b) If $p_i = 1$, then $p_j = 0$ for all $j \in E$ with $j \neq i$, and $p_i = 1 = \frac{v_i(\boldsymbol{p})}{v_i(\boldsymbol{p})} = \frac{v_i(\boldsymbol{p})}{\sum_{j\in E} v_j(\boldsymbol{p})}$. c) If $0 < p_i < 1$, it follows from the assumption that $\overline{v}_i(\boldsymbol{p}) = \overline{v}_j(\boldsymbol{p})$ for all $j \in E$ with $p_j > 0$. Furthermore: $\overline{v}_i(\boldsymbol{p}) > 0$ because $\sum_{j\in E} v_j(\boldsymbol{p}) > 0$ is required by definition 3.4. Hence: $p_i = \frac{\overline{v}_i(\boldsymbol{p})\cdot p_i}{\overline{v}_i(\boldsymbol{p})} = \frac{\overline{v}_i(\boldsymbol{p})\cdot p_i}{\overline{v}_i(\boldsymbol{p})\cdot\sum_{j\in E} p_j} = \frac{\overline{v}_i(\boldsymbol{p})\cdot p_i}{\sum_{j\in E} \overline{v}_j(\boldsymbol{p})\cdot p_j} = \frac{v_i(\boldsymbol{p})}{\sum_{j\in E} v_j(\boldsymbol{p})}$. □

Proposition 3.10 implies that a frequency distribution $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ is an element of the matching law solution only if the average value functions $\overline{v}_i$ and $\overline{v}_j$ of any two elements $i, j \in E$ with $p_i > 0$ and $p_j > 0$ intersect at the point $\boldsymbol{p}$. Therefore, condition (3.11) simplifies the application of the matching law to a problem of distributed choice. In table 3.1, the right-sided plot shows that the curves $\overline{v}_1$ and $\overline{v}_2$ intersect at $p_1 = 0.7$, which is an element of the matching law solution. Furthermore, since this is the only intersection, there are no further elements of $\mathcal{P}$ for which the matching law holds (apart from the elements specified by observation 3.8).

The characterisation of the matching law by condition (3.11) is widely known in behavioural psychology (e.g. Herrnstein, 1997). But the presentation of this equivalence relationship has mostly been incomplete and restricted to a small set of situations. Proposition 3.10 demonstrates that the equivalence of condition (3.11) and condition (3.7) holds in any situations that can be specified as a problem of distributed choice. It also emphasises that the matching law refers to alternatives that are chosen with strictly positive frequencies. Alternatives with zero frequency can be disregarded. Especially, the exclusive choice of a single alternative is always in line with the matching law (observation 3.8). This observation enlarges the set of choice distributions that are explained by the matching law.

Proposition 3.10 also clarifies that a researcher can concentrate on any subset $F \subset E$ of readily observable alternatives without compromising the credibility of the results. If the matching law holds for all alternatives E, then it must also hold for the observed ones F. Accordingly, if the matching law cannot be found in F, it does not hold for any superset of F.

Another benefit of the formulation of proposition 3.10 is the derivation of the following point: *the prediction of the matching law is independent of the particular representation of the preferences.* As stated in section 3.1.2, the functions $u_{e_1}, \ldots, u_{e_m}$ are unique up to scale in case that they are continuous on $[0, \infty)$. If there is another set of continuous functions $w_{e_1}, \ldots, w_{e_m}$ that additively represent the preference relation of the actor, then there exists an $a \in (0, \infty)$ with

$$w_j(x_j) = a \cdot u_j(x_j) \text{ for every } j \in E \text{ and } x_j \in [0, \infty).$$

It follows that, given a $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ and any $i, j \in E$ with $p_i, p_j > 0$,

$$\frac{u_i(\pi_i(g(\boldsymbol{p})))}{p_i} = \frac{u_j(\pi_j(g(\boldsymbol{p})))}{p_j} \Leftrightarrow \frac{w_i(\pi_i(g(\boldsymbol{p})))}{p_i} = \frac{w_j(\pi_j(g(\boldsymbol{p})))}{p_j}.$$

This proves the following proposition.

Proposition 3.12. *Let (E, g, u) be a problem of distributed choice and u be a continuous utility function that represents an actor's preferences. If there is another continuous, additive, and non-negative utility function w that represents the actor's preferences, then*

$$M(E, g, u) = M(E, g, w).$$

3.3 Conclusion

Given an adequate specification of the economic consumer problem, it is possible to apply the matching law as a solution. In this regard, the matching law becomes a property of some frequency distributions $\boldsymbol{p} \in \mathcal{P}$ and can be used to derive empirically testable hypotheses. The advantages of this approach are the existence of a clear definition of the matching law and a greatly enhanced falsifiability of the derived hypotheses. Moreover, this framework allows the matching law to be seen as an alternative to standard solutions of economic theory.

Nevertheless, it should be noted that, since the matching law requires a period of repetitions, it is best applied to situations of routine behaviour. This point was made by Tallman and Gray (1990, p. 422) who argued that "[b]ehaviorists have provided a viable framework for explaining choices that are made in day-to-day routine situations, whereas the subjectivists [e.g. rational choice theorists] explain choices under novel, or a least nonroutine, conditions."

Chapter 4

Optimal and suboptimal matching

With its incorporation into microeconomic theory, the matching law can be compared to optimal behaviour. In line with previous research, optimal behaviour is shown to imply the matching law under certain conditions (section 4.1). For example, given an experiment with concurrent ratio or interval schedules, the matching law holds for any outcome that conforms to the optimal point of a CES utility function (section 4.2).

However, an often noted property of the matching law is that the behaviour is not necessarily optimal (Rachlin et al., 1980; Vaughan and Herrnstein, 1987; Herrnstein, 1990a,b; Herrnstein and Prelec, 1991). Section 4.3 presents several examples in which the matching law and optimal behaviour diverge. A comparison of the matching law and the Nash equilibrium concludes this chapter.

4.1 Optimal matching

Given a problem of distributed choice (E, v), an *optimisation of the overall value of reinforcement* means that an actor selects a distribution $\boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}$ that maximises the total value $v(\boldsymbol{p}) = \sum_{j \in E} v_j(\boldsymbol{p})$. It was argued by several authors (Rachlin et al., 1976; Staddon and Motheral, 1978) that maximising behaviour often leads to a prediction of the matching law. Herrnstein (1982) noted that both matching and maximising predict the same distribution $\boldsymbol{p} \in \mathcal{P}$ if the average values $\overline{v}_j(\boldsymbol{p})$ of all alternatives $j \in E$ are independent of $\boldsymbol{p}$.

When considering the choice between two alternatives 1 and 2 with constant average values of reinforcements $\overline{v}_1 = \beta_1 \in (0,\infty)$ and $\overline{v}_2 = \beta_2 \in (0,\infty)$, the overall value is optimised by any p_1 that maximises

$$\sum_{j\in E} v_j(\boldsymbol{p}) = \overline{v}_1(\boldsymbol{p}) \cdot p_1 + \overline{v}_2(\boldsymbol{p}) \cdot (1-p_1) = p_1 \cdot (\beta_1 - \beta_2) + \beta_2.$$

Hence, the optimal distributions are given by

$$\arg\max_{\boldsymbol{p}\in\mathcal{P}} v(\boldsymbol{p}) = \begin{cases} \{(1,0)\} & \text{if } \beta_1 > \beta_2, \\ \{(0,1)\} & \text{if } \beta_1 < \beta_2, \\ \mathcal{P} & \text{if } \beta_1 = \beta_2. \end{cases}$$

In comparison, the matching law solution looks slightly different:

$$M(E,v) = \begin{cases} \{(1,0),(0,1)\} & \text{if } \beta_1 \neq \beta_2, \\ \mathcal{P} & \text{if } \beta_1 = \beta_2. \end{cases}$$

If $\beta_1 = \beta_2$, the set of maximising distributions is equal to the matching law solution. In the case of $\beta_1 \neq \beta_2$, either $(1,0)$ or $(0,1)$ is optimal, but the matching law allows any of the two elements $\{(1,0),(0,1)\}$. This means that the optimal solution implies the matching law, but the converse does not hold. Even if there is a best answer (e.g. $\beta_1 > \beta_2$), the exclusive choice of the inferior alternative is consistent with the matching law.

In situations in which the average values are not constant, the optimal solution may still imply the matching law. The following proposition expresses this result for a broad class of value functions.

Proposition 4.1. *Let (E,v) be a problem of distributed choice. If there exist $\alpha \in \mathbb{R}$, $\gamma \in [0,\infty)$, and $(\beta_j)_{j\in E} \in \mathbb{R}^E$ such that, for all $\boldsymbol{p} = (p_j)_{j\in E} \in \mathcal{P}$ and every $e \in E$ with $p_e > 0$, it holds that*

$$v_e(\boldsymbol{p}) = \beta_e \cdot (p_e)^\alpha + \gamma \cdot p_e, \tag{4.2}$$

then

$$\arg\max_{\boldsymbol{p}\in\mathcal{P}} v(\boldsymbol{p}) \subseteq M(E,v).$$

Proof. The optimal solution of the situation is given by

$$\arg\max_{\boldsymbol{p}\in\mathcal{P}} v(\boldsymbol{p}) = \arg\max_{\boldsymbol{p}\in\mathcal{P}} \left(\sum_{j\in E} v_j(\boldsymbol{p})\right) = \arg\max_{\boldsymbol{p}\in\mathcal{P}} \left(\gamma + \sum_{j\in E} \beta_j \cdot p_j^\alpha\right).$$

The extrema of $\gamma+\sum_{i\in E}\beta_i\cdot p_i^\alpha$ constrained by $\boldsymbol{p} = (p_j)_{j\in E} \in \mathcal{P}$ can be found by the method of Lagrange multipliers extended to the Karush–Kuhn–Tucker conditions (see e.g. Hauser, 2012, p. 9). The conditions are compactly written as

$$\begin{aligned} (p_j)_{j\in E} &\in \mathcal{P} && (4.3)\\ 0 &= \alpha\cdot\beta_i\cdot p_i^{\alpha-1} + \lambda + \mu_i, \forall i \in E && (4.4)\\ 0 &= \mu_i \cdot p_i, \forall i\in E, && (4.5)\end{aligned}$$

with $\lambda \in \mathbb{R}$ and $(\mu_i)_{i\in E} \in \mathbb{R}^E$ being Lagrange multipliers. The method of Lagrange multipliers states that, for any point $\boldsymbol{p} = (p_j)_{j\in E} \in \mathcal{P}$ that maximises $v(\boldsymbol{p})$, there must exist $\lambda \in \mathbb{R}$ and $(\mu_i)_{i\in E} \in \mathbb{R}^E$ such that $((p_j)_{j\in E}, \lambda, (\mu_i)_{i\in E})$ satisfy conditions (4.4) and (4.5). In the following, it is shown that the set of all points $(p_j)_{j\in E} \in \mathcal{P}$ that satisfy these conditions is a subset of the matching law solution. Let $(p_j)_{j\in E} \in \mathcal{P}$ and $i \in E$. If $p_i = 0$, condition (4.5) holds for this i, and condition (4.4) becomes true by setting $\mu_i = -\lambda$. If, on the other hand, $p_i > 0$, μ_i must equal zero (because of conditon (4.5)), and condition (4.4) reduces to

$$-\lambda = \alpha\cdot\beta_i\cdot p_i^{\alpha-1}.$$

This means that conditions (4.4) and (4.5) lead to

$$\forall i,j\in E \text{ with } p_i,p_j>0: \beta_i\cdot p_i^{\alpha-1} = \beta_j\cdot p_j^{\alpha-1},$$

which implies the condition of proposition 3.10:

$$\forall i,j\in E \text{ with } p_i,p_j>0: \underbrace{\beta_i\cdot p_i^{\alpha-1}+\gamma}_{=\overline{v}_i(\boldsymbol{p})} = \underbrace{\beta_j\cdot p_j^{\alpha-1}+\gamma}_{=\overline{v}_j(\boldsymbol{p})}. \qquad (4.6)$$

This proves the proposition. □

A special case of proposition 4.1 ($E = \{1, 2\}$ and $\gamma = 0$) was stated by Rachlin et al. (1980, p. 365). The proposition extends the earlier result to problems of distributed choice with an arbitrary number of alternatives and with $\gamma \geq 0$. The case of $\gamma = 0$ was studied by Rachlin et al. (1980) because it corresponds to ratio or interval schedules of reinforcement and a CES utility function (see section 4.2). By allowing $\gamma > 0$, the proposition also covers situations that were analysed by Herrnstein (1997, pp. 204, 277, 282).

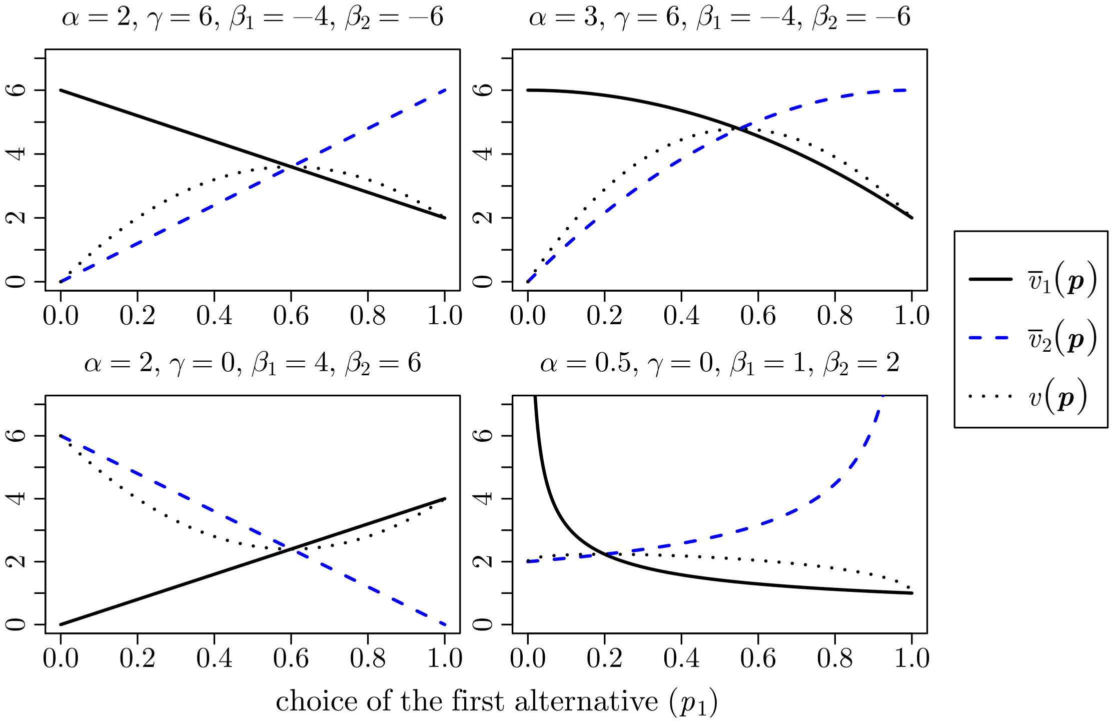


Figure 4.1: The situation of proposition 4.1

Some examples of the situation of proposition 4.1 are seen in figure 4.1. The choice is between two alternatives. The x-axis depicts the frequency p_1 of choosing the first alternative. The average values of each alternative $\overline{v}_1(\boldsymbol{p})$ and $\overline{v}_2(\boldsymbol{p})$, and the total value $v(\boldsymbol{p})$ are drawn for each $\boldsymbol{p} = (p_1, p_2) \in \mathcal{P}$ and for a set of different parameter combinations. Since definition 3.4 requires that the component value functions $\{v_j\}_{j \in E}$ are non-negative, the set of parameters of proposition 4.1 is limited to $(\beta_j)_{j \in E} \geq -\gamma$ if $\alpha \geq 1$ and to $(\beta_j)_{j \in E} \geq 0$ if $\alpha < 1$.

Two examples with $\alpha > 1$ and $\gamma > 0$ are shown in the plots of the upper row of figure 4.1. The average values decrease if their relative frequencies of choice increase. The parameters $(\beta_1, \beta_2) = (-4, -6)$ stand for the rates of decrease, and α specifies the shape of the curves. In both cases, the point $\boldsymbol{p}$ with the maximum overall value $v(\boldsymbol{p})$ coincides with the intersection of the two curves of $\overline{v}_1$ and $\overline{v}_2$. The latter is an element of the matching law solution.

If $\alpha > 1$ and $\beta_1, \beta_2 > 0$, the average values increase with their frequencies of choice. This is pictured in the left plot of the lower row. There exist some situations that can be modelled by $\alpha > 1$ and $\beta_1, \beta_2 > 0$. For instance, the average value of playing a musical instrument or speaking a foreign language may increase with the frequency of choosing this activity. In the particular example of figure 4.1, the maximum of v is at $p_1 = 0$, which is an element of the matching law solution. The remaining plot of figure 4.1 pictures an example with $\alpha < 1$ and $\gamma = 0$. In this case, β_1 and β_2 must be greater than zero. Situations with $\alpha < 1$ and $\gamma = 0$ are examined in the next section.

4.2 Constant elasticity of substitution utility

In chapter 2, multiple factors that cause deviations from the strict matching law (equation (2.1)) were mentioned. One factor that leads to undermatching and is captured by the generalised matching law (equation (2.3)) is the level of substitutability of the reinforcements (Baum and Nevin, 1981; Green and Freed, 1993). According to Rachlin et al. (1980), the level of substitutability can be taken into account by a CES utility function. The arguments of Rachlin et al. (1980) are retraced and elaborated in this section.

Given a set of alternatives $E = \{1, 2, \ldots, m\}$, $m \in \mathbb{N}$, the budget function g of equation (3.2) is supposed to describe behavioural experiments that use ratio or interval schedules (Rachlin et al., 1980, p. 362). In accordance with the situation of proposition 4.1, the reinforcement schedules of all alternatives are required to be of the same kind: $r_1 = \cdots = r_m =: r$. This means that, with $0 < r \leq 1$ and $b_1, b_2, \ldots, b_m > 0$,

$$g(\boldsymbol{p}) = \left(b_1 \cdot p_{e_1}^r, b_2 \cdot p_{e_2}^r, \ldots, b_m \cdot p_{e_m}^r\right), \text{ for all } \boldsymbol{p} = (p_{e_1}, p_{e_2}, \ldots, p_{e_m}) \in \mathcal{P}. \quad (4.7)$$

Furthermore, the preferences of an actor are assumed to follow a *constant elasticity of substitution (CES) utility function* (Arrow et al., 1961; Dixit and Stiglitz, 1977). This function is given by

$$u(\boldsymbol{x}) = \left(\sum_{j\in E} a_j \cdot x_j^\rho\right)^{\frac{1}{\rho}}, \text{ for all } \boldsymbol{x} = (x_j)_{j\in E} \in \mathcal{X}, \tag{4.8}$$

with $\rho \leq 1$, $\rho \neq 0$, and $a_j > 0$, for all $j \in E$. A transformation of this function to $u(\boldsymbol{x})^\rho$ does not change its extrema. If $\rho > 0$, it is a positive monotonic transformation with $u(\boldsymbol{x}) > u(\boldsymbol{y}) \Rightarrow u(\boldsymbol{x})^\rho > u(\boldsymbol{y})^\rho$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. If $\rho < 0$, maxima become minima and vice versa because $u(\boldsymbol{x}) > u(\boldsymbol{y}) \Rightarrow u(\boldsymbol{x})^\rho < u(\boldsymbol{y})^\rho$.

The transformed utility function $u(\boldsymbol{x})^\rho$ is clearly additive and non-negative on $\mathcal{X} = [0,\infty)^E$. The concatenation of the budget function g and the utility function u^ρ leads to the following component value functions:

$$v_j(\boldsymbol{p}) = a_j \cdot \left(b_j \cdot p_j^r\right)^\rho, \text{ for each } j \in E \text{ and } \boldsymbol{p} = (p_j)_{j\in E} \in \mathcal{P}. \tag{4.9}$$

With $\beta_j = a_j \cdot b_j^\rho$, $\alpha = r \cdot \rho$, and $\gamma = 0$, this corresponds to the situation of proposition 4.1. It was shown in the proof that all extrema (not only the maxima) of the value function $v(\boldsymbol{p}) = \sum_{j\in E} v_j(\boldsymbol{p})$ are elements of the matching law solution. Consequently, *the behaviour of an actor conforms to the matching law in case of CES preferences and a situation described by the budget function* (4.7).

A prominent example of a CES utility function is, with $\rho = 1$, the linear utility function:

$$u(\boldsymbol{x}) = \sum_{j\in E} a_j \cdot x_j, \text{ for all } \boldsymbol{x} = (x_j)_{j\in E} \in \mathcal{X}.$$

Situations that are modelled by linear functions cover reinforcements with *infinite elasticity of substitutability*, which means that they are perfect substitutes. This was evident in the original experiments of Herrnstein (1961), in which strict matching was discovered. The reinforcers were of the same kind and, hence, perfect substitutes. But also experiments with differences in the quality of reinforcements can be modelled by a linear utility function as long as the parameters a_j are set to *appropriate* values.

In case of $\rho < 1$, a finite elasticity of substitutability is modelled, and the resources, which are used as reinforcements, somehow complement each other. If, for example, two different kinds of food are used as reinforcements and both resources are subject to satiation, the subjective value of an additional unit of either resource decreases with its repeated consumption. When continuously consuming the first kind, receiving the second kind from time to time actually increases the subjective value of the first one. Therefore, a mix of both resources results in a higher average value than the consumption of a single resource.

In the limit $\rho \to 0$, equation (4.8) approaches the Cobb-Douglas utility function (see e.g. Saito, 2011). It has the following form:

$$u(\boldsymbol{x}) = \prod_{j \in E} x_j^{a_j}, \text{ for all } \boldsymbol{x} \in \mathcal{X}, \tag{4.10}$$

with $a_j > 0$, for all $j \in E$.

Cobb-Douglas preferences can be described by a additive utility function because function (4.10) is unique up to positive monotone transformation. Therefore, the logarithmic transformation represents the same preferences:

$$\log u(\boldsymbol{x}) = \sum_{j \in E} a_j \log x_j, \text{ for all } \boldsymbol{x} = (x_j)_{j \in E} \in \mathcal{X}.$$

However, the components $u_j(x_j) = a_j \log x_j$ take negative values if $x_j \in (0, 1)$. Since this holds true for any positive linear transformation of the components, there is no additive representation of Cobb-Douglas preferences with non-negative components $u_1, \ldots, u_m$.

Even though Cobb-Douglas preferences cannot be exactly represented by an additive and non-negative utility function, they can be arbitrarily closely approximated by

$$u(\boldsymbol{x}) = \sum_{j \in E} a_j \cdot x_j^{\rho}, \text{ for all } \boldsymbol{x} = (x_j)_{j \in E} \in \mathcal{X} \tag{4.11}$$

and small $\rho > 0$. In combination with the budget function g of equation (4.7), equation (4.11) corresponds to the situation of proposition 4.1. Consequently, behaviour of an actor with Cobb-Douglas preferences in a situation described by the budget function (4.7) approximates the matching law.

A similar result is obtained in the limit of $\rho \to -\infty$. The CES utility function approaches the Leontief utilities (see e.g. Saito, 2011), which are given by

$$u(\boldsymbol{x}) = \min_{j \in E} \left\{ \frac{x_j}{a_j} \right\}, \text{ for all } \boldsymbol{x} = (x_j)_{j \in E} \in \mathcal{X}.$$

Leontief preferences are not representable by an additive function, but they can be approximated by equation (4.11) and a small $\rho < 0$. Therefore, actors with Leontief preferences approach the matching law in the situation of equation (4.7).

A situation with $\rho < 0$ occurs if the reinforcement of one alternative is available only in combination with the reinforcement of another alternative. This leads to an inversion of the strict matching law because the less reinforced alternative is chosen more often than the highly reinforced alternative. An example was studied by Hursh (1978). In experiments with rhesus monkeys, food and water were provided as reinforcements of two alternatives. It was observed that a higher rate of food as reinforcement led to an increase in responding to the water alternative. Due to the fact that no food or water was provided outside of the experiment, food was valuable only in combination with water.

An inversion of the matching law can also be caused by situational properties. This is modelled by the budget function of equation (4.7) and $r < 0$. For example, the choice of one alternative may release food that remains behind a barrier. Only the choice of another alternative sometimes opens the barrier. In other words, the reinforcement of one alternative is not experienced until the other alternative is chosen. Consequently, if the reinforcement rate of the first alternative is increased, a subject chooses the second alternative more often.

4.3 Suboptimal matching

In regard to the opposite direction of proposition 4.1, an element $\boldsymbol{p}$ of the matching law solution does generally not result in an optimal value $v(\boldsymbol{p})$. Given the situation of proposition 4.1 with two alternatives ($E = \{1, 2\}$) and $\alpha = \gamma = 0$, it was already shown that matching is not always optimal: if $\beta_1 > \beta_2$, the exclusive choice of alternative 2 ($\boldsymbol{p} = (0, 1)$) is an element of the matching law solution but no optimal distribution of choice.

Moreover, the optimal distribution is no element of the matching law solution in many situations that differ from the ones of proposition 4.1. An instance is specified by the following component value functions if $\gamma_i \neq \gamma_j$ whenever $i \neq j$:

$$v_i(\boldsymbol{p}) = \beta_i \cdot p_i^{\alpha} + \gamma_i \cdot p_i, \text{ for all } \boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}. \tag{4.12}$$

In figure 4.2, two examples with $\alpha = 2$ are given. The choice is between two alternatives. The first plot pictures a situation with $\gamma_1 = 6$ and $\gamma_2 = 10$. If $p_1 = 0$, $p_1 = 1$, or $p_1 = \frac{1}{6}$, the behaviour corresponds to the matching law, whereas the optimal distribution requires that $p_1 = \frac{1}{3}$. A difference between optimal distribution and matching law is more clearly seen in the second plot, in which also the slope parameters β_1 and β_2 vary. The matching law solution is given by $\left\{(1,0), (0,1), \left(\frac{1}{3}, \frac{2}{3}\right)\right\}$ and the optimal distribution by $\left(\frac{7}{12}, \frac{5}{12}\right)$.

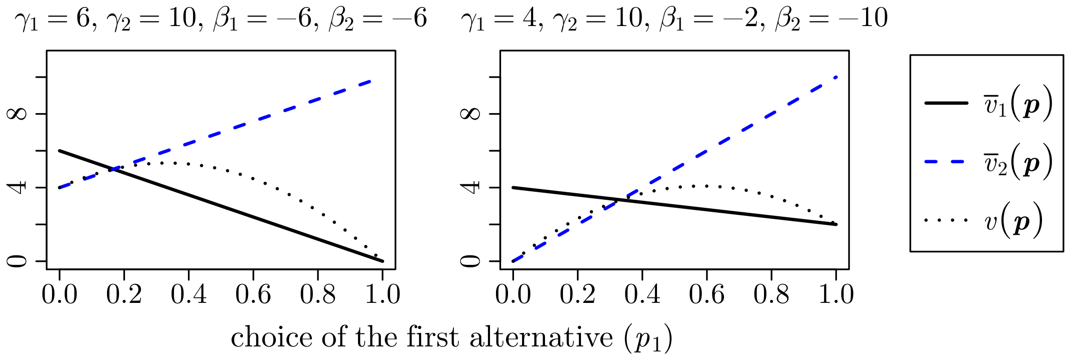


Figure 4.2: The situation of condition (4.12) with $\alpha = 2$

Furthermore, if different schedules of reinforcement are used in an experiment, the third parameter α of equation (4.2) varies such that $\alpha_i \neq \alpha_j$ whenever $i \neq j$:

$$v_i(\boldsymbol{p}) = \beta_i \cdot p_i^{\alpha_i} + \gamma_i \cdot p_i, \text{ for all } \boldsymbol{p} = (p_j)_{j \in E} \in \mathcal{P}. \tag{4.13}$$

Two instances of this situation are represented by the plots of figure 4.3. The situation of the left-sided plot implies an optimal distribution at $p_1 = 0.45$, which is no element of the matching law solution. A more noticeable difference is found in the second plot, in which the highest average value is obtained at $p_1 = 0.5$.

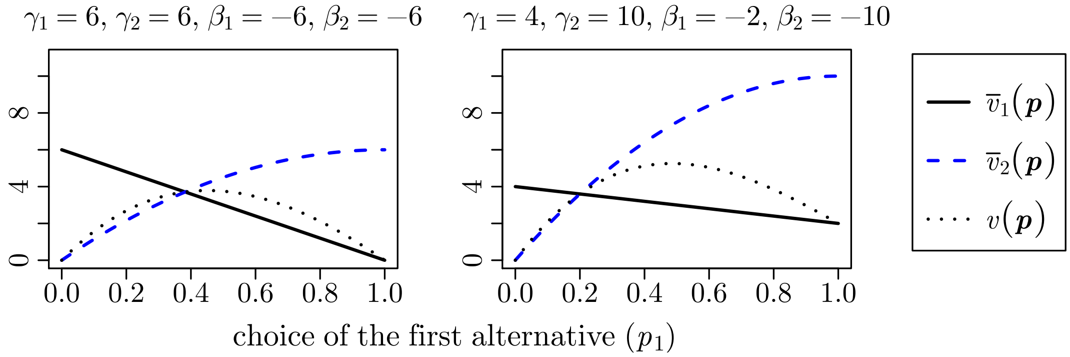


Figure 4.3: The situation of condition (4.13) with $\alpha_1 = 2$ and $\alpha_2 = 3$

In the three applications that follow, the optimal behaviour of a problem of distributed choice diverges from the matching law. One application comes from Herrnstein and Prelec (1991) and concerns the choice between two consumable goods. Second, the penalty kick example of chapter 2 is revisited. The third application is the study of addictive behaviour by Herrnstein and Prelec (1992).

4.3.1 The choice between pizza and salad

A situation that resembles the left-sided plot of figure 4.2 may arise in regard to the daily selection of food for lunch (Herrnstein and Prelec, 1991). In a highly simplified model, an actor chooses between two different items, e.g. pizza (alternative 1) and salad (alternative 2). A distribution $\boldsymbol{p} = (p_1, p_2)$ gives the relative frequencies of choice over a period of several days. As pictured in the left-sided plot of figure 4.2, the actor shows a general preference for salad ($\gamma_2 > \gamma_1$). Because of satiation or deprivation effects, the average values of pizza $\overline{v}_1$ and salad $\overline{v}_2$ change with the respective frequencies of choice. For example, if pizza is chosen every day, its average value is lower than if chosen every second day.

The matching law predicts the exclusive choice of pizza, the exclusive choice of salad, or any distribution $\boldsymbol{p}$ with $p_1 > 0$, $p_2 > 0$, and $\overline{v}_1(\boldsymbol{p}) = \overline{v}_2(\boldsymbol{p})$, which is, in the given example, $\boldsymbol{p} = \left(\frac{1}{6}, \frac{5}{6}\right)$. The actor maximises the overall value of lunch by choosing a mix of both alternatives with $p_1 = \frac{1}{3}$. This means that, if the actor chooses optimally, the average value of pizza $\overline{v}_1(\boldsymbol{p})$ is lower than the average value of salad $\overline{v}_2(\boldsymbol{p})$. It may seem unreasonable to choose pizza regularly if the average value of salad is actually higher. Similar examples gave rise to a series of articles by Herrnstein and colleagues that argue for the matching law (and, more specifically, for a learning model that leads to the matching law) instead of optimality as a fundamental solution concept of human choice behaviour (see the collection of papers in Herrnstein, 1997, part III).

4.3.2 Penalty kicks

In chapter 2, the strict matching law was illustrated by the example of penalty kicks during football games. The kicker was assumed to choose between the left and the right side of the goal. Either choice is reinforced if the ball ends up inside of the goal. If modelled by a problem of distributed choice (E, v), the set of alternatives E contains the choices L for the left and R for the right side of the goal. The specification of the value function v depends on the particular kicker as well as on the environment. Assuming that the worth of a goal is independent of the chosen side, the utility of a success can be normalised to one. This means that the average value $\overline{v}_i$ of an alternative $i \in E$ resembles its *success rate*, which is the probability of scoring a goal after choosing alternative i. Figure 4.4 shows two concrete specifications of v.

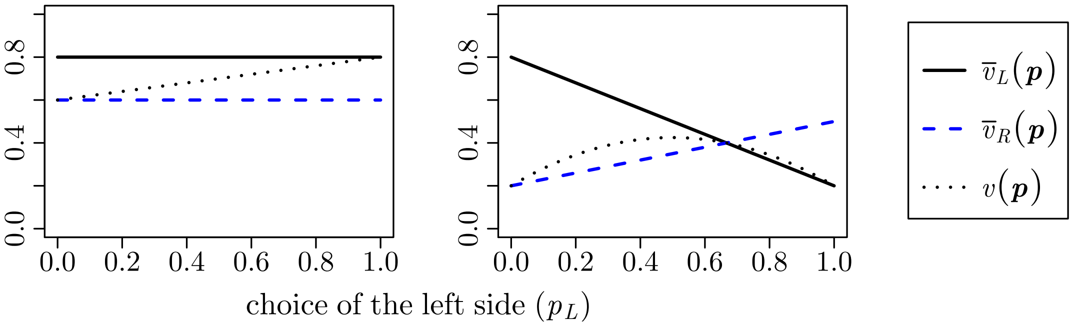


Figure 4.4: Examples of the penalty kick situation

In the first plot, the success rate is higher when choosing the left side of the goal. Both rates are independent of the frequencies of choice $\boldsymbol{p} = (p_L, p_R)$, and the kicker maximises the overall success by always choosing the left side of the goal. This conforms to the matching law.

More realistically, the success rate of an alternative $i \in E$ decreases with its frequency of choice p_i. For example, the goalkeeper might choose the left side of the goal with the same relative frequency as the kicker chooses this side. This is modelled by the second plot of figure 4.4 (the choices of kicker and goalkeeper are assumed to be independent of each other). The optimal response of the kicker is at $p_L = 0.5$, which results in $v(0.5, 0.5) = 0.425$. However, at this point, the success rate on the left side is higher than the success rate on the right side: $\overline{v}_L(0.5, 0.5) = 0.5 > 0.35 = \overline{v}_R(0.5, 0.5)$. It is plausible to assume that the kicker is going to increase the frequency of choosing the left side until the success rates are equal at $p_L = \frac{2}{3}$. This is a prediction of the matching law.

4.3.3 Addictive behaviour

In light of the previous examples, it is concluded that, on the one hand, the matching law implies the maximisation of immediate rewards. On the other hand, long-term disadvantages develop because of a reduced overall value of reinforcement. This is especially evident in the study of addictive behaviour.

Addictive behaviour is assumed to consist of the repeated choice between the consumption of an addictive substance (A) and any non-addictive alternative (N). According to Herrnstein and Prelec (1992), addictive substances are characterised by an *immediate* benefit, which means that the subjective value of consuming the addictive substance is higher than not consuming. This is shown in the left-sided graph of figure 4.5: the average value of the addictive substance $\overline{v}_A$ is always higher than the average value of the alternative behaviour $\overline{v}_N$.

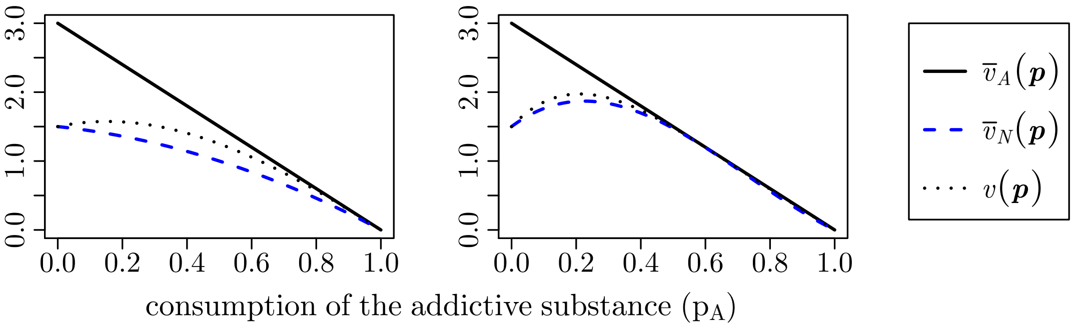


Figure 4.5: Average values of addictive (A) and alternative (N) behaviour

Similar to any other consumable good, the average value $\overline{v}_A$ of the addictive substance decreases with p_A. Another property of addictive substances is that their repeated consumption lowers the value of activities that do not involve these substances. Thus, the average value of the alternative behaviour $\overline{v}_N$ also decreases with the frequency p_A of consuming the addictive substance. As shown in the left-sided graph of figure 4.5, the optimal choice is a moderate consumption of the addictive substance. This is not a matching law equilibrium, which predicts either no or constant consumption.

Herrnstein and Prelec (1992) stated that the shape of a value function depends on the addictive substance. Substances such as alcohol, which is modelled by the right-sided graph of figure 4.5, may have different effects. It is often argued that a moderate level of consuming alcohol increases the average value of alternative behaviour because it solves problems that derive, for example, from stress or shyness. Furthermore, the value function $\overline{v}_N$ may be tangent to $\overline{v}_A$, or both functions intersect. But the optimal point is not stable in respect to the matching law as long as the average value of alcohol exceeds the average value of the alternative.

4.4 The matching law and the Nash equilibrium

Brenner and Witt (2003) analysed two-person strategic games that are repeatedly played with different partners. The authors demonstrated that the *Nash equilibrium of the stage game* corresponds to the matching law if the payoffs of the stage game do not change with the distribution of previous decisions. A Nash equilibrium is specified by a probability distribution over the alternatives such that the expected payoffs of all alternatives with strictly positive probability are equal. If the stage game is repeated independently, this property mirrors the statement of proposition 3.10 because the relative frequencies approach the probabilities of choice and the expected payoffs are approximated by the average values.

A difference between a two-person strategic game and a problem of distributed choice is that the latter does not explicitly model the decisions of the other person. In the example of penalty kicks, the goalkeeper's decisions were subsumed in the dependence of the average values $\overline{v}_L$ and $\overline{v}_R$ on the frequency distribution $\boldsymbol{p}$. The goalkeeper was assumed to take the previous decisions of the kicker into account.

Hence, from the kicker's point of view, the success rates $\overline{v}_L$ and $\overline{v}_R$ depend on the own frequency distribution $\boldsymbol{p}$.

goalkeeper		left	right
kicker	left	0.2	0.8
	right	0.5	0.2

Table 4.1: Score probabilities of a kicker in a penalty kick situation

If penalty kicks are modelled by a repeatedly played two-person strategic game, the payoffs of the stage game are typically independent of previous decisions (see e.g. Chiappori et al., 2002; Palacios-Huerta, 2003; Berger and Hammer, 2007). Table 4.1 shows a sample game by specifying a kicker's probabilities of scoring a goal given the own and the goalkeeper's decision. The probability of scoring is higher if the goalkeeper chooses not the same side as the kicker. Furthermore, since the kicker's right foot is stronger than his left foot, the score probability on the left side is higher than on the right side.

If the game of table 4.1 is repeatedly played, a problem of distributed choice can be defined. The value function depends on the strategy of the goalkeeper. For example, the goalkeeper may adopt the kicker's relative frequency p_L of choosing the left side of the goal. The value functions of this case were depicted in the right-sided plot of figure 4.4. As already noted, the matching law solution diverges from the individually optimal solution.

The Nash equilibrium requires that, given a choice probability p_L of the kicker, the goalkeeper selects a choice distribution $(q_L, q_R) \in \mathcal{P}$ that maximises his expected payoff (because the game is equivalent to a zero-sum game). In the game of table 4.1, this means that the goalkeeper always chooses the right side of the goal if $p_L < \frac{1}{3}$ and the left side of the goal if $p_L > \frac{1}{3}$. In the case of $p_L = \frac{1}{3}$, the goalkeeper is indifferent between both alternatives and chooses the left side with probability $q_L = \frac{2}{3}$, such that the kicker is also indifferent with respect to his alternatives. Figure 4.6 shows this situation from the perspective of the kicker. The optimal point ($p_L = \frac{1}{3}$) corresponds to an element of the matching law solution because $\overline{v}_L(\frac{1}{3}, \frac{2}{3}) = \overline{v}_R(\frac{1}{3}, \frac{2}{3})$.

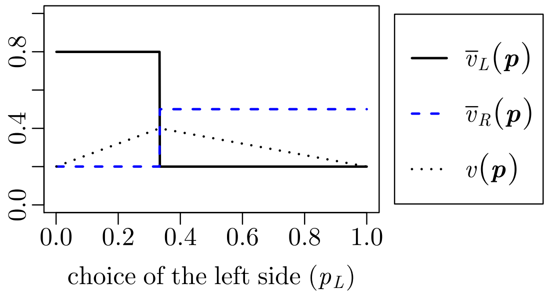


Figure 4.6: The penalty kick situation under Nash equilibrium condition

However, this result is based on the assumptions that the payoffs of the stage game are stable and the game is repeated independently. Brenner and Witt (2003) provided proof that, if a stage game is continuously played with the same partner and the payoffs depend on the distribution of previous choices, the *Nash equilibrium of the repeated game* is different from the matching law.

4.5 Conclusion

The relationship between the matching law and optimal behaviour depends on the particular problem of distributed choice. There is a class of situations in which the optimal distribution of choice implies behaviour that is in accordance with the matching law. But, in general, the matching law diverges from optimal behaviour. Further analyses of the relationship between optimisation and the matching law can be found in Baum (1981), Vaughan (1981), or Herrnstein (1997).

The matching law is different from maximisation because it neglects the long-term effects of previous decisions. More specifically, Sakai and Fukai (2008b) showed that the difference between optimisation and the matching law can be mathematically expressed by the change in the value v that stems from the change in behaviour $\boldsymbol{p}$. A similar point was made by Loewenstein et al. (2009): if actors are not able to draw a connection between past behaviour and future rewards, the matching law is consistent with the principle of utility maximisation.

Chapter 5

Melioration learning

Even though the integration of the matching law into consumer theory enables its usage as micro-level assumption, the application on the social level is still limited. One point was already made in chapter 1. The matching law describes a relationship between decisions and reinforcements that are observed over a period of time. Nothing is said about an actor's disposition of choosing one of the alternatives at a particular point in time. Hence, any hypothesis that is derived from the matching law must refer to aggregated individual behaviour.

Second, the matching law solution generally contains multiple elements because the exclusive choice of each alternative is always included (observation 3.8). There is no immediate criterion that marks one element of the matching law solution as more likely than another. Moreover, in case of $n \in \mathbb{N}$ actors and $m \in \mathbb{N}$ alternatives for each actor, there are at least m^n different outcomes that correspond to the matching law. Without a theory that selects between different elements of the matching law solution, its application on the social level is impeded.

Finally, in section 2.2.4, a macro-level regularity was derived from the matching law under the assumptions that reinforcement is external and all actors are perfectly informed about the reinforcements of other actors. The analysis becomes more complex if information is limited or if the reinforcement of one actor depends on the choices of other actors. In order to arrive at predictions for this kind of social situation, additional assumptions about the processing and evaluation of available information are required.

As stated in chapter 1, a solution to these limitations is the acceptance of a mechanism of decision-making, or a learning rule, that results in the matching law on the individual level and can be applied to the derivation of social phenomena. A wide range of existing theories might be applicable. In consideration of the behaviouristic origin of the matching law, it is focused on learning models. The best known economic models, such as regret matching, fictitious play, or Bayesian learning, were suggested as normative processes that lead to some form of individually optimal behaviour in the long run (see e.g. Young, 2004). Psychological models of learning, on the other hand, try to represent the development of human behaviour as realistic as possible while keeping it analytically tractable. According to Staddon and Cerutti (2003, p. 134), most of the psychological learning models describe processes that are consistent with the matching law.

One of the processes that are supposed to result in the matching law was introduced by Herrnstein and Vaughan (1980) and called *melioration*. In the next section, the ideas of the authors are summarised. Afterwards, an algorithm is introduced that mathematically expresses the informal description of melioration (section 5.2). It is shown that this model implies the matching law if relatively strong assumptions about the situation hold. In section 5.3, the application of melioration in sociological research is justified. Additionally, differences to other learning models and empirical studies are discussed.

5.1 Previous research

Generally speaking, melioration learning states that the development of behavioural tendencies is controlled by past experiences. More specifically, a particular behaviour is strengthened because of a high average value of previous events that were perceived as consequences of this behaviour.

In the original literature, melioration learning was defined similarly vaguely. For example, in one of the first articles, melioration meant that “behavior shifts toward higher local rates of reinforcement” (Herrnstein, 1997, p. 75). A **local reinforcement rate** was defined as “the reinforcement actually obtained from an alternative [..] divided by the time allocated to it” (Herrnstein, 1997, p. 76). In terms of definition 3.5, the local reinforcement rate can be equated with the *average* *value* of an alternative. Elsewhere, Vaughan and Herrnstein (1987) more formally describe the process of melioration by a differential equation. Let there be a two-element choice set $\{1, 2\}$. Given a point in time $t \in (0, \infty)$, $p_i(t) \in [0, 1]$ stands for the relative frequency of having chosen alternative $i \in \{1, 2\}$. Vaughan and Herrnstein (1987) state that the frequency $p_1(t)$ changes over time in accordance with the following equation:

$$\frac{dp_1(t)}{dt} = f\left(\hat{v}_1(t) - \hat{v}_2(t)\right). \tag{5.1}$$

The function $f(\cdot)$ is differentiable and strictly monotonically increasing. It must also hold that $f(0) = 0$. The term $\hat{v}_i(t)$ $(i \in \{1, 2\})$ stands for the local reinforcement rate of alternative i at time t. Herrnstein (1990a, p. 219) states that

> "[t]he melioration process continues until the stronger response displaces all others, or, because the reinforcement returns from an alternative may depend on its level of occurrence, equilibrium is attained with several alternatives left in the response set, each yielding the same returns per unit at a given allocation among them [..]."

If applied to a problem of distributed choice $(\{1, 2\}, v)$ (definitions 3.4 and 3.5), equation (5.1) transforms to

$$\frac{dp_1(t)}{dt} = f\left(\overline{v}_1(\boldsymbol{p}(t)) - \overline{v}_2(\boldsymbol{p}(t))\right), \tag{5.2}$$

with $\boldsymbol{p}(t) := (p_1(t), p_2(t))$ for all $t \in (0, \infty)$. It follows that the quotation of Herrnstein (1990a) is a verbal interpretation of the condition of proposition 3.10, which was shown to be equivalent to the matching law.

Equation (5.1) identifies melioration learning as a change in aggregated behaviour. Depending on the particular function $f(\cdot)$ and the situation, it describes a development of the distribution of relative frequencies $(p_1(t), p_2(t))$. Because $f(0) = 0$, melioration guarantees that the matching law is a stable state. And since $f(\cdot)$ is assumed to be a continuous and strictly monotonically increasing function, it seems plausible that this state is reached eventually (see also Vaughan, 1985, p. 387). However, this must not be true for every $f(\cdot)$ and every situation.

The distribution $(p_1(t), p_2(t))$ might change in accordance with equation (5.1) but never reach the matching law. Moreover, the function $f(\cdot)$ should depend on $\hat{v}_1(t) - \hat{v}_2(t)$ *and* $p_1(t)$, such that $(p_1(t), p_2(t))$ remains a frequency distribution with $p_1(t), p_2(t) \geq 0$ and $p_1(t) + p_2(t) = 1$.

Without specifying the function $f(\cdot)$ of equation (5.1), the melioration learning rule remains vague and the long-term behaviour cannot be properly analysed. A learning model is more accurate if it specifies a decision rule instead of a general change in aggregated behaviour. The decision rule may be deterministic or probabilistic, for example "given condition c, choose alternative a" or "given condition c, choose alternative a with probability q_a".

In the past, mathematically rigorous representations of melioration learning were suggested. For example, Brenner and Witt (2003, p. 432) substituted the relative frequency $p_1(t)$ of equation (5.1) by the *probability* $q_1(t) \in [0, 1]$ of choosing alternative 1 at time t and define melioration learning as

$$q_1(t + T) = q_1(t) + q_1(t)(1 - q_1(t)) \cdot \alpha \cdot (\hat{v}_1(t) - \hat{v}_2(t)). \tag{5.3}$$

In equation (5.3), $T \in (0, \infty)$ stands for the length of a time period between two choices, and α is a given constant (Brenner and Witt, 2003, p. 433). In connection with the rule "choose alternative 1 at time t with probability $q_1(t)$", equation (5.3) implies a change in the relative frequencies of choice that is *roughly* in line with equation (5.1). Since decisions are probabilistic, it is possible for the relative frequency $p_1(t)$ to decreases even if $\hat{v}_1(t) > \hat{v}_2(t)$.

Sakai et al. (2006) and Loewenstein (2010) formalised melioration learning by iterative processes that are similar to equation (5.3). They showed that the steady states of their learning rules correspond to the matching law. Loewenstein (2010) even based his rule on a theory of neural activity and synaptic plasticity. Similar neurobiological approaches to melioration and the matching law can be found in Soltani and Wang (2006) and Simen and Cohen (2009). The authors used neural networks to model a relationship between reinforcement and neural activity. In situations of repeated decision-making, the output of these models conforms to the melioration process and the matching law.

5.2 The model

In contrast to the previous models of melioration, this section presents an algorithm that is perfectly consistent with equation 5.1 and does not require probabilities of choice or neural networks. It is focused on the mental processing of reinforcement rates, which are directly evaluated when making decisions. An extensive comparison to other learning models is given in section 5.3.4.

More precisely, melioration is suggested to be formalised by an instance of the Q-learning algorithm (Watkins, 1989) with ε-greedy strategy. According to Sakai et al. (2006, p. 1102), greedy Q-learning is an "extreme case of melioration". Apart from that, no explicit connection between Q-learning, on the one hand, and melioration or the matching law, on the other hand, has previously been made. An implicit account of this connection is found in the work of Thuijsman et al. (1995) and Seth (2002). The authors analysed foraging strategies that are similar to Q-learning and compatible with the matching law in some situations.

Q-learning is an off-policy form of temporal-difference (TD) learning and originates from an area of research in artificial intelligence that is called *reinforcement learning* or RL (Sutton and Barto, 1998). While TD models were initially used to represent classical conditioning (Sutton and Barto, 1990), they can be "applied to stochastic sequential decision tasks to produce an analog of instrumental learning" (Barto et al., 1990, pp. 541-542).

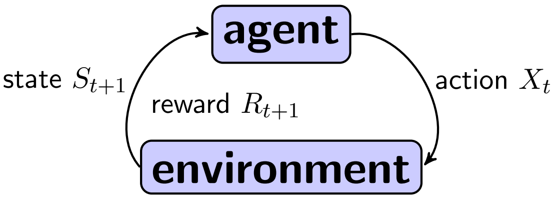


Figure 5.1: A general model of sequential decision-making

A general model of *sequential decision-making* was given by Sutton and Barto (1998, p. 52) and is illustrated in figure 5.1. The actor, who is called *agent*, interacts with the *environment* by repeatedly emitting an action and awaiting a response. The response of the environment consists of a *state* and a *reward*. A finite set E of choice alternatives and a finite set $\mathcal{S}$ of states of the environment are assumed. The interaction takes place along discrete time steps $t = 1, 2, \ldots$. At each time step t, the agent inspects the current state $S_t \in \mathcal{S}$ and chooses an element $X_t \in E$ from the set of alternatives. Subsequently, a real-valued reward R_{t+1} is received, and the state changes to $S_{t+1} \in \mathcal{S}$.

5.2.1 Markov decision processes

In compliance with most of the previous work (Sutton and Barto, 1998; Wiering and van Otterlo, 2012), it is assumed that the reinforcement learning situation can be modelled by a Markov decision process. Markov decision processes were introduced by Bellman (1957, cited by Sutton and Barto, 1998, p. 16). They denote particular stochastic models of the situation of figure 5.1. This means that $\{S_t\}_{t=1}^{\infty}$, $\{X_t\}_{t=1}^{\infty}$, and $\{R_t\}_{t=2}^{\infty}$ are stochastic processes in which the random variables have values in $\mathcal{S}$, E, and $\mathbb{R}$, respectively.

The transition between states is defined by a probability distribution. It specifies the probability of observing state $s \in \mathcal{S}$ at time $t+1$ given the realisation of all past states $S_t, \ldots, S_1$ and choices $X_t, \ldots, X_1$:

$$Pr(S_{t+1} = s \mid S_t = s_t, X_t = x_t, \ldots, S_1 = s_1, X_1 = x_1).$$

The Markov property requires that this distribution depends only on the state and action of the previous time step t. If, additionally, the probabilities do not change with time, the state transitions can be modelled by a **transition function** $p_{s,s'} : E \to [0, 1]$ for each $s, s' \in \mathcal{S}$. At any time t, the probability of the next state $S_{t+1} = s' \in \mathcal{S}$, given the current state $S_t = s \in \mathcal{S}$, the current action $X_t = e \in E$, and all previous states and actions, is

$$Pr(S_{t+1} = s' \mid S_t = s, X_t = e, \ldots, S_1 = s_1, X_1 = x_1) = p_{s,s'}(e).$$

It is required that $\sum_{s' \in S} p_{s,s'}(e) = 1$ for each $e \in E$ and $s \in \mathcal{S}$.

In an analogous manner, the expected value of the next reward R_{t+1} depends only on the current state and the current action. The reward can, therefore, be modelled by a **reward function** $r_s : E \to \mathbb{R}$, for each $s \in \mathcal{S}$, such that the expected value of the next reward is

$$\mathbb{E}[R_{t+1} \mid S_t = s, X_t = e, \ldots, S_1 = s_1, X_1 = x_1] = r_s(e).$$

The following definition (adapted from van Otterlo and Wiering, 2012, p. 12) summarises the assumptions that are made for the reinforcement learning situation.

Definition 5.4. *Let $\mathcal{S}, E$ be finite sets, $p_{s,s'} : E \to [0,1]$, for each $s, s' \in \mathcal{S}$, and $r_s : E \to \mathbb{R}$, for each $s \in \mathcal{S}$. A* ***Markov decision process*** *is given by the stochastic processes $\{S_t\}_{t=1}^{\infty}$, $\{X_t\}_{t=1}^{\infty}$, and $\{R_t\}_{t=2}^{\infty}$ if the random variables S_t, X_t, and R_t have values in $\mathcal{S}$, E, and $\mathbb{R}$, respectively, and if, for all $t \in \mathbb{N}$,*

1. *$Pr(S_{t+1} = s' | S_t = s, X_t = e, \ldots) = p_{s,s'}(e)$ for all $s, s' \in \mathcal{S}, e \in E$, and*

2. *$\mathbb{E}[R_{t+1} \mid S_t = s, X_t = e, \ldots] = r_s(e)$ for all $s \in \mathcal{S}, e \in E$.*

The decisions of an agent are modelled by a **policy** $\pi := \{q_s\}_{s \in \mathcal{S}}$ with

$$q_s : E \to [0,1] \text{ and } \sum_{e \in E} q_s(e) = 1, \text{ for each } s \in \mathcal{S}.$$

Given a state $s \in \mathcal{S}$, $q_s(e)$ stands for the probability of choosing alternative $e \in E$. A policy is said to be **optimal** for a Markov decision process if it maximises the *return* of this process. In case of a decision process without determinable end, the **return** at time $t_0 \in \mathbb{N}$ is defined by a discounted sum:

$$\sum_{i=0}^{\infty} \gamma^i R_{t_0+i+1}.$$

The **discount rate** $\gamma \in [0,1)$ implies a decreasing interest in future rewards.

Let π be a policy and $t_0 \in \mathbb{N}$ be any point in time. The expected return of taking action $e \in E$ in state $s \in \mathcal{S}$ and following policy π thereafter is given by

$$Q^{\pi}(s,e) := \mathbb{E}_{\pi}\left[\sum_{i=0}^{\infty} \gamma^i R_{t_0+i+1} \middle| S_{t_0} = s, X_{t_0} = e\right].$$

$\mathbb{E}_{\pi}[\cdot]$ denotes the expected value under the policy π. Q^{π} is called **action-value function for policy** π (Sutton and Barto, 1998, p. 69). Note that this function is independent of the time t_0 because the functions $p_{s,s'}$,and r_s are independent of t_0. The **optimal action-value function** is defined by

$$Q^*(s,e) := \max_{\pi} Q^{\pi}(s,e), \text{ for all } s \in S, e \in E.$$

If all elements of a Markov decision process are known, the action-value function can be calculated for a particular policy π. Furthermore, it is possible to determine an optimal policy by dynamic programming (Sutton and Barto, 1998, pp. 89-110). But also without the knowledge of the transition function $p_{s,s'}$ or the reward function r_s, for any $s, s' \in \mathcal{S}$, an optimal policy is obtainable. This requires multiple encounters with the situation and an appropriate strategy that iteratively updates estimates of $Q^*(s,e)$.

5.2.2 The melioration algorithm

There has been extensive research on reinforcement learning methods that guarantee an agent to approach the optimal policy without knowing the functions $\{p_{s,s'}\}_{s,s' \in \mathcal{S}}$ and $\{r_s\}_{s \in \mathcal{S}}$ of a Markov decision process (an introduction is given by Sutton and Barto, 1998, and a more recent review by Wiering and van Otterlo, 2012). There is no best method. Even though most algorithms converge to the optimal policy, there are differences in efficiency, which is measured by the number of repeated encounters that are required to be close to the optimal policy.

While the search for optimal behaviour is intriguing, this is not the goal of the present chapter. On the contrary, it is looked for a formal model of melioration learning, which is not necessarily optimal. Fortunately, one instance of a reinforcement learning method is very similar to the ideas of Vaughan and Herrnstein (1987) about melioration learning. This instance is called *Q-learning with* $\gamma = 0$ *and* ε*-greedy strategy.*

The Q-learning algorithm was introduced by Watkins (1989) and is one of the most basic and popular methods to estimate action-value functions if $\{p_{s,s'}\}_{s,s' \in \mathcal{S}}$ and $\{r_s\}_{s \in \mathcal{S}}$ are not known (van Otterlo and Wiering, 2012, p. 31). It uses a table of **Q-values**, $Q_t(s,e)$ for each $s \in \mathcal{S}$ and $e \in E$, that are iteratively updated. Initially, all Q-values ($Q_1(s,e)$) are set to zero. For every round $t \in \mathbb{N}$, there are realisations of S_t and X_t. Given the subsequent reward $R_{t+1} = y$ and the new state $S_{t+1} = s'$, the Q-values are updated by the following rule:

$$Q_{t+1}(s,e) = \begin{cases} Q_t(s,e) + \alpha_t(s,e) \cdot (y + \gamma \cdot V_t(s') - Q_t(s,e)) & \text{, if } S_t = s, X_t = e, \\ Q_t(s,e) & \text{, else.} \end{cases}$$

Each sequence $(\alpha_t(s,e))_{t=1}^{\infty}$ is from $[0,1]$, and $V_t(s') := \max_{e' \in E} Q_t(s',e')$. It can be shown that, for each $s \in \mathcal{S}$ and $e \in E$, $Q_t(s,e)$ converges towards the optimal values $Q^*(s,e)$ as $t \to \infty$. Besides an appropriate convergence of the sequences $(\alpha_t(s,e))_{t=1}^{\infty}$ towards zero, it is usually assumed that the rewards R_t (or at least their variances) are bounded and that each state-action pair occurs infinitely often (Watkins and Dayan, 1992; Jaakkola et al., 1994; van Otterlo and Wiering, 2012).

Let $K_t(s,e)$ indicate the frequency of having chosen $e \in E$ in state $s \in \mathcal{S}$ before time step $t \in \mathbb{N}$. This means that $K_1(s,e)$ is zero for each $s \in \mathcal{S}$ and $e \in E$. If $\gamma = 0$ and $\alpha_t(s,e) = \frac{1}{K_t(s,e)+1}$, Q-learning reduces to

$$Q_{t+1}(s,e) = \begin{cases} Q_t(s,e) + \frac{1}{K_t(s,e)+1} \cdot (y - Q_t(s,e)) & \text{, if } S_t = s, X_t = e, \\ Q_t(s,e) & \text{, else.} \end{cases} \tag{5.5}$$

Let $s \in \mathcal{S}$ and $e \in E$, and $y_1, y_2, \ldots, y_{K_t(s,e)} \in \mathbb{R}$ denote all rewards that were received after choosing action $e \in E$ in state $s \in \mathcal{S}$ before time step $t \in \mathbb{N}$. In the words of Vaughan and Herrnstein (1987), this means that, if $K_t(s,e) > 0$,

$$\frac{1}{K_t(s,e)} \sum_{i=1}^{K_t(s,e)} y_i$$

denotes the *local reinforcement rate* of action e in state s at time t. Given equation (5.5), it holds that

$$Q_t(s,e) = \begin{cases} \frac{1}{K_t(s,e)} \sum_{i=1}^{K_t(s,e)} y_i & \text{, if } K_t(s,e) > 0, \\ 0 & \text{, if } K_t(s,e) = 0, \end{cases} \quad \text{for all } t \in \mathbb{N}, s \in \mathcal{S}, e \in E.$$

This is evident after transforming the first row of equation (5.5) (see also Sutton and Barto, 1998, p. 37):

$$Q_{t+1}(s,e) \;=\; \tfrac{1}{K_t(s,e)+1} \cdot (y + K_t(s,e) \cdot Q_t(s,e)) = \tfrac{1}{K_t(s,e)+1} \cdot \left(y + \textstyle\sum_{i=1}^{K_t(s,e)} y_i\right).$$

Consequently, a Q-value $Q_t(s, e)$ gives the local reinforcement rate of action $e \in E$ in state $s \in \mathcal{S}$ at time $t \in \mathbb{N}$. If, additionally, an agent always chooses an action with the currently highest value of $Q_t(s, e)$, the relative frequency of this action increases as required by equation (5.1). Therefore, *updating the local reinforcement rates by equation* (5.5) *and always choosing the action with the highest Q-value* resembles *melioration learning.*

Algorithm 5.2.1 The melioration learning algorithm

```
Require: exploration rate $\varepsilon \in (0, 1)$, set of alternatives $E$
$t \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    observe state $s$
    if $s$ is a new state then
        initialise $Q_t(s, e) \leftarrow 0$, for all $e \in E$
        initialise $K_t(s, e) \leftarrow 0$, for all $e \in E$
    end if
    $\varepsilon_s \leftarrow \frac{\varepsilon}{1+\sum_{j \in E} K_t(s,j)}$
    if $\varepsilon_s >$ random number between 0 and 1 (uniform distribution) then
        chose a random action $e \in E$ using a uniform distribution
    else
        choose action $e \in E$ greedily using the Q-values
    end if
    observe reward $y$
    $K_{t+1}(s, e) \leftarrow K_t(s, e) + 1$
    $Q_{t+1}(s, e) \leftarrow Q_t(s, e) + \frac{1}{K_{t+1}(s,e)} \cdot (y - Q_t(s, e))$
    for all $s' \neq s$ and $e' \neq e$ do
        $K_{t+1}(s', e') \leftarrow K_t(s', e')$
        $Q_{t+1}(s', e') \leftarrow Q_t(s', e')$
    end for
until termination
```

Algorithm 5.2.1 gives the formal model of melioration learning. The main differences to the ideas of Vaughan and Herrnstein (1987) are the distinction between different states of the environment and the exploration rate ε. The maximally exploiting strategy of *greedily* selecting an action with the highest Q-value has the disadvantage of exclusively choosing a single alternative as soon as it has been reinforced by a strictly positive reward. A trade-off between the exploitation of the currently best actions and the exploration of other actions can be made by the ε-greedy strategy (Sutton and Barto, 1998, p. 28). The parameter $\varepsilon \in (0, 1)$ specifies the level of exploration. More specifically, the ε-greedy strategy states that the currently best action is chosen with probability $1 - \varepsilon$, and a random action otherwise. If an agent maintains an exploration parameter ε_s for each state $s \in \mathcal{S}$ and these parameters decrease as specified in line 9 of algorithm 5.2.1, every state-action pair appears infinitely often and the behaviour is greedy in the limit $t \to \infty$ (Singh et al., 2000, p. 304).

Let there be a state $s \in \mathcal{S}$ that occurs infinitely often. For any two actions $e_1, e_2 \in E$ that are chosen with a strictly positive relative frequency in this state, algorithm 5.2.1 implies that the Q-values $Q_t(s, e_1)$ and $Q_t(s, e_2)$ approach each other in the long run with probability one. This resembles the condition of the *matching law*, which was given in proposition 3.10.

Proposition 5.6. *Consider a Markov decision process with* $Var(R_t) < \infty$*, for all* $t \in \mathbb{N} \setminus \{1\}$. *It is assumed that an agent follows algorithm 5.2.1 in order to update a table of Q-values:* $Q_t(s, e)$ *for each* $s \in \mathcal{S}$ *and* $e \in E$. *For any state* $s \in \mathcal{S}$ *that is visited infinitely often and any two* $e_1, e_2 \in E$*:*

$$Pr\left(\lim_{t\to\infty} Q_t(s, e_1) - Q_t(s, e_2) = 0 \middle| \lim_{t\to\infty} \tfrac{1}{t} K_t(s, e_1) > 0, \lim_{t\to\infty} \tfrac{1}{t} K_t(s, e_2) > 0\right) = 1.$$

Proof. Let $i \in \{1, 2\}$ and

$$\alpha_t(s, e_i) := \begin{cases} \frac{1}{K_t(s,e_i)} & \text{, if } S_t = s, X_t = e_i, \\ 0 & \text{, else.} \end{cases}$$

From $\lim_{t\to\infty} \frac{1}{t} K_t(s, e_i) > 0$ follows that $\lim_{t\to\infty} K_t(s, e_i) = \infty$. It also holds that $\sum_{t=1}^{\infty} \alpha_t(s, e_i) = \infty$ and that $\sum_{t=1}^{\infty} (\alpha_t(s, e_i))^2 < \infty$ (Riemann zeta functions). This means that theorem 1 of Singh et al. (2000, p. 294) can be applied. It states that $Q_t(s, e_i)$ converges towards $Q^*(s, e_i)$, as $t \to \infty$, with probability one.

It is assumed that $Q^*(s, e_1) > Q^*(s, e_2)$. Because of the convergence of $Q_t(s, e_1)$ and $Q_t(s, e_2)$, there must exist a $t_0 \in \mathbb{N}$ such that for all $t > t_0$: $Q_t(s, e_1) > Q_t(s, e_2)$

with probability one. According to algorithm 5.2.1, this implies that action e_2 is chosen with probability ε_s and independently of the previous choices at every time step t with $t > t_0$. And since ε_s converges towards zero as $t \to \infty$, it follows that

$$\mathbb{E}\left(\mathbf{1}_{\{X_t=e_2,S_t=s\}} \mid \mathbf{1}_{\{X_{t-1}=e_2,S_{t-1}=s\}}, \ldots, \mathbf{1}_{\{X_1=e_2,S_1=s\}}\right) \xrightarrow{\text{a.s.}} 0 \text{ as } t \to \infty.$$

Because the variance of $\mathbf{1}_{\{X_t=e_2,S_t=s\}}$ is between zero and one for all $t \in \mathbb{N}$, the *stability theorem* of Loève (1978, p. 53) yields:

$$\frac{1}{t}\sum_{i=1}^{t} \mathbf{1}_{\{X_i=e_2,S_i=s\}} - \mathbb{E}\left(\mathbf{1}_{\{X_i=e_2,S_i=s\}} \mid \mathbf{1}_{\{X_{i-1}=e_2,S_{i-1}=s\}}, \ldots, \mathbf{1}_{\{X_1=e_2,S_1=s\}}\right) \xrightarrow{\text{a.s.}} 0.$$

Since

$$\frac{1}{t}K_t(s,e_2) = \frac{1}{t}\sum_{i=1}^{t} \mathbf{1}_{X_i=e_2,S_i=s} \xrightarrow{\text{a.s.}} 0,$$

a premiss is violated. In an analogous manner, a contradiction can be drawn from the assumption $Q^*(s,e_1) < Q^*(s,e_2)$. This proves the proposition. □

Two short remarks conclude this section about the melioration learning algorithm. First, it follows from the proof of proposition 5.6 that, instead of a Markov decision process, the convergence of $Q_t(s,e_i)$, for every $i \in \{1,2\}$, would be a sufficient condition of the proposition. Also in situations that violate the Markov property, for example by containing a more complex reward structure, the behaviour of an agent approaches the matching law as long as the Q-values converge. It is also not necessary that the Q-values converge to the optimal action-value function $Q^*(s,e_i)$. The proof works with any finite limit.

Second, since $\gamma = 0$, the optimal action-value function is given by

$$Q^*(s,e_i) = \max_{\pi}\left\{\mathbb{E}_\pi\left[R_{t_0+1} \mid S_{t_0} = s, X_{t_0} = e_i\right]\right\} = r_s(e_i),$$

for all $i \in \{1,2\}$ and $t_0 \in \mathbb{N}$. This expression is usually not understood as an optimal result but as *myopic* outcome (Sutton and Barto, 1998, p. 58). A similar conclusion was made at the end of the previous chapter: the matching law is not optimal because it neglects the long-term effects of previous decisions.

5.3 Discussion

This section contains several arguments that justify the usage of melioration learning in sociological research. Despite its recurrent appearance in the psychological literature, there are very few references to melioration in sociological papers. Nevertheless, learning mechanisms are relevant, especially in the fields of *Rational Choice Sociology* (Braun and Gautschi, 2011) and *Analytical Sociology* (Hedström and Bearman, 2009). In fact, there has been some effort to introduce reinforcement learning models to sociological theory (Macy and Flache, 2009, pp. 250-251). Melioration could constitute one of these learning models. As shown in the next chapters, it can serve as an explanation of social phenomena. Moreover, it is argued in section 5.3.1 that the basic ideas of melioration learning are deeply rooted in past sociological work. Additionally, the usage of the exploration rate in algorithm 5.2.1 is motivated in section 5.3.2.

Another justification of the melioration model stems from the limitations of the matching law. It was pointed out in the beginning of this chapter that there are multiple equilibrium outcomes that conform to the matching law. Which outcome finally emerges depends on the initial conditions as well as on the particular rule of decision-making. Furthermore, it is of interest how an equilibrium state arises and what dynamics to expect if perturbations appear. Similar arguments for the usage of learning models are found in the economic literature (e.g. Camerer and Ho, 1999). However, in economics it is usually asked: “Does behaviour converge to an optimum?” (Friedman et al., 1995, p. 164) rather than “Does behaviour converge to the matching law?” Whether the first or the second question is more relevant is an empirical question. Research regarding the empirical status of melioration learning and its stable states is summarised in section 5.3.3.

In section 5.3.4, the particular melioration learning model of section 5.2.2, which means Q-learning with ε-greedy, is discussed. Generally, Q-learning was regarded as too simple to account for real human behaviour (Shteingart and Loewenstein, 2014), and more appropriate learning models have been suggested in the past. Therefore, advantages and disadvantages of the present implementation of melioration learning are listed. Also a comparison with popular economic models of learning is presented.

5.3.1 Melioration and past sociological research

In the past, several attempts were made to include findings from behavioural psychology into sociological theory. There are numerous books and articles about the potentials of a *behavioural sociology* (Homans, 1961; Burgess and Bushell, 1969; Opp, 1972; Kunkel, 1975; Hamblin and Kunkel, 1977; Michaels and Green, 1978; Molm, 1981). An explicit connection between behavioural psychology and sociology is found in the early works on social exchange theory (Emerson, 1969, 1972a; Molm and Wiggins, 1979). In particular, Emerson (1972b) attempted to build a theory of power and dependence in exchange networks entirely on axioms and propositions from operant conditioning.

However, in social exchange theory as well as in other fields, behavioural sociology did not hold out against other schools of thought, such as rational choice theory. A possible reason is the (previous) lack of a formal framework that allows the derivation of empirically testable and falsifiable hypotheses from behavioural regularities. In fact, most approaches that are able to derive testable hypotheses about social behaviour are based on rational choice theory. Since there are many situations in which matching behaviour clearly deviates from individually optimal behaviour (chapter 4), the melioration model should be regarded as an alternative to rational choice or as a theory of bounded rationality.

In contrast to many rational choice models, melioration does not require full information or high cognitive skills. It is close to two other models of learning that were recently introduced to sociological theory: the Bush-Mosteller and the Roth-Erev model (Macy and Flache, 2002). The predictions of those models differ from game-theoretic solutions: "[a]pplied to social dilemmas, both the [Bush-Mosteller] and Roth-Erev models identify a key difference with analytical game-theoretic solutions: the existence of a cooperative equilibrium that is not Nash equivalent, even in Stag Hunt games where mutual cooperation is also a Nash equilibrium" (Macy and Flache, 2002, p. 7230). The difference between these two models of learning and the melioration model is elaborated in section 5.3.4.

Finally, the matching law was considered as an explanation of social behaviour in early sociological theory. More specifically, Homans (1974, pp. 21-22) introduced the matching law as a quantification of his first proposition of individual behaviour (success proposition). Additionally, his *rationality proposition* (Homans, 1974, p. 43) can be seen as informal description of melioration learning:

> "In choosing between alternative actions, a person will choose that one for which, as perceived by him at the time, the value, $[V^H(e)]$, of the result, multiplied by the probability, $[p^H(e)]$, of getting the result, is the greater."

Whether this proposition is equal to melioration depends on the meaning of the phrase "as perceived by him at the time". In regard to the probability $p^H(e)$, Homans (1974, p. 44) clarifies: "[i]n case of actions repeated over time, one of the determinants of his perception will be the actual frequency with which the past action has been followed by the reward." Hence, the probability $p^H(e)$ of getting a result corresponds to the fraction of past choices of $e \in E$ that were followed by a reinforcement. Formally:

$$p^H(e) = \frac{S(e)}{K(e)}.$$

$K(e)$ is the frequency of choosing alternative e (correlating to the definition of $K_t(s, e)$ above, without the time index t and the state s), and $S(e)$ denotes the number of choices with a result ($S(e) \leq K(e)$).

Homans (1974) is less clear about the meaning of the value $V^H(e)$ of the result of an action. A possible assumption is that the value "as perceived by him at the time" is equal to the mean value of past rewards:

$$V^H(e) = \frac{1}{S(e)} \sum_{i=1}^{S(e)} y_i,$$

with $\sum_{i=1}^{S(e)} y_i$ standing for the sum of rewards that were received for action e. It is clear from the quotation that $V^H(e)$ includes only strictly positive rewards. Therefore, it is divided by $S(e)$ and not by $K(e)$. This distinguishes $V^H(e)$ from the local reinforcement rate $\frac{1}{K(e)} \sum_{i=1}^{S(e)} y_i$, which also includes the zero values of missing reinforcements. Homans (1974) captures this difference by the second factor $p^H(e)$, such that

$$\frac{1}{K(e)} \sum_{i=1}^{S(e)} y_i = \frac{1}{S(e)} \sum_{i=1}^{S(e)} y_i \cdot \frac{S(e)}{K(e)} = V^H(e) \cdot p^H(e).$$

It follows that the rationality proposition is equivalent to the process of melioration learning. Consequently, also the *success proposition* and the *value proposition* of Homans (1974, pp. 16,25) are entailed by the melioration process because they are implied by the rationality proposition. Furthermore, the *stimulus proposition* (Homans, 1974, pp. 22-23) is included by mapping different states (elements of $\mathcal{S}$) to different stimuli. It should also be noted that Herrnstein (1979) locates any changes in value, as described by Homans' (1974, p. 29) *deprivation-satiation proposition*, in the definition of the rewards (the stochastic process $\{R_t\}_{t=2}^{\infty}$).

5.3.2 The exploration rate

An important difference between the ideas of Vaughan and Herrnstein (1987) and algorithm 5.2.1 is the *exploration rate*. While it was introduced as a technical trick that enabled the convergence of the Q-values, the exploration rate is actually a realistic assumption about human behaviour. First, this is evident from the observation of undermatching, which can be explained by exploration (sections 2.2.1 and 2.3). Second, actors are generally assumed to make mistakes, misinterpret situations, and occasionally try different behaviour (e.g. Selten, 1975; Bendor, 1987). Multiple studies showed that the assumption of random perturbations in human behaviour has significant effects on social outcomes (Macy and Tsvetkova, 2015): it can lead to more efficiency, improve the predictability, and alter processes of cultural assimilation, ethnic segregation, and diffusion of innovations.

In particular, it was argued that success depends on fine-tuning the trade-off between exploration and exploitation (March, 1991). While no exploration undermines the adaptation to changes in the environment and compromises the long-term prosperity, high exploration misses out on the gains of emitting the currently best actions. It is, therefore, conceivable that humans maintain a small but effective level of exploration that balances environmental changes and helps to learn profitable behavioural regularities.

5.3.3 Empirical studies

Empirical evidence for melioration learning was found on different levels. On the neural level, studies confirmed the underlying logic of melioration learning. In particular, experiments with monkeys indicated that neural activities are in line with the assumptions of melioration learning: individuals mentally represent the subjective values of available alternatives and choose the most desirable action (Dorris and Glimcher, 2004; Sugrue et al., 2004).

Further proof of melioration learning stems from behavioural experiments. These results are less profound because other learning mechanisms can explain the same observations. But there are only a few studies on the neural level. Therefore, it is focused on behavioural experiments in the following.

Given a problem of distributed choice, not all elements of the matching law solution emerge from melioration learning. In the first example of table 3.1 (chapter 3), the melioration process with strictly positive exploration rate ($\varepsilon > 0$) results only in the distribution $\boldsymbol{p} = (1, 0)$. Since $\overline{v}_1(\boldsymbol{p}) > \overline{v}_2(\boldsymbol{p})$ for any point $\boldsymbol{p} = (p_1, p_2)$ with $p_1 > 0$, the Q-value of the second alternative is eventually lower than the Q-value of the first alternative, and, hence, the relative frequency p_2 approaches zero. Similarly, in the second example of table 3.1, the melioration process converges to the distribution $\boldsymbol{p} = (0.7, 0.3)$ because $\overline{v}_1(\boldsymbol{p}) > \overline{v}_2(\boldsymbol{p})$ for any point $\boldsymbol{p} = (p_1, p_2)$ with $0 < p_1 < 0.7$ and $\overline{v}_1(\boldsymbol{p}) < \overline{v}_2(\boldsymbol{p})$ in case of $0.7 < p_1 < 1$. The other two distributions of the matching law solution $(0, 1)$ and $(1, 0)$ are feasible only if the actor starts with one alternative and never tries the other one ($\varepsilon = 0$).

In both examples, the distribution that follows from melioration is optimal in comparison with the other elements of the matching law solution. Instead of melioration learning, one could assume that individuals always end up in the matching law distribution with the highest average value. But in some situations, melioration is in conflict with the optimal matching law distribution. An often studied example of this situation is seen in figure 5.2. This example is sometimes called "Harvard game" (Vaughan and Herrnstein, 1987). It describes an experiment in which the subject chooses between two alternatives (1 and 2). While the average value $\overline{v}_1$ of the first alternative is always three points higher than the average value $\overline{v}_2$ of the second alternative, both values decrease with the relative frequency p_1. The dotted line shows the total average value v.

Two distributions correspond to the matching law: the exclusive choice of the first alternative ($\boldsymbol{p} = (1, 0)$) and the exclusive choice of the second alternative ($\boldsymbol{p} = (0, 1)$). According to melioration learning, the first distribution is *more likely*

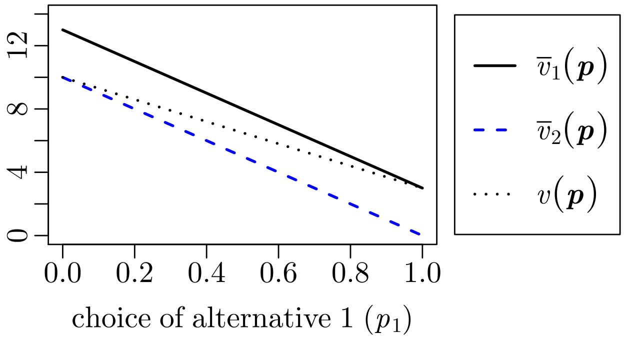


Figure 5.2: Illustration of the Harvard game

because an actor continuously increases the relative frequency of the alternative with the currently highest average value. A maximising individual, on the other hand, should consistently choose the second alternative. In Herrnstein (1997), various experiments that implement this or a similar procedure were reported. Most of these and other studies (e.g. Vaughan, 1981; Mazur, 1981) showed a clear tendency to a matching distribution that deviates from the optimum (see also Mazur, 2001, ch. 14.2). In the example of figure 5.2, this means that almost all subjects nearly exclusively chose the first alternative.

However, a majority of these experiments were conducted with pigeons. It is reasonable to suspect that humans are capable of maximising the overall reward (Herrnstein, 1997, p. 93). Depending on the complexity of the situation, humans may understand the implications of a decision or have information about the causal processes between choice distribution and reinforcement. This claim was tested in a series of experiments by Herrnstein et al. (1993). Similar to the experiment of figure 5.2, the amounts of reinforcement changed with the distribution of previous choices. Additionally, the experimental procedures differed in the recognisability of the relationship between the distribution of previous choices and current reinforcements. For example, the relative frequency of previous choices was indicated to some of the subjects, or a hint about the reinforcement mechanisms was given. The experiments showed that the easier it was for the subjects to discover the underlying mechanism, the closer the behaviour was to the optimum.

Furthermore, it was reasoned by Antonides and Maital (2002) that the relationship between recognisability and optimality depends on the cognitive skills of the subjects. Experiments revealed that insufficient information about the underlying mechanism of reinforcement is offset by the educational background of the subjects (students of economics vs. students of engineering).

The experiments of Herrnstein et al. (1993) also indicated a learning process that led from a suboptimal to an optimal distribution. Similar processes were found in experiments by Tunney and Shanks (2002). But in other studies, neither matching nor maximising was observed (Savastano and Fantino, 1994). Instead, the subjects randomised between the choice alternatives. The authors hypothesised that the subjects did not fully understand the cues that indicated the optimal choice (Savastano and Fantino, 1994, p. 459). A lack of understanding might also explain the finding that suboptimal matching was predominant if the previous choices influenced the delay (Herrnstein et al., 1993) or the probability of reinforcement (Tunney and Shanks, 2002; Neth et al., 2005) instead of its amount. Estimating the time delay or probability, instead of the amount, increases the cognitive demand and impedes the recognition of the optimal behaviour.

In the opinion of Herrnstein (1997, p. 205), the actual prediction of melioration learning depends on the capability of the individual to redefine the choice alternatives. If a subject is able to redefine the alternatives from pure actions ("action 1 vs. action 2") to distributions of actions ("90% action 1 vs. 80% action 1 vs. 70% action 1 vs. .. "), the long-term reward increases. Herrnstein (1997, p. 205) argued that a matching individual with unlimited capability of generalising the choice alternatives is a maximising individual. As indicated in the previously mentioned studies, different factors can limit this capability.

In summary, melioration learning predicts animal behaviour very well. In studies with humans, the results are mixed. Cognitive skills and the accessibility of information seem to be intervening factors. If choice alternatives can be redefined or if the underlying mechanisms are recognised, humans are able to optimise their behaviour. But in many situations, their behaviour is in line with melioration.

This finding has practical implications. For example, Yechiam et al. (2003) showed that behavioural modification is more successful if a training program accounts for melioration instead of optimisation. Given the former, the strengthening of certain behaviour depends on an appropriate reinforcement structure and not on the emphasize of anticipated future outcomes.

5.3.4 Comparison with other learning models

Brenner (2006) distinguished two main types of learning models: *reinforcement learning models* and *belief learning models* (a similar categorisation is found in Camerer, 2003, ch. 6). Melioration was categorised as reinforcement learning because it is less cognitive demanding than most belief learning models. However, melioration is also more sophisticated than simple reinforcement models, such as the Bush-Mosteller model. It is a form of reinforcement learning and still assumes that actors have beliefs about the consequences of different actions (Brenner, 2006, pp. 904-905). This seemingly contradictory description of melioration is explained in the following. Furthermore, problems of the present implementation of melioration and possible solutions are identified.

Belief learning models

In the algorithm of section 5.2.2, an agent learns the value $Q^*(s, e)$ of an action $e \in E$ in state $s \in \mathcal{S}$. In some way, this value constitutes a *belief* about the environment. Since the agent responds to these beliefs in an optimal way, melioration learning can be seen as a rudimentary form of belief learning.

But in other aspects, melioration learning is different from the general belief learning model. In most models, the formation of beliefs does not stop at the level of actions. On the contrary, the values of actions are usually externally given, and beliefs about the reinforcement mechanism or the behaviour of other agents are acquired. For example, in a two-person game-theoretic situation, the agents may know the structure of the game and learn the strategy of the opponent. A general belief learning algorithm for this situation is given by the following pseudocode (Shoham and Leyton-Brown, 2009, p. 196):

```
Initialize beliefs about the opponent's strategy
repeat:
    Play a best response to the beliefs
    Observe the opponent's actual choice and update beliefs accordingly
```

One example of belief learning is **fictitious play**: "in fictitious play, an agent believes that his opponent is playing the mixed strategy given by the empirical distribution of the opponent's previous actions" (Shoham and Leyton-Brown, 2009, p. 195). In other words, the agent remembers the decisions of the opponent, forms the corresponding relative frequencies, and chooses an action with the highest expected reward assuming that the relative frequencies resemble the opponent's probabilities of choice.

Fictitious play differs from melioration learning because the latter does not consider the behaviour of the opponent and ignores the expected future reward of an action. Instead, it is focused on the average rewards of past actions, and no mental model of the situation is built. An extensive model of the situation is also attained by **Bayesian learning**, which is another belief learning model (e.g. Young, 2004, ch. 7). With Bayesian learning, agents are even able to learn the transition probabilities of a Markov decision process (Vlassis et al., 2012).

Reinforcement learning models

In contrast to belief learning, reinforcement learning is a very simple idea about behavioural change. It can be summarised by Thorndike's *law of effect*: "pleasure stamps in, pain stamps out". More specifically, behaviour that is followed by a positive experience is likely to reoccur, but, if provoking negative reactions, it diminishes over time. Two examples of reinforcement learning are the Bush-Mosteller and the Roth-Erev model (Roth and Erev, 1995; Skyrms and Pemantle, 2000; Flache and Macy, 2002).

The **Bush-Mosteller model** was developed by Bush and Mosteller (1964) and states that a probability of choice changes linearly with the level of satisfaction. Given the notations of sections 5.1 and 5.2, an actor chooses an element $e \in E$ at time $t \in \mathbb{N}$ with probability $q_e(t) \in [0, 1]$. After receiving a reward $y_t \in \mathbb{R}$, the probability is updated by

$$q_e(t) = q_e(t-1) + \begin{cases} (1 - q_e(t-1)) \cdot \sigma(y_t) & \text{if } \sigma(y_t) \geq 0 \\ q_e(t-1) \cdot \sigma(y_t) & \text{if } \sigma(y_t) < 0 \end{cases} . \tag{5.7}$$

The function $\sigma(\cdot)$ expresses the level of satisfaction with the result y_t. Equation 5.7 can be found in a similar form in Macy and Flache (2002, p. 7231) or Izquierdo et al. (2007, p. 262).

The dynamics of Bush-Mosteller learning differ from the dynamics of melioration. This is seen when comparing the probabilities of choosing an action. Let $s \in \mathcal{S}$ be a state and $F_t(s) := \arg\max_{e' \in E} Q_t(s, e')$ denote the set of alternatives with the highest Q-value at time t. According to algorithm 5.2.1, the probability of choosing e at time t is:

$$Pr\big[X_t = e \mid S_t = s\big] = \frac{1 - \varepsilon_s}{|F_t(s)|} + \frac{\varepsilon_s}{|E|}. \tag{5.8}$$

The Bush-Mosteller model implies that $Pr\big[X_t = e \mid S_t = s\big] = q_e(t)$. In comparison with equation 5.8, behaviour that follows the Bush-Mosteller model changes gradually, for the probability of choice $q_e(t)$ instead of the set F_t is adjusted at each time step. Furthermore, the dynamics of equation 5.7 depend on the level of satisfaction with the outcome y_t. This is in line with the ideas of Simon (1955) about boundedly rational behaviour but requires additional assumptions. In the past, the level of satisfaction was implemented by comparing the actual outcome to an *aspiration level* (e.g. Macy and Flache, 2002). This aspiration level is a key factor and significantly affects the long-term behaviour (Macy, 1991; Macy and Flache, 2002). Moreover, Bendor et al. (2007) showed that, if aspiration levels are exogenously fixed or respond to past experiences, any outcome of a class of iterated games is stable. Consequently, not only the dynamics but also the stable states of the Bush-Mosteller and the melioration model are different.

The **Roth-Erev model** describes another process of reinforcement learning. Given any state $s \in \mathcal{S}$, its basic form (Roth and Erev, 1995, p. 172) specifies the probabilities of choice by the fraction of accumulated values $K_t(s, e) \cdot Q_t(s, e)$:

$$q_e(t) := \frac{K_t(s, e) \cdot Q_t(s, e)}{\sum_{j \in E} K_t(s, j) \cdot Q_t(s, j)}. \tag{5.9}$$

The right side of equation (5.9) resembles the right side of the matching law (equation (3.7)). It may have been this coincidence that led to the association of the matching law with the Roth-Erev model (Erev and Roth, 1998, p. 861, and Skyrms, 2010, p. 12). But when comparing the left-sided terms, equations (3.7) and (5.9) specify different quantities. In case of the matching law, it is the relative frequency of choosing an alternative over a period of time. The Roth-Erev model, on the other hand, describes the probability of choice at a particular point in time. The relative frequencies approach the probabilities in the long run if the probabilities converge. Given a Markov decision process and Q-learning, this holds true. Hence, in combination with Q-learning, the Roth-Erev model leads to the matching law and constitutes an alternative to melioration. Nevertheless, the dynamics of both models are different, which is evident from equations 5.8 and 5.9. An exemplary comparison of melioration and the Roth-Erev model in multi-agent situations is presented in chapter 6.

The Bush-Mosteller and the Roth-Erev model are just two of many models of reinforcement learning. There is a whole research field in the computer sciences that engages in the development and analysis of reinforcement learning (RL) (Sutton and Barto, 1998). Algorithm 5.2.1 describes a relatively trivial instance of the Q-learning method, which is, in turn, just one of several RL methods. While computer scientists are mainly interested in the design of autonomous software that is able to control its environment in a meaningful way, some of the work on RL is concerned with the accurate representation of human learning in situations of sequential decision-making.

As pointed out in the previous section, melioration learning accounts for observed behaviour and neural activity in situations of repeated choice (see also Sakai et al., 2006, p. 1092). But generally, there is “tremendous heterogeneity in reports on human operant learning” (Shteingart and Loewenstein, 2014, p. 94). In particular, melioration was often regarded as too simple to accurately represent the complexity of human decision-making (Barto et al., 1990, p. 593). For example, experiments with animals revealed that changes in behaviour occur rapidly with changes in the rates of reinforcement (Gallistel et al., 2001; Sugrue et al., 2004). This indicates that subjects maintain a *temporally local* representations of reinforcement rates, which include only the most recent outcomes. Corrado et al. (2005) presented a dynamic model of choice that can account for rapid changes. Their **linear-nonlinear-Poisson model** “differs from melioration with respect to both the quantity that drives behavioural change and the temporal window over which that quantity is computed” (Corrado et al., 2005, p. 611).

Other models of reward-driven behaviour were listed in Sakai et al. (2006). All of them exhibit the matching law in their steady states. For example, Sakai and

Fukai (2008a) argued for the usage of **actor-critic learning** because it exhibits the matching law in steady states and no direct representation of the average values (Q-values) is needed. Further advantages of this model are, first, that it requires minimal computation and, second, that it has useful properties in competitive and non-Markov cases (Sutton and Barto, 1998, p. 153).

A somewhat different decision rule was proposed by McDowell (2013b). The author modelled learning as an **evolutionary process within the individual**. A population of different behavioural alternatives was assumed, and successful behaviour was selected and "reproduced". The relative number of an alternative in this population resembled the probability of choice. In a series of simulations, McDowell and colleagues could show that this model leads to matching behaviour (McDowell, 2004; McDowell and Caron, 2007; McDowell et al., 2008; McDowell and Popa, 2010). Also behaviour that deviates from the matching law was reproduced if this behaviour had been observed in laboratory experiments.

In summary, the empirical status of melioration learning is disputed and alternative models of learning have been suggested. Because most of the mentioned learning models entail the matching law in their steady states, there is no theoretical reason to prefer one model over the others. If the most realistic representation is wanted, a **neural network model** (e.g. Loewenstein and Seung, 2006) is most appropriate because it is closer to the physical basis of human decision-making.

Any decision about the *best* model should be made by assessing its empirical adequacy. Since the long-term behaviour is similar, empirical findings of the dynamics (transient effects and temporal structures) must be used to distinguish between different models (Sakai and Fukai, 2008b, p. 241). First steps in this direction have already been made (Gallistel et al., 2001; Sugrue et al., 2004; Gureckis and Love, 2009). But more research is needed.

Even if melioration learning turns out to be too simple, it may serve as starting point of further investigations. Instead of using another learning model, the melioration algorithm of section 5.2.2 can be adjusted in order to account for empirical results. For example, a rapid change in behaviour is facilitated if, instead of all previous encounters, a smaller time window is used to calculate the Q-values. Alternatively, the speed of learning is increased by adding *eligibility traces* (Sutton and Barto, 1998, ch. 7).

A rapid change in behaviour takes also place if a new state of the environment is recognised. But this involves another problem of the melioration algorithm: the discrimination of environmental states. Similar to the value of actions, the state of the world should not be given to the agents. The agents must learn which environmental aspects are relevant for the reinforcement mechanisms and whether two stimuli indicate the same state or two different states. Thus, a **theory of stimulus discrimination** is needed to complement the reinforcement learning process (Shteingart and Loewenstein, 2014). Some approaches have already been studied (e.g. Sakai et al., 2006). There are also extensions of RL techniques that can deal with continuous state and action sets (van Hasselt, 2012) or with limited information about the current state of the environment (Spaan, 2012).

Q-learning, like other reinforcement learning methods, is goal-dependent. The Q-values are learned with respect to particular preferences, which are expressed by the subjective values of the results. If the preferences change, new Q-values must be learned. In contrast, agents can use their experiences to build a mental representation of the environment. Subsequently, they are able to update behaviour in the pursue of new goals. For example, **associative learning** denotes a process in which the agents learn associations between states of the environment. There is some progress in the integration of associative learning techniques in models of reinforcement learning (Alonso and Mondragón, 2006; Veksler et al., 2014). Similarly, agents can be designed to learn the transition and reward functions of a Markov decision process (Sutton and Barto, 1998, ch. 9; Hester and Stone, 2012).

With these extensions, the simple ideas of reinforcement learning are abandoned. The presence of a mental image of the environment and the processing of this image are properties of belief learning models. It can be speculated that, in the end, a *good* model of human learning should integrate elements of both reinforcement and belief-based models (see also Camerer and Ho, 1999).

Regret matching

Regret matching is another learning model that is found in economic literature (e.g. Young, 2004, ch. 2). In comparison with most reinforcement learning models, it takes *counterfactual* reinforcements into account. Let $\{R_t\}_{t=2}^{\infty}$ be the stochastic process of actually obtained rewards and let $\{R_t(e)\}_{t=2}^{\infty}$ denote the stochastic process of partly counterfactual rewards that would be obtained if action $e \in E$ is chosen at time t. The regret from not having chosen an action $e \in E$ is defined by

$$Z_t(e) := \frac{1}{t-1} \sum_{i=2}^{t} R_i(e) - R_i.$$

Regret matching specifies the probability of choosing action e at time t by

$$q_e(t) := \frac{Z_t(e)^+}{\sum_{j \in E} Z_t(j)^+}.$$

The +-sign indicates the non-negative part of the regret. It was shown that regret matching ensures no regret in the limit of t (Hart and Mas-Colell, 2000):

$$\limsup_{t \to \infty} Z_t(e) \leq 0 \text{ for all } e \in E.$$

Also without the knowledge of the counterfactual rewards $\{R_t(e)\}_{t=2}^{\infty}$, regret can be eliminated (Young, 2004, pp. 22-24, referring to Foster and Vohra, 1993). Similar to algorithm 5.2.1, this requires a stochastically independent exploration of the consequences of different actions.

Regret matching constitutes an alternative to melioration learning because, first, it seems that any state without regret implies the matching law. Second, regret matching has been shown to eliminate regret irrespective of the dynamics of the environment. This is an advantage over melioration learning, which needs some kind of stability (e.g. a Markov decision process). A formal analysis of both claims will be the subject of future work.

5.4 Conclusion

Melioration learning was categorised as reinforcement learning with rudimentary beliefs about the environment. Similar to other learning models, the matching law is a stable state of melioration. Due to its simplicity, several problems of the melioration algorithm exist and were outlined in the previous section. This chapter concludes by listing several advantages of the advocated model of melioration.

1. Algorithm 5.2.1 was theoretically derived from the work of Vaughan and Herrnstein (1987). In contrast to other models of melioration, it is closer to its original formulation (equation (5.1)), which stated that the relative frequencies, and not the probabilities of choice, change in accordance with the differences in reinforcement rates.

2. The model omits probabilities of choice. Therefore, it is not necessary to assume that humans behave stochastically (apart from the exploration rate).

3. No neural network is needed, which simplifies its implementation.

4. Q-learning is one of the most popular and widely studied RL techniques. Many results about its convergence properties already exist and can be appropriated for an application in social theory.

5. Melioration learning is connected to past sociological theory because it resembles the rationality proposition of Homans (1974).

6. Q-learning is one of the simplest RL methods. In the following chapters, social implications are derived by agent-based simulations. Simplicity is a desired property of this kind of simulation models (Axelrod, 1997, p. 18). The rules of decision-making should be kept simple in order to ease the understanding of the results and to reduce the time of computation.

7. Although experiments revealed deviations from the melioration model on the individual level, its predictions might be sufficiently accurate on a social level. Only if deviations are observed on the social level, the introduction of more advanced behavioural assumptions is justified.

8. The parameter γ of Q-learning supplies a convenient way to switch between optimal and matching behaviour. Melioration ($\gamma = 0$) implies *myopic* behaviour because it considers only the short-term effects of an action. The adjustment to $\gamma > 0$ accounts for long-term effects and guarantees optimal behaviour in the long run.

Chapter 6

Two-person melioration learning

Whereas the previous chapters largely referred to individual decision-making, this chapter finally deals with the explanation of social phenomena by the matching law. The simplest social situation is the interaction between two persons. In past sociological research, these situations were analysed, for example, by means of game theory. In this regard, a situation was represented as a game between two players. Subsequently, particular game-theoretic solutions were used to derive predictions about the players' behaviour.

Based on this approach, the melioration algorithm of the previous chapter is applied to various two-person games that have been studied in the sociological literature. It is tested whether melioration leads to similar predictions as game theory, for instance, to dominant actions, a Nash equilibrium, or a maximin outcome. If this is the case, the game-theoretic concepts are justified by a learning model that makes less strict assumptions about available information and the players' cognitive skills. Additionally, in games with multiple equilibria, the problem of equilibrium selection is solved.

However, the melioration algorithm is not guaranteed to approach a game-theoretic solution or even to provide a prediction at all. In the previous chapter, the convergence of melioration learning was based on a certain kind of *stationarity* of the situation. This stationarity is unlikely to be found in social settings in which multiple persons interact and reinforcements are contingent upon the decisions of everyone. Consequently, the behaviour of the actors may not converge.

Existing results about the convergence of Q-learning in multi-agent reinforcement situations have been mixed (Nowé et al., 2012, p. 451). While equilibria were reached in some instances of the prisoner's dilemma or the coordination game (Sandholm and Crites, 1995; Claus and Boutilier, 1998; Gomes and Kowalczyk, 2009), the behaviour failed to converge in others. The results depended on the reward structure of the situation as well as the method of decision-making. For example, contingent upon the particular instance of the prisoner's dilemma, the ε-greedy version of Q-learning may or may not converge (Wunder et al., 2010, theorem 6). On the other hand, the behaviour is guaranteed to converge if another strategy of exploration is assumed (Kianercy and Galstyan, 2012, p. 7).

In the following, only particular examples of two-person situations are explored by computer simulations. Some authors were able to analytically investigate the whole parameter space of two-person two-action situations. But slightly different implementations of Q-learning were used. More specifically, Kianercy and Galstyan (2012) applied the Boltzmann strategy instead of ε-greedy. Wunder et al. (2010) simplified the ε-greedy version by assuming a continuous time process with *infinitesimal* Q-learning, which meant that $\alpha_t(s, e) \to 0$ at all time $t \in (0, \infty)$ (compared to $\alpha_t(s, e) \to 0$ as $t \to \infty$; see section 5.2.2). In the case of algorithm 5.2.1, computer simulations are more convenient.

A connection to the previous literature is established by comparing melioration to the learning model of Roth and Erev (1995). It is focused on the Roth-Erev model because it is similar to melioration. Both models take a "mechanistic perspective on learning", which means that "[p]eople are assumed to learn according to fixed mechanisms or routines" (Brenner, 2006, p. 903). They were both categorised as reinforcement learning and have been connected to the matching law (section 5.3.4). Most learning models that were mentioned by Fudenberg and Levine (1998) or Young (2004) differ strongly from melioration. In their most common forms, those models make different assumptions about the available information (section 5.3.4). This impedes a direct comparison with melioration because the models cannot be applied to the same situation. It is possible to modify either side in order to account for more or less information about the environment. But this touches another broad research area (Sutton and Barto, 1998, ch. 9; Nowé et al., 2012) and will be subject of future work.

Other models of reinforcement learning, such as Bush-Mosteller (section 5.3.4) or experience-weighted attraction (Camerer and Ho, 1999), require additional assumptions and the specification of further parameters. In case of melioration and Roth-Erev, simple versions with only one parameter (the exploration rate) exist. Furthermore, learning by imitation (Nowak and May, 1992) and evolutionary algorithms (Maynard Smith, 1976) are less eligible for a comparative study because interactions with more than one partner are needed. The following simulations of two-person games can be adapted to an evolutionary context (e.g. Axelrod, 1987) or to a spatial game with imitation (Nowak and May, 1992). While this is also a promising area of future research, baseline results about simple two-person games are acquired in this chapter.

6.1 The simulation framework

In order to analyse actors who interactively learn by melioration, an extension of NetLogo was written. NetLogo (Wilensky, 1999) is a programming environment for the development of agent-based simulations. There are several reasons for the usage of NetLogo as simulation platform:

1. NetLogo is free, open source, extendible, and well documented.
2. NetLogo is a convenient platform to perform experiments with simulation models. It is possible to set parameters interactively, to monitor the behaviour of agents, and to plot and export results. The platform also contains a tool for the systematic variation of model parameters.
3. With a graphical user interface and, perspectively, a Web front end, NetLogo provides easy access to the simulations.
4. The language of NetLogo is relatively easy to learn and use, for it is based on Logo - "an educational programming language, originally designed to train youngsters" (Bianchi and Squazzoni, 2015, p. 301).
5. NetLogo is currently the most popular platform for agent-based modelling (Bianchi and Squazzoni, 2015, p. 301).

The first three points refer to practical advantages that are partly found in other simulation platforms as well (Railsback et al., 2006; Gilbert, 2008; Lytinen and Railsback, 2012). The last two points are relevant because simulation results are prone to stem from mistakes in the model or implementation (Gilbert, 2008, pp. 38-44). The usage of simulations in scientific research requires the possibility of understanding the code and replicating the results. Both operations are facilitated by using a widely known and easily accessible simulation language.

Nevertheless, NetLogo has disadvantages in case of advanced simulations. First, the execution of complex calculations and simulations with many agents may run slower on NetLogo than on other simulation platforms (Railsback et al., 2006, p. 619). For instance, the platforms MASON and Repast are more suited for advanced simulations because they are able to distribute computations among several processors. Second, the simplified programming environment of NetLogo forces the user to write all code in a single file. It is difficult to apply common programming techniques, such as modularity and reusability of code.

Both disadvantages of NetLogo can be reduced by employing its extensions API and its controlling API (application programming interface). The extensions API allows the implementation of new procedures in Java or Scala. These procedures are, subsequently, available in NetLogo. With the controlling API, a NetLogo model can be started and controlled from another program.

The following simulations make use of both APIs in order to parallelise the calculations, employ existing Java-libraries, and organise reusable code. The software tool, which was written for the simulations, is called *ql-extension*. Since it is an extension of NetLogo, it cannot run without it. The name emphasises that melioration is a special case of Q-learning. It is possible to analyse other Q-learning algorithms with this NetLogo extension. The installation and usage of the ql-extension are comprehensively described in appendix A.

The design of the simulation framework as a NetLogo extension facilitates the research process in several ways. First, because of the advantages of NetLogo, the specification and execution of simulation models are greatly simplified. Second, the usage of the extensions API allows the separation of different parts of the model. More specifically, the decision-making algorithm (e.g. melioration learning) is the same in all of the following simulations and is sourced out to the extension. A researcher who uses the ql-extension does not need to worry about this part of the model. The definition of the situation, on the other hand, changes with the particular research question, and the researcher must be able to easily adjust it.

When employing the ql-extension, the main task of the researcher is the specification of two components of the situation: a group structure and a reward function. The details of implementing either component are given in appendix A. Basically, the *group structure* indicates who interacts with whom. It can be statically set at the beginning of the simulation or change dynamically during the simulation. The *reward function* takes the decisions of a group of agents and calculates their rewards. After the specification of both components, the simulation can be started. It runs in discrete time steps. At each time step, all agents choose one of the given alternatives, and the reward function is called for each group to obtain the outcome of the decisions.

6.2 The actor models

The ql-extension implements different algorithms of decision-making. In this chapter, two of them are compared: stateless melioration learning and the Roth-Erev model. Also the Boltzmann (softmax) version of Q-learning and the ε-greedy version with decay in exploration are implemented (details are given in appendix A.1). As already mentioned in the introduction, Boltzmann exploration entails different results than ε-greedy. If the decay in exploration is enabled, either implementation of Q-learning converges to greedy behaviour. A rudimentary comparison of Boltzmann and ε-greedy is included in appendix B.1.

6.2.1 Stateless melioration learning

Algorithm 6.2.1 contains a description of stateless melioration learning. The main differences between algorithm 6.2.1 and algorithm 5.2.1 (previous chapter) are that, first, no states of the environment are distinguished and that, second, the exploration rate does note decay over time.

The constant exploration rate permits the acquisition of sufficient experience with the situation. Even if algorithm 5.2.1 guarantees that each alternative is

Algorithm 6.2.1 The stateless melioration learning algorithm

```
Require: exploration rate $\varepsilon \in (0,1)$, set of alternatives $E$
t ← 0
initialise $Q_1(e) \leftarrow 0$, for all $e \in E$
initialise $K_1(e) \leftarrow 0$, for all $e \in E$
repeat
    $t \leftarrow t + 1$
    if $\varepsilon >$ random number between 0 and 1 (uniform distribution) then
        chose a random action $e \in E$ using a uniform distribution
    else
        choose action $e \in E$ greedily using the Q-values
    end if
    observe reward $y$
    $K_{t+1}(e) \leftarrow K_t(e) + 1$
    $Q_{t+1}(e) \leftarrow Q_t(e) + \frac{1}{K_{t+1}(e)} \cdot (y - Q_t(e))$
    for all $e' \neq e$ do
        $K_{t+1}(e') \leftarrow K_t(e')$
        $Q_{t+1}(e') \leftarrow Q_t(e')$
    end for
until termination
```

chosen infinitely often, this property is impeded by the finite nature of every simulation. Keeping the exploration rate high supports the convergence of the Q-values to the expected rewards. However, this also means that algorithm 6.2.1 does not converge to greedy behaviour in the long run and that proposition 5.6 does not hold for algorithm 6.2.1.

Hence, the matching law must be evaluated empirically. This is done by the *standard deviation of normalised Q-values after accounting for the exploration rate.* The normalised Q-value of alternative $e \in E$ is defined by

$$\hat{Q}_t(e) := \frac{Q_t(e)}{\max_{j \in E} |Q_t(j)|}.$$

With $m := |E|$ being the number of alternatives, the following statistic is used to account for the exploration rate:

$$S_t(e) := \sqrt{\frac{mt}{\varepsilon(1 - \frac{\varepsilon}{m})}} \left(\frac{K_t(e)}{t} - \frac{\varepsilon}{m} \right).$$

If an alternative $e \in E$ was chosen only because of the exploration rate, the probability distribution of this statistic is approximately standard normal (see e.g. Collett, 1999, p. 20), and only 1% of the observations are greater than 2.33. Conversely, if this statistic is strictly greater than 2.33, it is concluded that the alternative was chosen deliberately, and the corresponding Q-value is included into the calculation of the standard deviation. More specifically, the compliance of an agent's behaviour with the matching law is measured by the NetLogo implementation of the sample standard deviation `sd` (`http://ccl.northwestern.edu/netlogo/docs/dict/standard-deviation.html`):

$$d_{ML} := \texttt{sd}\left\{\hat{Q}(e) \mid S(e) > 2.33\right\}.$$

If $d_{ML} < 0.05$, it is said that the matching law holds empirically.

6.2.2 The Roth-Erev model

Roth and Erev (1995) introduced a simple reinforcement learning model that did "a surprisingly good job of reproducing the major features of the experimental data" (Roth and Erev, 1995, p. 165). Subsequent studies confirmed this conclusion, especially in comparison to predictions of the Nash equilibrium (Ochs, 1995; Erev and Roth, 1998; Slonim and Roth, 1998; Feltovich, 2000).

Algorithm 6.2.2 specifies the Roth-Erev learning algorithm. Similar to algorithm 6.2.1, the actor holds a table of Q-values $\{Q_t(e)\}_{e\in E}$ that reflect the previous experiences with the alternatives. In Roth and Erev (1995, p. 172), the Q-values are called *propensities*. At each time step, an alternative $e \in E$ is chosen with probability $\frac{Q_t(e)}{\sum_{e'\in E} Q_t(e')}$. The parameter ε maintains some level of exploration.

There are two small differences between algorithm 6.2.2 and the original model of Roth and Erev (1995). First, *gradual forgetting* is not considered because the melioration algorithm omits this feature as well. Second, the exploration quantity $\frac{\varepsilon}{|E|-1} \cdot y$ is not added to just the "adjacent" but to all alternatives. The latter was also done in Erev and Roth (1998, p. 863) in case of two-action games or if there is no apparent linear order of the alternatives. In one example of the following sections, a linear order is apparent. It is, nevertheless, not taken into account, for it would distort the comparability with melioration learning.

Algorithm 6.2.2 The Roth-Erev learning algorithm

```
Require: exploration rate $\varepsilon \in (0,1)$, set of alternatives $E$
$t \leftarrow 0$
initialise $Q_1(e) \leftarrow 1$, for all $e \in E$
repeat
    $t \leftarrow t + 1$
    choose action $e \in E$ stochastically using the probabilities $\left\{ \frac{Q_t(e)}{\sum_{e' \in E} Q_t(e')} \right\}_{e \in E}$
    observe reward $y$
    $Q_{t+1}(e) \leftarrow Q_t(e) + (1 - \varepsilon) \cdot y$
    for all $e' \neq e$ do
        $Q_{t+1}(e') \leftarrow Q_t(e') + \frac{\varepsilon}{|E|-1} \cdot y$
    end for
until termination
```

6.3 Two-person games

In the remaining of this chapter, algorithms 6.2.1 and 6.2.2 are analysed in the context of different two-person games. It is tested whether the aggregated choices correspond to the matching law and which outcomes of the game occur. Most of the two-person games are known from economic literature. All games are presented in normal-form. The two players are labelled by "x" and "y". Capitalised letters or integers depict the alternatives. Further assumptions are listed in the following:

- For each game, a simulation with 20 000 pairs of agents is run. Every agent interacts with the same partner during the whole simulation.
- Half of the pairs of agents employ algorithm 6.2.1 (melioration learning). The other half uses algorithm 6.2.2 (Roth-Erev).
- Every agent repeatedly chooses one of the alternatives according to algorithm 6.2.1 or 6.2.2 until 1 000 choices have been made.
- Agents observe the set of alternatives and the reward of each choice. They are not aware of the structure of the game or the partner's choices.
- A payoff matrix shows the mean rewards. The actual reward is drawn from a normal distribution with a standard deviation of one.
- The exploration rate ε is set to 0.1 and does not decrease with time (simulations with different settings are presented in appendix B).

Statistical tests are omitted in the comparison of the two algorithms. Even if one excepts the quasi-randomness, which is simulated by the computer, as sufficiently close to real randomness, the application of statistical tests is largely unnecessary. Since there are 10 000 pairs of agents for each algorithm, any standard test would mark a difference as low as 150 pairs as statistically significant. For example, in the histogram of figure 6.1, the first two bars at (A, A) show a difference of 178 pairs. It remains a task to the reader to decide whether the reported differences in numbers are theoretical or practical *significant*.

In the following, three classes of two-person games are distinguished. The first class contains games in which at least one of the players has a (weakly) dominant alternative. Second, games without dominant alternatives but with several pure Nash equilibria are considered. The last class covers games with exactly one mixed Nash equilibrium. This division is not exhaustive, but it clarifies important properties of two-person melioration learning.

6.3.1 Games with dominant alternatives

An alternative of one player is said to be *dominant* if the choice of this alternative comes with a mean reward that is strictly greater than the mean reward of any other alternative given one choice of the partner and greater than or equal to the mean reward of any other alternative given the other choices of the partner (cf. *weak dominance* in Shoham and Leyton-Brown, 2009, p. 77). A representative example of this first class of games is the *prisoner's dilemma*. This game describes a situation in which the individually preferred outcome is the socially least desired one. In the example of figure 6.1, alternative B is dominant for both players. The outcome (B, B) is, therefore, a Nash equilibrium. All other outcomes are optimal.

In figure 6.1, the frequency distribution of pairs of agents at the 1 000th round of the simulation is shown. It is distinguished between pairs of agents who learn by melioration (mel) and pairs of agents who use the Roth-Erev model (RE). Both types of agents predominantly choose the Nash equilibrium. In case of melioration learning, there is no mixing, and each agent chooses B exclusively. Because of the exploration rates, also the non-equilibrium outcomes (A, B) and (B, A) occur. The frequencies approximate the expected ones: $10\,000 \cdot \frac{\varepsilon}{2} \cdot (1 - \frac{\varepsilon}{2}) = 475$. Therefore,

		y	
		A	B
x	A	(5,5)	(0,10)
	B	(10,0)	(2,2)

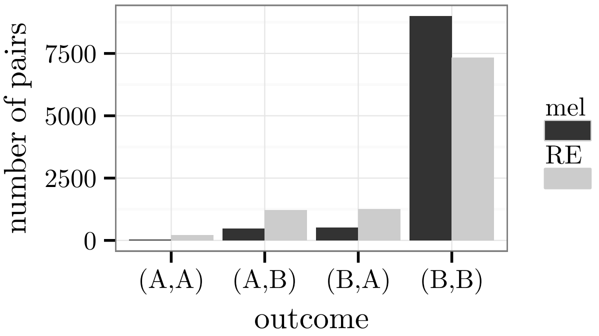


Figure 6.1: A prisoner's dilemma and simulation results

the measures d_{ML} are strictly less than 0.01, and the matching law holds for all agents who learn by melioration. Agents who use the Roth-Erev model show higher frequencies of non-equilibrium outcomes, and the matching law measures d_{ML} are greater than 0.1 for the majority of agents.

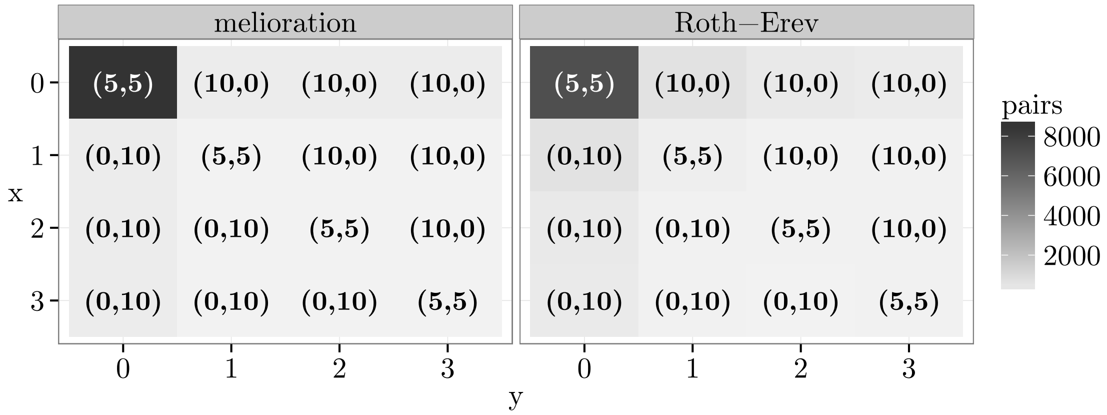


Figure 6.2: The game "guess $\frac{2}{3}$ of the average" and simulation results

Another example of a game with dominant alternative is called "guess $\frac{2}{3}$ of the average". Figure 6.2 contains a discrete version of this game with four alternatives. The rules state that each agent tries to guess $\frac{2}{3}$ of the average of the guesses. The agent who is closest to this value "wins" the game. In the particular example of figure 6.2, one can choose an integer between 0 and 3. The choice of alternative 0 is dominant. The reward table and the simulation results are displayed in the same plot by heat maps. The background colour of a cell is light grey if only few pairs of agents choose this outcome at the 1 000th round of the simulation. It is close to black if many pairs do so. The heat maps show that almost all agents learn to choose the dominant alternative 0, which constitutes the only Nash equilibrium. Similar to the prisoner's dilemma, agents who use the Roth-Erev model end up slightly more often in non-equilibrium outcomes.

While both actor models imply a tendency towards the dominant alternative, the results are less pronounced for Roth-Erev. This effect is more clearly seen in the game of figure 6.3. Alternative A is dominant for player x, and alternative B is dominant for player y. Hence, the outcome (A, B) is a Nash equilibrium. Additionally, (A, A) and (B, B) are Nash equilibria, which are not *payoff-dominated* by (A, B) because they involve the same mean rewards (Harsanyi and Selten, 1992, p. 81). The simulations reveal that all agents prefer the first equilibrium (B, A) instead of (A, A) and (B, B).

		y	
		A	B
x	A	(10,10)	(0,0)
	B	(10,10)	(10,10)

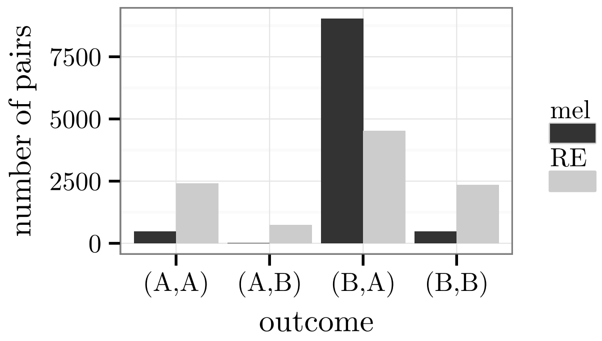


Figure 6.3: A game with three optimal Nash equilibria

In case of the melioration algorithm, the choice of the dominant alternative is due to the exploration rate. Exploration guarantees that the fourth outcome (A, B) is selected occasionally, especially at the beginning of the simulation. This means that, for player x, the average value of alternative A is between 0 and 10. The value of alternative B, on the other hand, is approximately 10. The reverse holds for player y, which leads to the combination (B, A) in rounds without exploration.

The choice of (B, A) is also most likely given the Roth-Erev model. But each agent maintains a relatively high probability of choosing the other alternative. This probability does not converge towards the exploration rate with further rounds of the simulation. Consequently, the agents of the Roth-Erev model significantly deviate from the matching law.

6.3.2 Games with multiple pure equilibria

A result of the previous section is that melioration learning leads to the choice of a dominant alternative even though the structure of the game is not known. A key factor is the exploration rate, which renders dominated alternatives inferior. In games without dominant alternative, this argument does not apply, and actors are not drawn to a single outcome. A Nash equilibrium is still guaranteed to exist (Nash, 1951). Games with a single Nash equilibrium are considered in the next section. In this section, games with more than one equilibrium are analysed.

A basic game with two or more Nash equilibria is the *coordination game*. It refers to a class of situations in which the agents prefer to coordinate their choices in some way. In the particular example of figure 6.4, the outcomes (A, A) and (B, B) are pure Nash equilibria, and (A, A) payoff-dominates (B, B) because of higher mean rewards (Harsanyi and Selten, 1992, p. 81). This game has an additional mixed equilibrium with probabilities $\left(A : \frac{4}{9}, B : \frac{5}{9}\right)$ for both players.

		y	
		A	B
x	A	(10,10)	(0,0)
	B	(0,0)	(8,8)

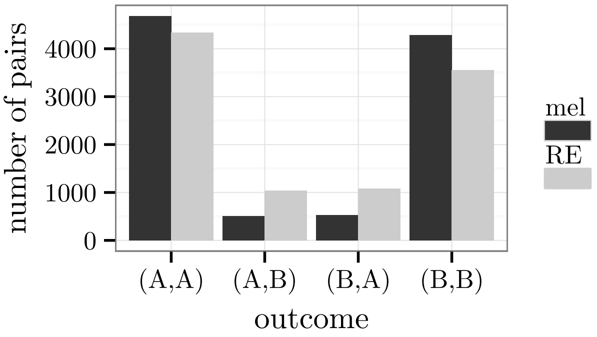


Figure 6.4: A coordination game and simulation results

In figure 6.4, the frequency distribution of pairs of agents at the 1 000th round of the simulation is shown. The agents choose mainly a pure Nash equilibrium and the payoff-dominant one with a slightly higher frequency. In other words, all pairs of agents are able to coordinate their choices. The deviations to (A, B) and (B, A) are due to the exploration rate and, similar to the previous simulations, more pronounced in case of the Roth-Erev model.

Further simulations reveal that the particular reward structure affects the distribution of agents among the two Nash equilibria. In particular, the frequency of

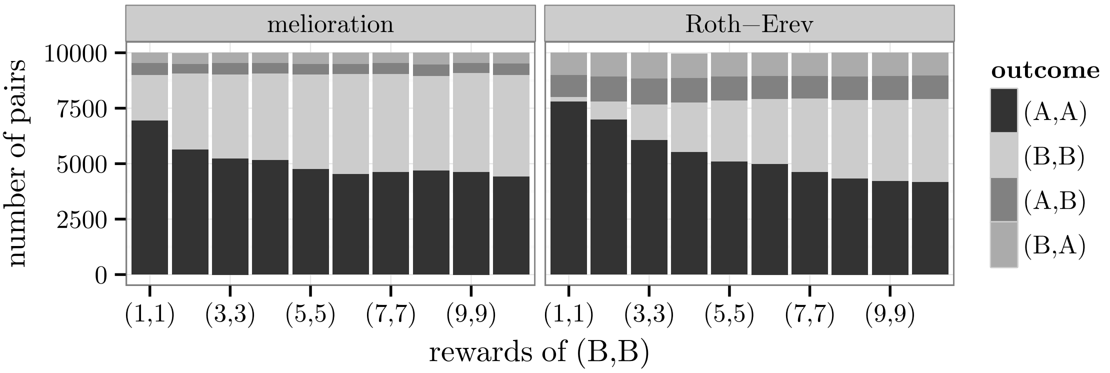


Figure 6.5: Relationship between the rewards of (B, B) and frequencies

the suboptimal equilibrium (B, B) depends on its expected rewards. As seen in figure 6.5, the higher the rewards, the higher this frequency.

The distribution of agents also changes with the rewards of the non-equilibrium outcomes (A, B) and (B, A). In the game of figure 6.6, these rewards are set by two parameters a and b. Depending on the difference $b - a$, the agents are more strongly drawn to either (A, A) or (B, B). If $a = 0$ and $b = 10$, almost all agents choose (B, B). The number of agents at (B, B) decreases with the difference $b - a$.

		y	
		A	B
x	A	(10,10)	(a,b)
	B	(b,a)	(8,8)

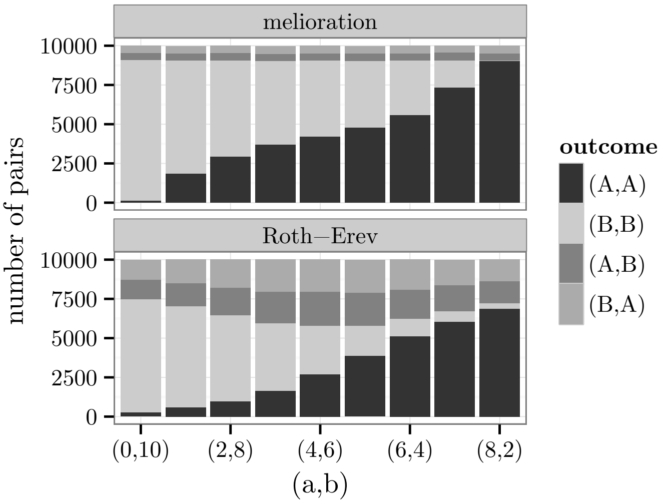


Figure 6.6: Relationship between non-equilibrium rewards and frequencies

This correlation can be explained by considering the melioration algorithm. The agents attach values $Q(A)$ and $Q(B)$ to the alternatives A and B regardless of the choice of the other agent. Because of the exploration rates, also the outcomes (A, B) and (B, A) are selected occasionally. This means that the value of action A increases with the reward a and that the value $Q(B)$ increases with b. Therefore, the tendency to choose (A, A) instead of (B, B) grows if either the reward a increases or the reward b decreases (or if both happens).

The dynamic of figure 6.6 reflects a preferences for the *maximin* alternative. An alternative is maximin if its choice leads to a maximum of all rewards that are minimal over the choices of the partner (Shoham and Leyton-Brown, 2009, p. 72). It is a low-risk strategy because, regardless of the decisions of the partner, another choice could be worse. If $a = 0$ and $b = 10$, B is the maximin alternative for both players. On the contrary, A is their maximin alternative if $a = 8$ and $b = 2$. In the case of $a = b = 5$, both alternatives are maximin. Because the latter implies an indifference between the alternatives, all four outcomes should occur with the same frequency. For the Roth-Erev model, this is approximately correct. But agents who learn by melioration still coordinate their actions, and slightly more agents end up in (A, A) (9.567 pairs) than in (B, B) (8.498 pairs).

battle of the sexes

		A	B
x	A	(5,10)	(0,0)
	B	(2,2)	(10,5)

game of chicken

		A	B
x	A	(5,5)	(2,10)
	B	(10,2)	(0,0)

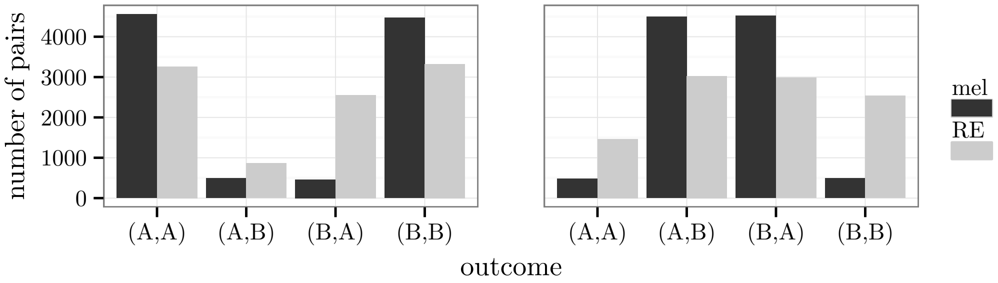


Figure 6.7: A “battle of the sexes” and a game of chicken

In conclusion, the melioration model is more successful in the coordination of actions than the Roth-Erev model. In case of the Roth-Erev model, non-equilibrium outcomes appear more frequently than predicted by the exploration rate. This is also apparent in the "battle of the sexes", which is a particular kind of coordination game. This game describes an interaction of two persons who plan to attend an event together (e.g. a concert or a cooking class). The partners have complementary preferences about two alternative events, but with an additional preference for attending the same one. A sample reward matrix is given by the left-sided table of figure 6.7. There are two pure and one mixed Nash equilibria: (A, A); (B, B); $\left(x : \left(A : \frac{3}{13}, B : \frac{10}{13}\right), y : \left(A : \frac{10}{13}, B : \frac{3}{13}\right)\right)$. Both pure equilibria are optimal. The outcome (B, A) consists of the maximin alternatives.

The simulations show that a pair of meliorating agents ends up in (A, A) or (B, B). Because of the symmetry of the game, there is no criterion that favours one of the two pure equilibria. Harsanyi (1977) calls this state *bargaining deadlock* between (A, A) and (B, B). While Harsanyi (1977, p. 279) suggests the third (mixed) equilibrium as solution of the game, melioration leads to an equal division of the pairs. If agents use the Roth-Erev model, also the suboptimal non-equilibrium outcome (B, A) appears frequently.

A similar effect arises in the *game of chicken* (right-sided table of figure 6.7). This game resembles a basic conflict between two parties that requires the retreat of at least one of them to be solved. In this case, agents who learn by melioration predominantly choose one of the two pure Nash equilibria: (A, B) or (B, A). The Roth-Erev model implies the regular choice of the worst outcome (B, B).

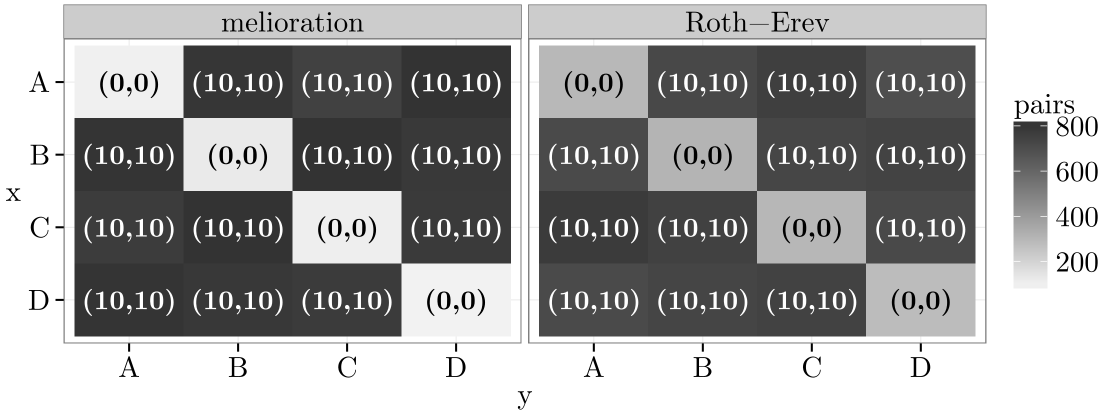


Figure 6.8: A dispersion game and simulation results

Finally, a game with more than two pure Nash equilibria is analysed. Figure 6.8 contains heat maps of a *dispersion game* with four alternatives. It is, in some respect, the opposite of a coordination game. Each agent prefers not to match the choice of the other agent. This means that all but the diagonal outcomes are optimal Nash equilibria. Consequently, most agents of the simulations are distributed evenly among the non-diagonal outcomes. Agents who use the Roth-Erev model end up slightly more often in non-equilibrium outcomes.

6.3.3 Games with a single Nash equilibrium

The games of this section have one strictly mixed Nash equilibrium. In contrast to games with pure equilibria, it takes a higher number of decisions until the behaviour of the agents has converged. Therefore, the following simulations are run with only 2 000 pairs of agents but for 20 000 rounds of the game. The relative frequencies of choice are calculated for the whole period of 20 000 rounds and for each agent separately. The convergence to the matching law is evaluated with respect to the whole period as well. Furthermore, a slightly higher exploration rate ($\varepsilon = 0.2$) is assumed because it supports the speed of convergence (the effect of different exploration rates is indicated in appendix B.1).

		y	
		A	B
x	A	(10,0)	(0,10)
	B	(0,10)	(10,0)

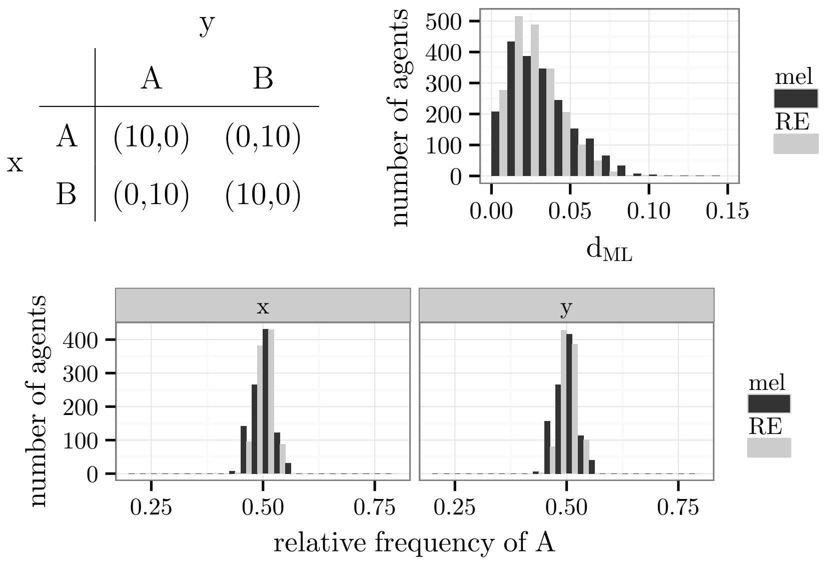


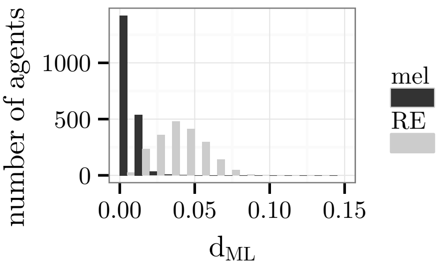
Figure 6.9: The game “matching pennies” and simulation results

First, zero-sum games are analysed. In these games, the gain of one of the players equals the loss of the other player. The reference point is not necessarily zero. For instance, in all of the following zero-sum games, the reference point is 5. This means that the default reward of any player is 5. If the outcome of an interaction is $(0, 10)$, the first player looses 5 while the second player gains 5. One example is the game "matching pennies" as shown in figure 6.9. The Nash equilibrium is given by the probabilities $(A : 0.5, B : 0.5)$ for both players.

According to figure 6.9, the behaviour of all agents converges to the matching law: the measures d_{ML} approximate zero for most of the agents. The values are even closer to zero if the simulation progresses or if a higher exploration rate is used (see appendix B.1). Figure 6.9 also contains histograms over the relative frequencies of alternative A. For both types of actors, the relative frequencies are in accordance with the probabilities of the mixed Nash equilibrium. The agents display a mix of the alternatives such that each one is chosen half of the time.

A similar result is obtained for the game "rock-paper-scissors", which is zero-sum with three alternatives per player. The game is specified in figure 6.10. The agents' behaviour approaches the matching law and the predictions of the mixed Nash equilibrium: $\left(A : \frac{1}{3}, B : \frac{1}{3}, C : \frac{1}{3}\right)$.

			y	
		A	B	C
	A	(5,5)	(10,0)	(0,10)
x	B	(0,10)	(5,5)	(10,0)
	C	(10,0)	(0,10)	(5,5)



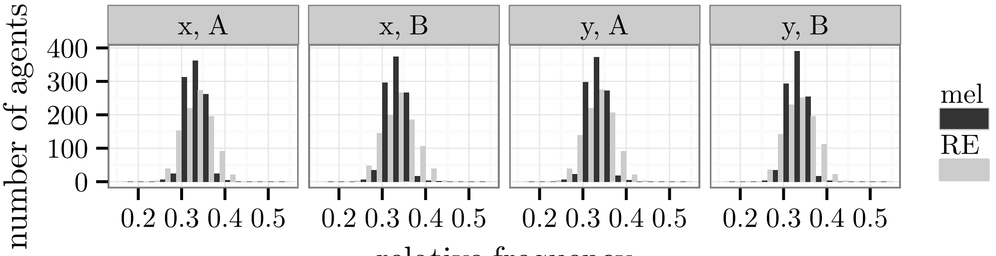
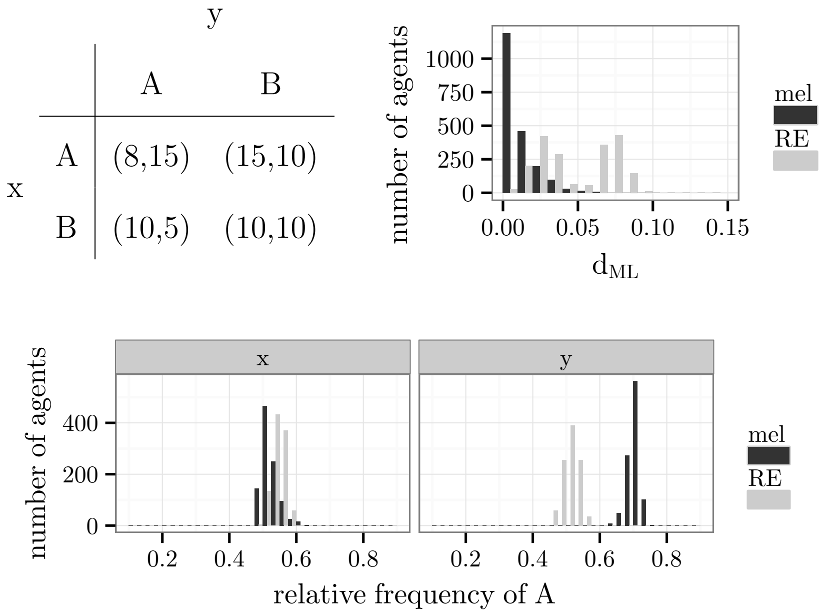


Figure 6.10: The game "rock-paper-scissors" and simulation results

A game that is not zero-sum is seen in figure 6.11. It has a single mixed Nash equilibrium at $\left(x : \left(A : \frac{1}{2}, B : \frac{1}{2}\right), y : \left(A : \frac{5}{7}, B : \frac{2}{7}\right)\right)$. In the past, this game was used to model the interaction between criminals and police (Tsebelis, 1990) and was, therefore, called *inspection game* (Rauhut, 2009). The criminal (player x) chooses between committing a crime (A) or not committing a crime (B). The police or inspector (player y) either inspects the suspect (A) or not (B). Committing a crime is beneficial only if no inspection takes place. An inspection is beneficial only if a crime occurs. The simulations show that agents who learn by melioration approach the matching law and the Nash equilibrium. The Roth-Erev model deviates from from both concepts.



		y	
		A	B
x	A	(8,15)	(15,10)
	B	(10,5)	(10,10)

Figure 6.11: An example of the inspection game and simulation results

Further simulations were run with different payoffs for player x given the outcome (A, A). This payoff refers to the punishment of a crime. The prediction of the Nash equilibrium for player x does not change with this reward, and the results of the simulations remain in line with the Nash equilibrium. Consequently, criminals who learn by melioration choose to commit a crime with a relative frequency of 0.5 regardless of the punishment.

Laboratory experiments with humans indicated that the level of punishment actually has an effect on the crime rate. More specifically, low punishment comes with a higher crime rate than high punishment (Rauhut, 2009). However, the experiments lasted for only 15 rounds of decision-making. If humans learn slowly, the behaviour may not have converged to a stable point yet. In figure 6.12, the temporal development of the relative frequencies of choosing A are shown for the two games of the experiment of Rauhut (2009). All agents use the melioration learning model, and the mean value of 1000 agents is plotted on a logarithmic scale of time. In case of low punishment (upper row), the Nash equilibrium (0.5) is approached from above. If punishment is high (lower row), the equilibrium is approached from below. Hence, there is a long period of time in which crime rates are higher for low punishment than for high punishment. Also the inspection rates conform qualitatively to the experimental results if it is focused on early rounds.

		y	
		A	B
x	A	(−1,5)	(5,0)
	B	(0,−5)	(0,0)

		A	B
x	A	(−20,5)	(5,0)
	B	(0,−5)	(0,0)

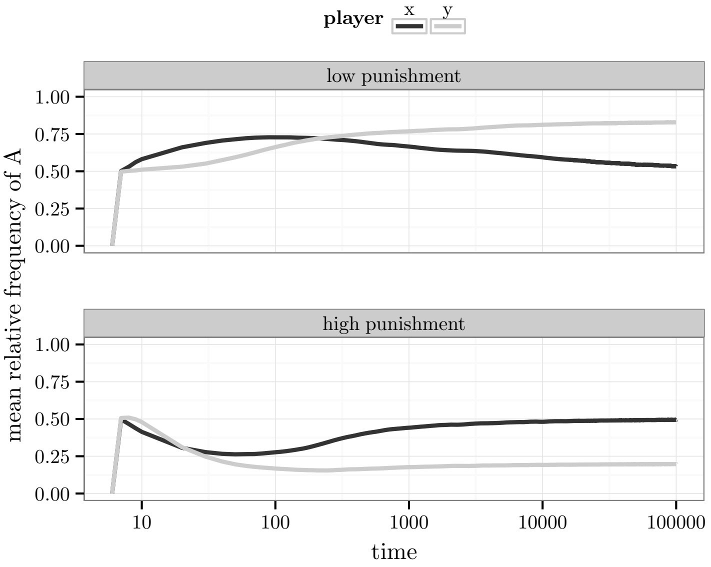
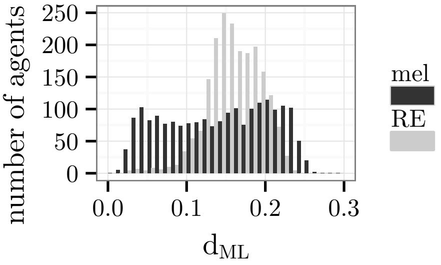


Figure 6.12: The inspection game with low or high punishment

Finally, there are some games in which the choices of agents who learn by melioration or the Roth-Erev model do not converge to a stable state. One example is presented in figure 6.13. This game is sometimes referred to as *Shapley's game* and known for its difficulties in regard to the convergence of learning algorithms (Abdallah and Lesser, 2006). It is similar to the game “rock-paper-scissors” except for the diagonal rewards, which are $(0,0)$ instead of $(5,5)$. The Nash equilibrium is given by $\left(A : \frac{1}{3}, B : \frac{1}{3}, C : \frac{1}{3}\right)$.

As seen in figure 6.13, the agents do not approach the matching law. The measures d_{ML} are significantly greater than zero after 20 000 choices and not converging towards zero if the simulation progresses. The lower plots of figure 6.13 depict the changes in relative frequencies of two particular players. If agents learn by melioration, the relative frequencies of all three alternatives rise and fall in sequence without any clear tendency towards convergence. This implies a constant change in outcomes: from (B,A) to (C,A) to (C,B) to (A,B) to (A,C) to (B,C) and back to (B,A). The time is on logarithmic scale, which means that the lengths of the waves increases with time. This happens because the Q-values are calculated as long-term averages. There is no decrease in the height of the waves, which could lead to a stable outcome. In case of the Roth-Erev model, the dynamic is slower, but no convergence is visible as well.

		y		
		A	B	C
	A	(0,0)	(10,0)	(0,10)
x	B	(0,10)	(0,0)	(10,0)
	C	(10,0)	(0,10)	(0,0)



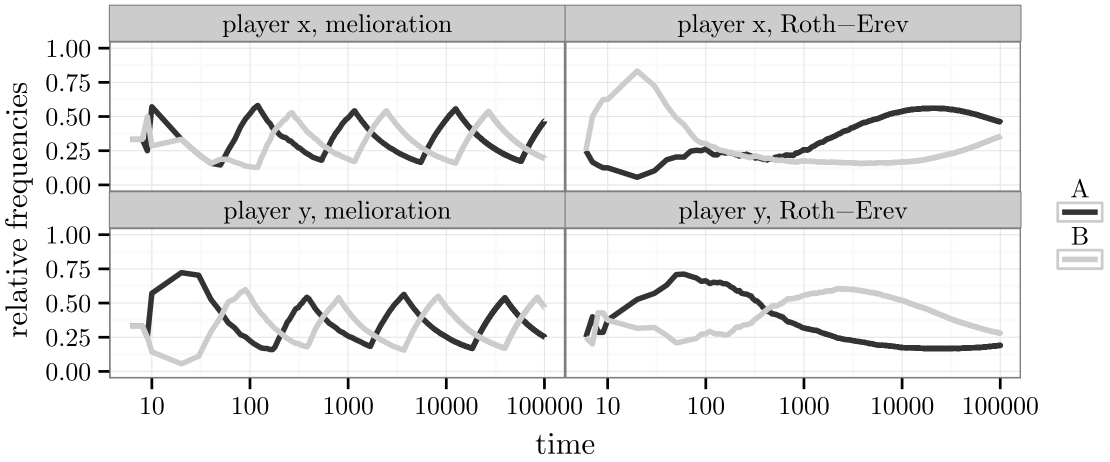


Figure 6.13: Shapley's game and simulation results

6.4 Conclusion

In this chapter, the melioration learning algorithm was analysed in the context of various two-person situations, for example the prisoner's dilemma, the coordination game, the game of chicken, and the inspection game. The results were largely in line with game-theoretical predictions. But less strict assumptions about available information and the agents' cognitive skills were required. Additionally, the problem of equilibrium selection was solved because the model specifies a distribution of outcomes for a given game.

The simulations indicated that agents who learn by melioration choose the dominant alternative of two-person games. If no alternative is dominant, mainly pure Nash equilibria occurred, and optimal ones were preferred but not chosen exclusively. More specifically, the structure of the game, which includes the rewards of non-equilibria, affected the distribution of outcomes.

If no pure equilibrium existed, the agents chose several alternatives with strictly positive probability. In some of the games, the long-term relative frequencies were in line with the matching law and corresponded to the mixed Nash equilibrium. But there are games that prevent the convergence of the agents' behaviour.

In case of the Roth-Erev model, the results were qualitatively similar. But many differences were discovered. For example, the probability of choosing a dominant alternative increased slowly, which led to deviations from the matching law. In coordination games, non-equilibrium outcomes appeared frequently, and even the worst outcome was chosen regularly in the game of chicken.

Chapter 7

Multi-person games

An advantage of the melioration learning algorithm is its applicability to a wide range of situations. Other authors have already analysed slightly different versions of Q-learning in the setting of two-person two-action games (Wunder et al., 2010; Kianercy and Galstyan, 2012). But there is no work about the convergence of Q-learning in games with more than two players. Because of the complexity of these situations, the convergence of a learning model to a dominant alternative or a pure Nash equilibria is more difficult than in two-person games. It is tested in the following sections whether melioration learning enables players of multi-person games to arrive at a Nash equilibrium or another steady state.

In section 7.1, n-way coordination games are investigated. Given one of these situations, a player simultaneously interacts with several partners in two-person coordination games. This model allows to explore the evolution of social conventions and institutions, for the actions of multiple persons must be synchronised in order to achieve an optimal result. The simulations reveal that, besides the reward structure of the game, also the network structure of interactions affects the group's ability to coordinate its members' choices.

In section 7.2, it is shown that the agents can learn to volunteer in the volunteer's dilemma. More specifically, the relative frequencies of volunteering are in line with the prediction of a pure or mixed Nash equilibrium. In the asymmetric version of the dilemma, results of the simulations are more intuitive than some of the game-theoretic predictions.

Finally, a multi-person prisoner's dilemma is explored. The actors sustain cooperation under favourable conditions. The exploration rate plays a decisive role in this game because incentives change if several players explore an action at the same time. Additionally, the option to punish defectors or to abstain from an interaction stabilises and even increases the level of cooperation.

7.1 N-way coordination games

N-way models refer to the repeated play of a two-person stage game with multiple partners (e.g. Macy, 1991, p. 826). Depending on the particular stage game, n-way models can be used to explain the evolution of conventions or social institutions. In the case of a two-person coordination game, a *convention* (or social institution) is said to be established if all members of a finite population agree on a pure Nash equilibrium (see e.g. Schelling, 1960; Young, 1998).

Since a coordination game has multiple pure equilibria, a stable state with convention may or may not emerge. Even if a convention is established, several outcomes are possible. Furthermore, a stable outcome may be *inefficient*, i.e. *payoff-dominated*, if its rewards are strictly smaller than the rewards of another equilibrium. Therefore, it is analysed which conditions enable the emergence of an *efficient* outcome, which means that there is no equilibrium point with strictly greater rewards (Harsanyi and Selten, 1992, p. 81).

7.1.1 Previous research

Young (1993) analysed n-way coordination games with random interactions. In this model, the players of each stage game were randomly drawn from a large population. Every actor knew the payoff structure of the stage game and considered a random sample from the set of previous interactions. The model did not necessarily describe a learning process. An agent could "ask around" instead of relying on own experiences (Young, 1993, p. 59). Decisions were made by selecting an alternative with the highest expected reward. The probabilities of a partner's decision were assumed to equal the corresponding relative frequencies in the random sample of previous interactions.

As result, Young (1993) found out that actors learn to play a *risk-dominant* equilibrium of the stage game. Risk-dominance is a theoretical property of an equilibrium that is not necessarily equivalent to payoff-dominance (Harsanyi and Selten, 1992, pp. 88-89). It takes into account that, if the decision of the partner is not known, the choice of one alternative may be less risky than the choice of another one (Harsanyi and Selten, 1992, pp. 82-84).

In a related study, Kandori et al. (1993) came to the same conclusion. The authors analysed the dynamics of n-way models by assuming a finite population and an evolutionary process that controlled the choices of its members. Given a certain class of coordination games, the unique stable state corresponded to a risk-dominant equilibrium of the stage game (Kandori et al., 1993, p. 46).

The models of Young (1993) and Kandori et al. (1993) are similar to melioration learning because the actors were assumed to be myopic, take past occurrences into account, and make random mistakes. But, in contrast to melioration, these actors knew the reward structure of the game and chose a "best-reply" to the given information (see also section 5.3.4).

In addition to interactions with random partners, Young (1998, ch. 6) considered n-way games in networks. In this case, every actor had a small group of potential partners, which were specified by a network. The information about previous encounters was limited to this group. Similar to the model with random interactions, only risk-dominant equilibria were stable states, and all members of a connected component of the network chose the same alternative. A connected component is a part of a network in which every member is reachable by every other member via a sequence of edges.

In other words, the particular structure of the network had no effect on the outcome of an n-way game. This result was due to random mistakes of the actors. If decision are deterministic, the structure of the network affects the likelihood of reaching an equilibrium (Buskens and Snijders, 2015). Even equilibria that are risk-dominated by another outcome may occur. This depends on the density, the centrality, and the segmentation of the network. Furthermore, it is possible that two different conventions coexist in some networks (Berninghaus and Schwalbe, 1996). This phenomenon, which is called *polarisation*, is even more likely if network connections can be endogenously changed by the actors (Buskens et al., 2008).

7.1.2 Results

It is tested whether conventions emerge in simulations with agents who learn by melioration (algorithm 6.2.1). Two different coordination games are analysed:

I		y	
		A	B
x	A	(10,10)	(0,0)
	B	(0,0)	(b,b)

II		y	
		A	B
x	A	(10,10)	(0,6)
	B	(6,0)	(b,b)

The parameter b is set to an element of $\{2, 4, 6, 8, 10\}$. Hence, (A, A) and (B, B) are pure Nash equilibria. In game I, the outcome (A, A) is efficient and risk-dominant as long as $b < 10$. If $b = 10$, both outcomes (A, A) and (B, B) are efficient, and there is no risk-dominance relationship between them. Also in game II, (A, A) is efficient. But it is risk-dominant only if $b < 4$. In case of $b = 4$, there is no risk-dominance relationship between (A, A) and (B, B). If $b > 4$, (B, B) risk-dominates (A, A), even in situations in which it is inefficient ($b < 10$).

It was shown in section 6.3.2 that, if these games are repeatedly played by the same two actors, they end up in one of the Nash equilibria (A,A) or (B,B). The latter outcome is observed even if it is payoff-dominated by the first one ($b < 10$). However, given a situation in which the players alternately interact with different partners, all agents should agree on a single alternative in order to avoid the inferior outcomes (A,B) and (B,A).

In each of the following simulations, 200 groups of agents are created. Every group consists of 50 agents and a network that specifies the structure of interactions. The vertices of the network represent the agents. An edge exists between two vertices if the corresponding agents repeatedly take part in the same coordination game. Since the games are symmetric, no differences between the vertices need to be made. The exploration rate is set to $\varepsilon = 0.1$.

To illustrate the effect of the network structure on the agents' behaviour, the small-world network (β-)model of Watts (1999, p. 67) is adopted. This model has two parameters: the average number of neighbours $\overline{d} \in \{2, 4, 6, \dots\}$ and the

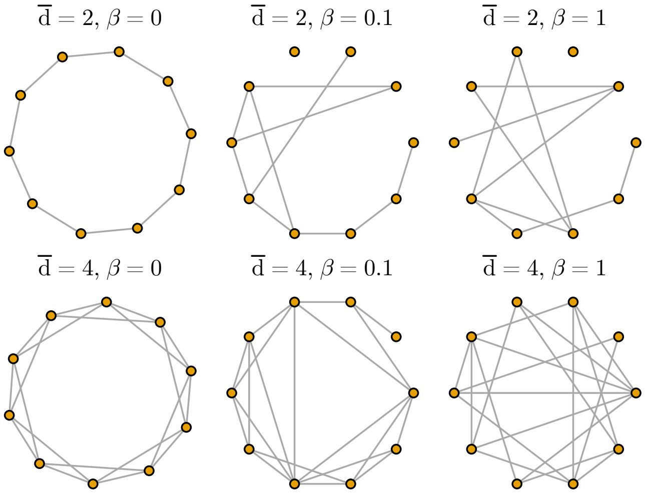

$\overline{d} = 2, \beta = 0$
$\overline{d} = 2, \beta = 0.1$
$\overline{d} = 2, \beta = 1$
$\overline{d} = 4, \beta = 0$
$\overline{d} = 4, \beta = 0.1$
$\overline{d} = 4, \beta = 1$

Figure 7.1: Small-world networks with different parameters

probability of rewiring $\beta \in [0, 1]$. With $\beta = 0$, the network is a perfect one-dimensional lattice in which each agent has exactly $\overline{d}$ neighbours (see figure 7.1). If β increases, more and more edges are rewired from a close neighbour to a random agent of the network. In case of $\beta = 1$, the network is random.

First, one-dimensional lattices with $\overline{d} = 2$ are analysed. This means that the network is a polygon. The grey histograms of figure 7.2 show the relative frequencies of alternative A at the 1 000th round of the simulation. A relative frequency is calculated for each of the 200 groups. In game I, the higher the rewards of (B, B), the more difficult it is for the agents to coordinate their decisions. If $b = 2$ or $b = 4$, at least 80% agents choose alternative A in almost all groups. With $b > 4$, there is an increasing number of groups which members choose alternative B regularly but not exclusively. This means that also suboptimal outcomes occur and that a convention is not established.

Unless there is no risk-dominant equilibrium (in case of $b = 10$), the agents' behaviour converges to the choice of alternative A in game I with further rounds of the simulations. This is indicated by the black histograms in some of the plots of figure 7.2. They present the relative frequencies of A at the 100 000th round of the simulation. Conversely, the agents predominantly learn to play alternative B in most of the instances of game II. Only if $b = 2$, the outcome (A, A) risk-dominates (B, B) and, hence, appears most often.

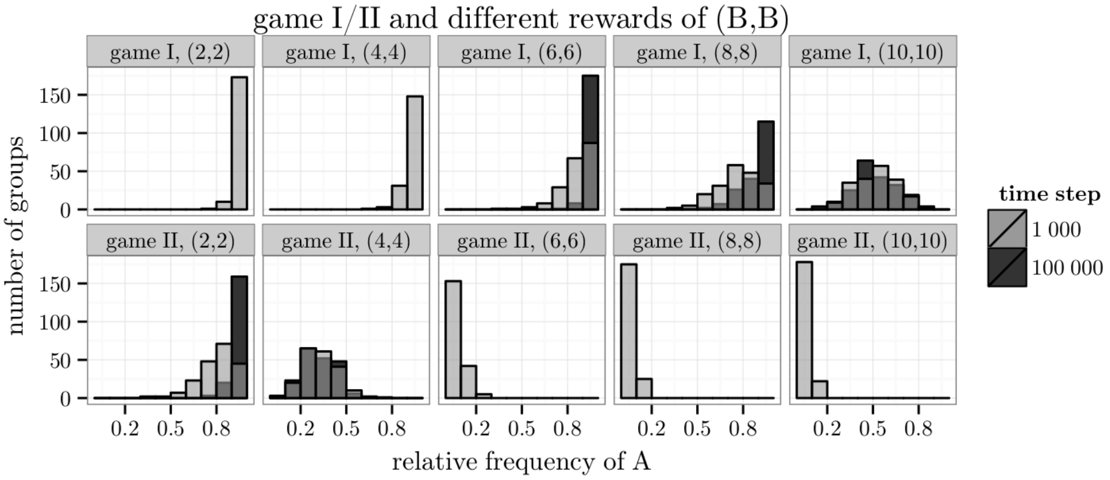


Figure 7.2: Histograms of small-world networks with $\overline{d} = 2$ and $\beta = 0$

In summary, the simulations confirm the result of Young (1998, p. 98): *the groups establish a convention by coordinating their members' choices to the risk-dominant equilibrium.* This holds true even if the risk-dominant equilibrium is inefficient (game II with $4 < b < 10$). Without risk-dominant equilibrium, a mix of several outcomes occurs. In game II with $b = 4$, the outcome (A, A) is efficient but seen less often than (B, B). In case that (A, A) and (B, B) are efficient (game I with $b = 10$), all outcomes are present with approximately the same frequency.

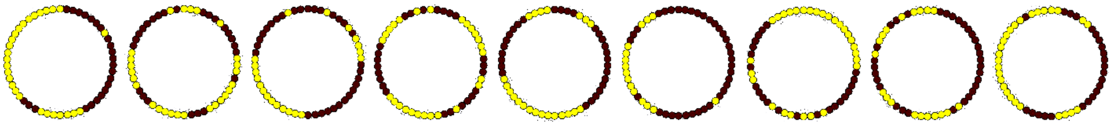


Figure 7.3: Simulation results for 9 networks with $\overline{d} = 2$, $\beta = 0$, and $b = 10$

Figure 7.3 contains nine of the 200 groups that played game I with $b = 10$. Different colours indicate different choices at the 1 000th round of the simulation. The agents are partitioned into clusters within which some agents deviate from the prevalent alternative because of the exploration rate. The clusters are fairly stable over time, for agents on the edge of a cluster have no incentive to change behaviour. These agents receive a reward of 10 from one of the partners and a reward of 0 from the other one. Switching to the other alternative would not change this pattern, unless exactly one of the two partners switches as well.

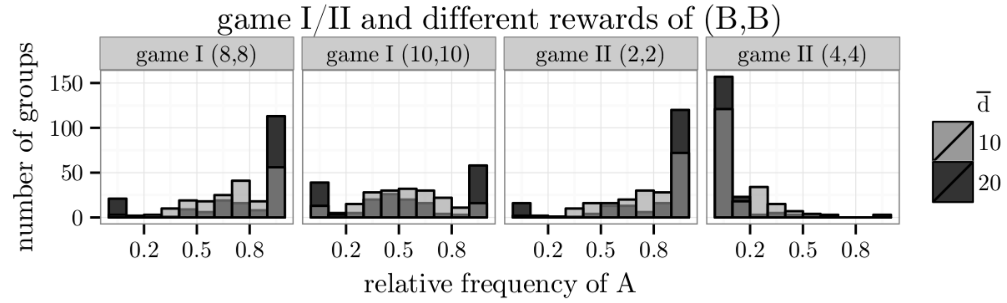


Figure 7.4: Histograms of small-world networks with $\beta = 0$

The difficulty of establishing a convention as well as the slowness of convergence in some games can be traced back to the restrictive structure of polygons (small-world networks with $\overline{d} = 2$ and $\beta = 0$). First, the convergence to a single alternative is accelerated by adding more connections to the network. Figure 7.4 shows this effect for the most problematic cases of the previous simulations: game I with $b \in \{8, 10\}$ and game II with $b \in \{2, 4\}$. The relative frequencies are measured at the 1 000th round of the simulation. A higher number of network partners enables a larger fraction of group members to choose the same alternative. Even in the case of game I with $b = 10$, approximately half of the groups can coordinate their choices within 1 000 rounds if $\overline{d} = 20$. In some groups, only alternative A is chosen, and, in other groups, only alternative B is chosen.

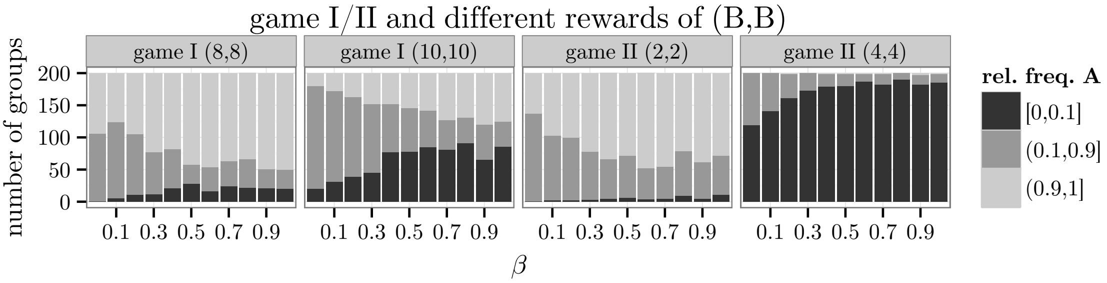


Figure 7.5: Histograms of small-world networks with $\overline{d} = 10$

Second, a high probability of rewiring facilitates the common choice of a single alternative. Similar to a larger number of partners, connections to random agents support the flow of information within the network. This enables a large fraction of the group to select the same alternative. Figure 7.5 shows the effect for networks with $\overline{d} = 10$ and different values of β. To make the plots accessible, the histograms are reduced to three intervals: $[0, 0.1]$, $(0.1, 0.9]$, and $(0.9, 1]$. It is evident that the frequency of the center interval decreases with β, which means that more and more agents agree upon a common alternative.

In conclusion, a large number of connections or a high level of randomness supports the establishment of a convention within an group. In game I with $b = 8$ and game II with $b = 2$, the expected convergence to alternative A is seen. In the games without risk-dominant outcome, the results differ. While the groups are equally divided among the two efficient outcomes in game I with $b = 10$, the agents settle on the inefficient outcome (B, B) in game II with $b = 4$.

7.2 The volunteer's dilemma

The volunteer's dilemma (Diekmann, 1985) is a generalisation of the game of chicken (figure 6.7) in respect to the number of players. It is sociologically relevant because it captures various social conflict situations, such as the call for help in an emergency (Diekmann, 1985, p. 606) or e-mail requests that are send to multiple recipients (Barron and Yechiam, 2002).

The volunteer's dilemma represents a situation with $n \in \mathbb{N}$ agents, each of whom must decide between volunteering or being idle. A collective good is provided as soon as one member of the group volunteers. This results in a utility $u \in (0, \infty)$ for every agent. But volunteering entails a cost $c \in (0, u)$. The rewards are listed in the following table from the perspective of any single agent:

		number of other volunteers				
		0	1	2	...	$n-1$
single agent	volunteer	$u-c$	$u-c$	$u-c$	$u-c$	$u-c$
	be idle	0	u	u	u	u

The volunteer's dilemma has n pure Nash equilibria in which exactly one agent volunteers. Additionally, there is a mixed Nash equilibrium. The probability of volunteering is given by

$$p = 1 - \left(\frac{c}{u}\right)^{\frac{1}{n-1}},$$

for each agent (Diekmann, 1985, p. 607). If the agents choose their actions independently of each other, there is no volunteer with probability

$$(1-p)^n = \left(\frac{c}{u}\right)^{\frac{n}{n-1}}.$$

In the following, it is investigated whether melioration learning leads to outcomes that correspond to the Nash equilibria.

7.2.1 Learning to volunteer

First, simulations are run for fixed groups of agents. This means that the same agents interact repeatedly in a volunteer's dilemma. This situation corresponds to any association, department, or group of friends that regularly needs a volunteer to, for example, complete a task or organise the annual Christmas party.

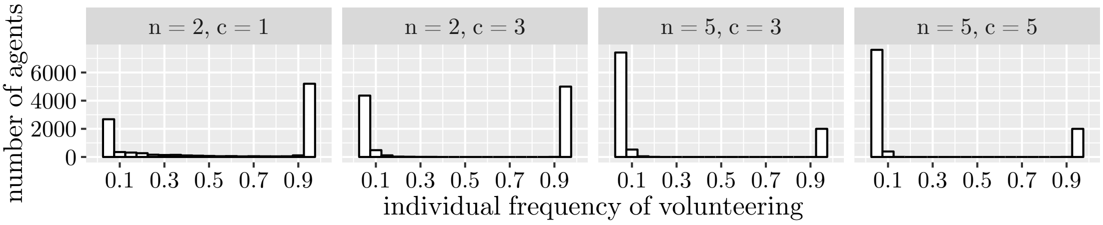


Figure 7.6: Histograms over frequencies of volunteering in fixed groups

The utility of the collective good is set to $u = 10$ and the exploration rate to $\varepsilon = 0.1$. Each simulation consists of 1 000 choices by 10 000 agents that are divided into groups of size n. The analysis reveals that the agents learn to coordinate their choices to a pure Nash equilibrium. This mirrors the results of the game of chicken in section 6.3.2. The behaviour of the agents is illustrated by four sample histograms in figure 7.6. More data is shown in appendix B.3. The x-axis depicts the *individual frequency of volunteering*, which refers to all choices of an agent during one simulation run (1 000 choices). While some of the agents always volunteer, the remaining agents are mostly idle.

Since there is exactly one volunteer in a pure equilibrium and all agents are equally likely to be this one, the individual likelihood of volunteering is $\frac{1}{n}$ and, hence, decreases hyperbolically with the group size. This is seen in the left-sided plot of figure 7.7. The relative frequency of volunteers decreases in accordance with the inverse of the group size. The difference between simulation results and the inverse of the group size can be explained by the exploration rate ε.

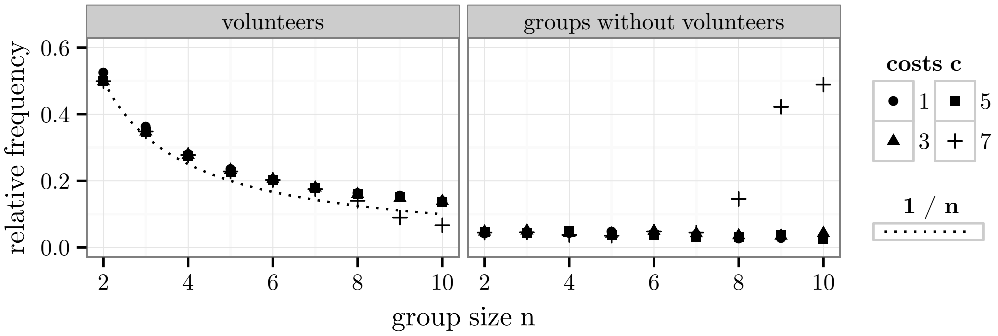


Figure 7.7: The effect of group size in fixed groups

The relative frequencies of volunteers are largely independent of the costs c. Only if $c = 7$ and $n \geq 8$, volunteering drops more rapidly. This is an implication of the exploration rate. Because the agents try different actions with probability $\varepsilon = 0.1$, the expected value of being idle is $10 \cdot \left(1 - \left(1 - \frac{\varepsilon}{2}\right)^n\right) = 10 \cdot (1 - 0.95^n)$ if nobody volunteers intentionally. With $c = 7$ and $n \geq 8$, this value is greater than the reward of volunteering ($u - c = 3$). Hence, melioration leads to a situation without volunteers except for the random volunteering that is done by exploration.

Finally, since there is generally one volunteer per group, the relative frequency of groups without volunteers is very low (unless $c = 7$ and $n \geq 8$). This is seen in the right-sided plot of figure 7.7.

7.2.2 Anonymous games

The simulations of the previous section are repeated in an anonymous setting, which means that the agents interact with different partners at each round. On the individual level, the results are similar. Most of the agents are still choosing one alternative exclusively. This is evident from figure 7.8, which, similar to figure

7.6, exhibits histograms of the individual relative frequencies of volunteering. It is roughly distinguished between three types of agents: agents who mostly volunteer (the interval $(0.9, 1]$), agents who almost never volunteer (the interval $[0, 0.1]$), and agents who occasionally volunteer (the interval $(0.1, 0.9]$). A majority of agents is either always volunteering or always idle. But there is also a considerable portion of agents who choose both alternatives.

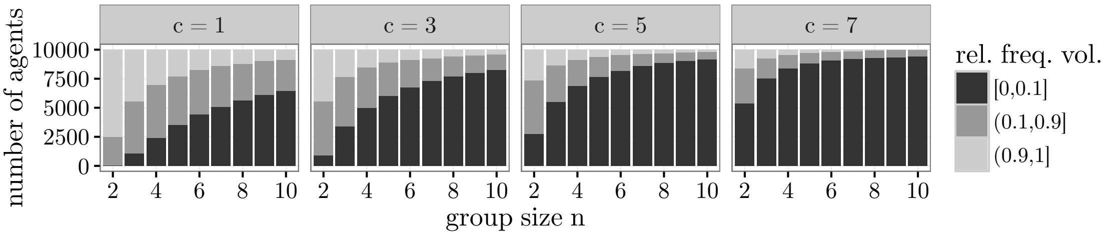


Figure 7.8: Histograms over frequencies in anonymous groups

In simulations with high costs of volunteering, the agents are either always idle or occasional volunteers. Therefore, the number of volunteers is not sufficient to obtain at least one volunteer per group. This is seen in the right-sided plot of figure 7.9. There are many idle groups if c is high. But also in the case of low costs, the rate of groups without volunteers is greater than in fixed groups.

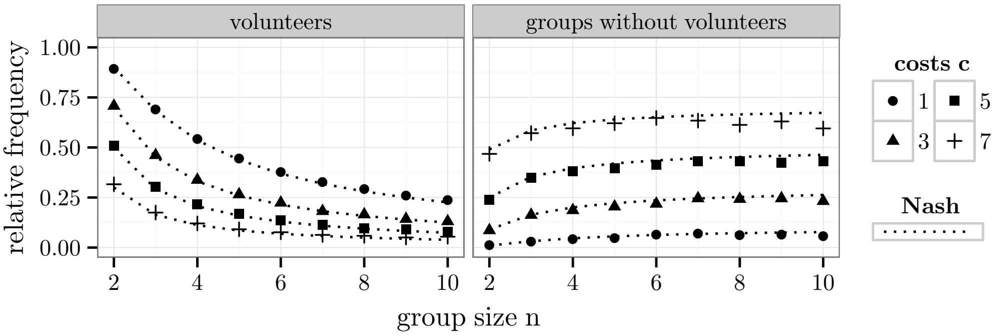


Figure 7.9: The effect of group size in anonymous groups

Additionally, it is evident from figure 7.9 that the relative frequencies of volunteering correspond to the mixed Nash equilibria. The simulation results are plotted as points, and the predictions of the Nash equilibrium are drawn as lines.

7.2.3 An asymmetric volunteer's dilemma

In an asymmetric version of the volunteer's dilemma, there are two types of players. One type is assumed to be stronger than the other one. In the following simulations, the costs of volunteering are cut in half for the strong agents. According to the mixed Nash equilibrium, the probability of a strong agent being idle is twice as high as the corresponding probability of weak agents (Diekmann, 1993, p. 77, eq. 4). This rather counter-intuitive hypotheses does not match empirical findings (Diekmann, 1993). A pure Nash equilibrium is a more plausible solution, especially if a strong agent is the single volunteer.

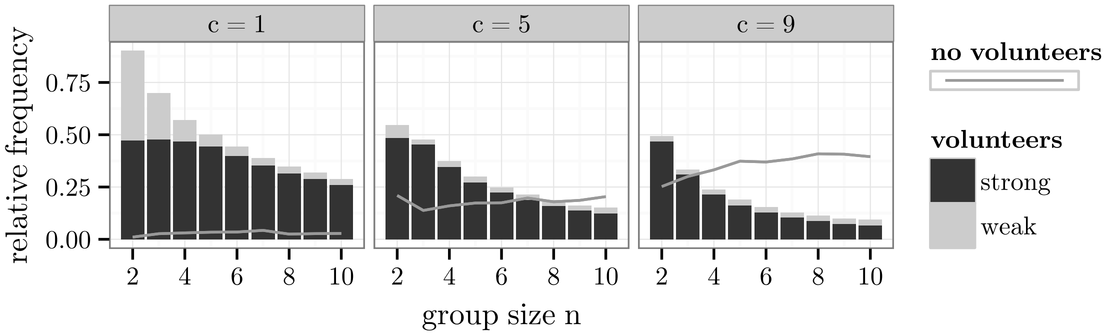


Figure 7.10: The effect of group size in the asymmetric volunteer's dilemma

Simulations were run with anonymous groups and half of the agents being strong. As seen in figure 7.10, mainly the strong agents volunteer. The relative frequencies of volunteering correspond approximately to the mixed Nash equilibrium of the *symmetric* version of the game with costs $\frac{c}{2}$.

In contrast to the predictions of the mixed equilibrium of the actual asymmetric game, the strong agents are more likely to volunteer than weak agents if they learn by melioration. In groups of size two, all the strong agents volunteer. With increasing group size, the rate of volunteers decreases at a similar rate as in the previous section. The weak agents cease to volunteer first.

Nevertheless, due to the anonymous groups condition, the agents cannot coordinate on a pure Nash equilibrium. Especially in case of large costs c, there is a high fraction of groups without volunteers. This is indicated by the grey lines in the plots of figure 7.10.

7.3 A multi-person prisoner's dilemma

The prisoner's dilemma is widely known and extensively studied not only in Sociology, but also in Economics and Biology. It describes a social situation in which the individually best action is not socially optimal. One reason for the game's popularity is the frequent occurrence of real-life situations that, on the one hand, resemble this game and, on the other hand, are important enough to desire an optimal solution (see e.g. Frank, 2011).

In section 6.3.1, the two-person version of the prisoner's dilemma was analysed. Simulations revealed that melioration leads to the only Nash equilibrium, which is socially suboptimal. At first sight, this result is not likely to change in similar interactions with multiple persons. Yet, the following simulations demonstrate under which conditions the agents learn to deviate from the Nash equilibrium.

7.3.1 The basic version

The two-person prisoner's dilemma can be extended to a similar situation with multiple persons. A particular instance of a multi-person prisoner's dilemma is called *public goods game* (e.g. Sigmund et al., 2001). In a public goods game, there is a group of $n > 1$ agents. Each agent chooses between *cooperation* and *defection*. Every cooperator pays one unit into a common pool, which is, subsequently, multiplied by a factor r, with $1 < r < n$. Let n_c denote the number of agents that choose to cooperate. Therefore, the common pool contains a total amount of $r \cdot n_c$. This amount is divided equally among all agents of the group, such that a cooperator obtains the payoff

$$\frac{r \cdot n_c}{n} - 1$$

and a defector gets

$$\frac{r \cdot n_c}{n}.$$

This model resembles the basic structure of the two-person prisoner's dilemma of figure 6.1. The fraction $\frac{r}{n}$ is the part of a contribution that is returned from the common good to a cooperator. The condition $1 > \frac{r}{n}$ implies that cooperation is dominated by defection. But since $r > 1$ and, hence, $\frac{r \cdot n}{n} - 1 > 0$, the establishment of the common good (with all agents contributing) is socially desired.

In the following simulations, a population of 10 000 agents is assumed, and the exploration rate is set to $\varepsilon = 0.1$. It is distinguished between anonymous and fixed groups. In the anonymous condition, groups are formed randomly before each round by drawing agents from a large population. Fixed groups, on the other hand, do not change. The same agents interact repeatedly with each other.

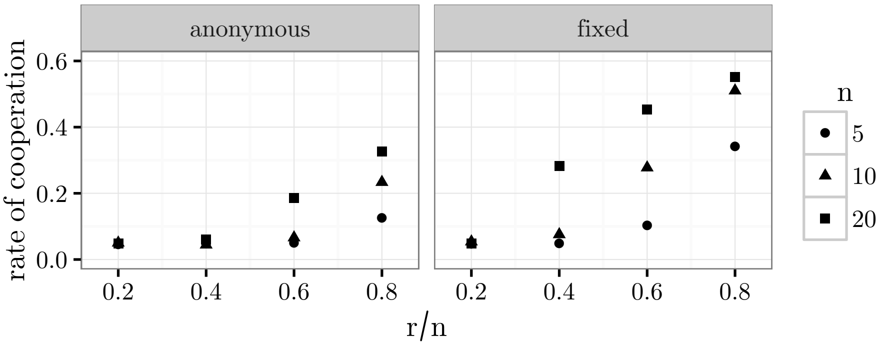


Figure 7.11: The rate of cooperation in the public goods game

The relative frequencies of cooperation are measured at the 1 000th round of the simulations and displayed in figure 7.11 for different group sizes n and reward ratios $\frac{r}{n}$. In case of randomly assembled groups, the rate of cooperation is mostly zero and increasing with n and $\frac{r}{n}$. Nevertheless, this development is only temporary. If the simulations are continued, the rate of cooperation approaches zero. This happens at a very slow rate (figure 7.12).

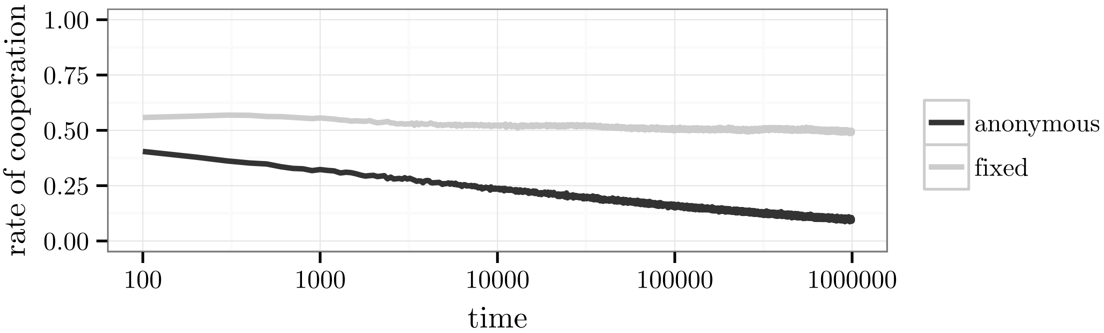


Figure 7.12: The rate of cooperation over time; $n = 20$; $\frac{r}{n} = 0.8$

In contrast, the high rates of cooperation in fixed groups do not decrease with further repetitions of the game (figure 7.12). In case of the simulation with $n = 20$ and $\frac{r}{n} = 0.8$, figure 7.13 reveals a deeper look at the frequencies of cooperation. The left-sided histogram contains the individual historical frequencies over the first 1 000 rounds. It indicates that there are two types of agents. The first type always chooses defection. The second type mainly cooperates. In the right-sided histogram, the historical frequencies are shown on the group level. On average, more than half of the members of any group cooperates. The mean of 0.57 is displayed by the vertical line.

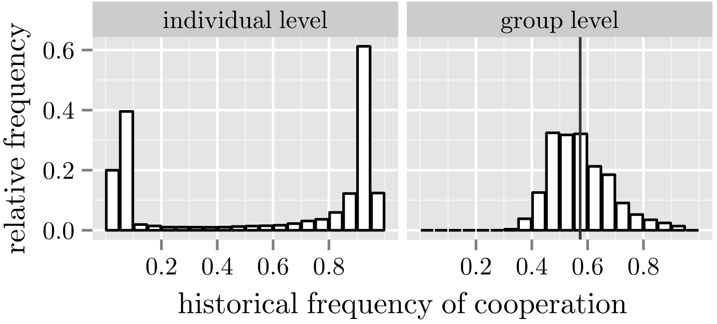


Figure 7.13: Histograms of the simulation with $n = 20$ and $\frac{r}{n} = 0.8$

According to the plots of figure 7.13, there is a substantial fraction of agents in each group who choose to cooperate. Melioration learning implies that these group members explore defection occasionally. Because of the structure of the prisoner's dilemma, the Q-value of defection should be greater than the Q-value of cooperation. Why do these cooperators not necessarily change to defection with further encounters of the prisoner's dilemma?

The alleged tendency to defection is based on a stationary point of reference. From the perspective of a cooperator, the reward of switching to defection is $(n_c - 1) \cdot \frac{r}{n}$, which is *greater* than the current reward: $n_c \cdot \frac{r}{n} - 1$. This inequality requires that no other group member changes his behaviour at the same time. With two cooperators simultaneously exploring defection, the reward is $(n_c - 2) \cdot \frac{r}{n}$. If $\frac{r}{n} > 0.5$, this is *smaller* than $n_c \cdot \frac{r}{n} - 1$. Since all agents explore their alternatives independently of each other, the probability of two or more agents changing their behaviour simultaneously increases with the group size.

When a cooperator explores defection, three types of states are possible. First, nothing else has changed compared to the previous round. Second, there are more cooperators than before. Third, there are more defectors. While defection results in a gain in the first two states, it may come with a loss in the third state. Therefore, the Q-value of defection is smaller than the Q-value of cooperation if the loss of the third state counterbalances the gains from the first two states. Besides the conditions of a large group and a high incentive of cooperation, this also requires that the third state occurs more frequently than the second state.

The last condition is plausible in the context of a prisoner's dilemma, for agents are still drawn to defection. The probability that more agents defect in the next round is always greater than the probability that more agents cooperate. Besides the random changes that are due to the exploration rate, agents deliberately choose defection because of a slightly higher Q-value. Nevertheless, because of the mechanism described above, the choice of defection comes with a loss on average and the Q-value of defection only temporarily exceeds the Q-value of cooperation.

This dynamic is seen in figure 7.14. It shows the Q-values of two representative agents of a fixed group with $n = 20$ and $\frac{r}{n} = 0.8$. It also contains the relative frequencies of cooperation measured over 25 000 rounds of the simulation (the numbers below the curves). In the first plot, the Q-value of defection is visibly greater than the Q-value of cooperation. This agent always defects and cooperates only in exploration rounds. The Q-values of the other agent are close to each other. This agent cooperates most of time but also defects deliberately apart from the exploration rate. Hence, the relative frequency of cooperation is less then 0.95.

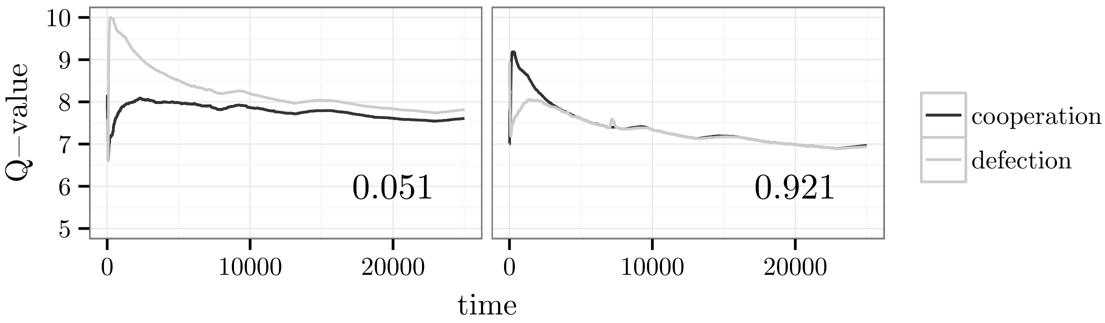


Figure 7.14: The Q-values of two members of a fixed group; $n = 20$; $\frac{r}{n} = 0.8$

7.3.2 The punishment of defection

In the past, multiple mechanisms have been suggested that *solve* the prisoner's dilemma by increasing the frequency of cooperation (an overview of studies with agent-based simulations was given by Gotts et al., 2003). One prominent example can be easily incorporated into the present model of melioration learning. This is the possibility of punishing agents who deviate from cooperation.

By itself, the option to punish is not sufficient to solve the prisoner's dilemma. It merely creates a second-order dilemma (Coleman, 1990, pp. 270-272). But with additional assumptions, punishment can be shown to support socially optimal outcomes (e.g. Axelrod, 1984; Boyd and Richerson, 1992; Brandt et al., 2003; Boyd et al., 2003; Hauert et al., 2007). In the following, it is demonstrated that melioration learning is one of these assumptions. In combination with punishment, it considerably mitigates the prisoner's dilemma.

In reference to Hauert et al. (2007), punishment is added to the present model by allowing a cooperator to impose a penalty $s > 0$ upon each defector. Every penalty comes at a cost $c \geq 0$. Let n_p denote the number of agents that choose to cooperate and punish, this implies that the defectors' payoff is reduced to

$$\frac{r \cdot n_c}{n} - s \cdot n_p$$

and that punisher are left with

$$\frac{r \cdot n_c}{n} - 1 - c \cdot (n - n_c).$$

First, it is focused on small groups ($n = 5$), which ended up with the lowest rates of cooperation in the previous section. The penalty s is set to 1. In this case, a defector gets the same payoff as a cooperator if exactly one agent carries out the punishment (the effect of lower penalties is illustrated in appendix B.4). The costs of punishment c are varied between 0 and 1, such that they never exceed s. Figure 7.15 shows results from simulations with different values of r and for the anonymous and fixed group condition. The category *punishment* stands for cooperators who also punish defectors. The horizontal lines mark the levels of cooperation without punishment (as in figure 7.11, but at the 25 000th round).

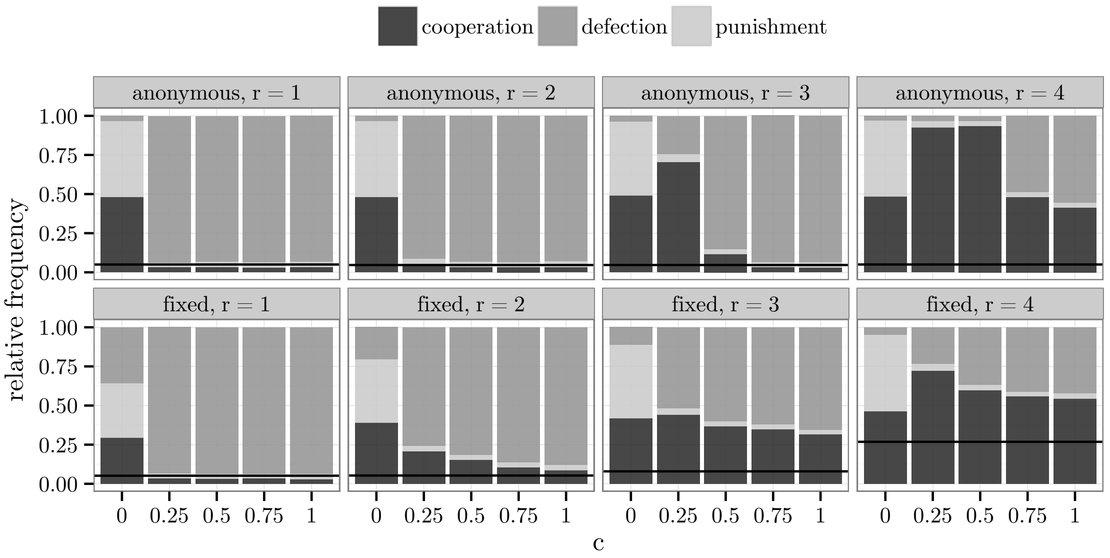


Figure 7.15: Frequencies at the 25 000th round with punishment; $n = 5$

Without costs of punishment, defection is almost eliminated. If $c > 0$, the results are mixed. The rates of cooperation approach zero in simulations with anonymous groups and low values of r. This is in accordance with previous studies (e.g. Sigmund et al., 2001; Hauert et al., 2007). The agents learn that punishment is costly. In the first plot of figure 7.16, the temporal development of the rates of choice is displayed for the simulation with $r = 2$ and $c = 0.25$. First, punisher switch to cooperation. Subsequently, cooperation is replaced by defection. In contrast, if $r = 4$ and costs are low, the rate of cooperation can be as high as 0.93. This level of cooperation is stable over time, which is apparent from the second plot of figure 7.16. In simulations with fixed groups and $r > 2$, a considerable number of agents is cooperating as well.

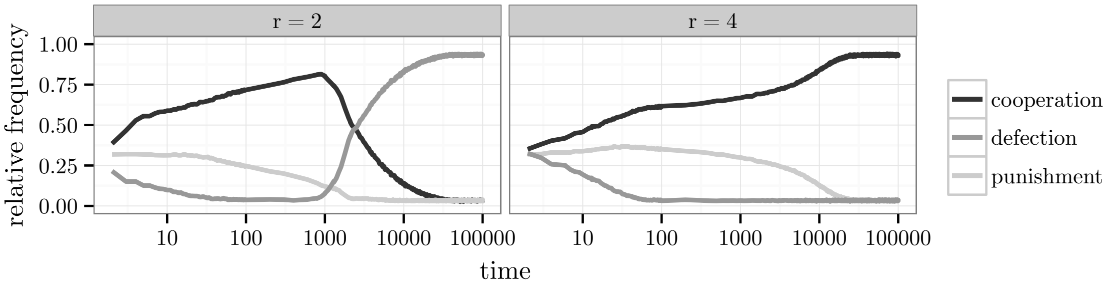


Figure 7.16: Frequencies over time in anonymous groups; $n = 5$; $c = 0.25$

Similar to the situations without punishment, the high rates of cooperation are explained by the exploration rate. If multiple cooperators simultaneously explore defection, the average reward of cooperation can be greater than the average reward of defection. The occasional penalty for defection is an additional incentive to cooperate. More specifically, there is a negative effect of punishment on defection even though its relative frequency decreases to the exploration rate (3.3%) in all simulations with $c > 0$.

The *accidental* punishment by a fellow group members increases the likelihood that the reward of defection is lower than the reward of cooperation. In combination with beneficial conditions, such as low costs of punishment or high values of r, this leads to significantly higher rates of cooperation than in situations without punishment. It is shown in appendix B.4 that the penalty s has an impact on the relationship between costs of punishment and cooperation.

Another relevant factor is the group size. The larger the group, the higher the number of agents per round that punish because of the exploration rate. This leads to outcomes with almost everyone cooperating (see figure 7.17). Compared to the horizontal lines, which mark the levels of cooperation without punishment, there is a significant raise in cooperation.

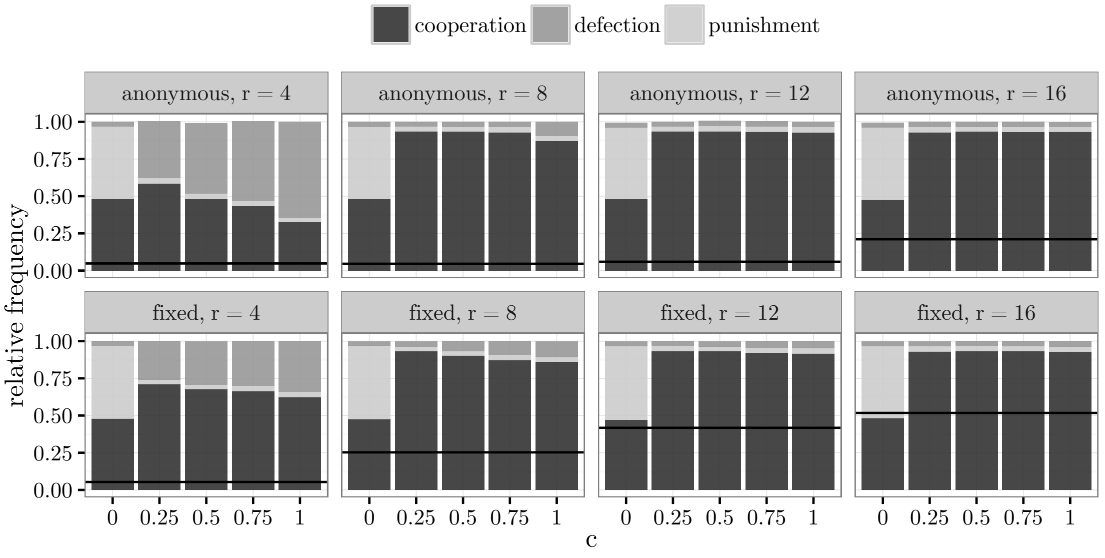


Figure 7.17: Frequencies at the 25 000th round with punishment; $n = 20$

7.3.3 The option to abstain from an interaction

Instead of the direct punishment of defectors, it may be possible to abstain from an interaction of the prisoner's dilemma. This is an indirect form of punishment since cooperators can avoid interactions with notorious defectors. The option to abstain from an interaction or, equivalently, the possibility of partner selection has been shown to increase the level of cooperation in the two-person prisoner's dilemma (Orbell and Dawes, 1993; Macy and Skvoretz, 1998; de Vos et al., 2001). In case of the public goods game, it was disclosed by Hauert et al. (2002) that this option allows cooperation to persist in spatially restricted interactions as well as in randomly formed groups. The authors deployed simulations with agents who imitated the choices of more successful agents. In the following, it is tested whether similar conclusions can be drawn for agents who learn by melioration.

It is assumed that, before an interaction takes place, each agent can choose between participating in the public goods game or receiving a default payoff l with $0 < l < r - 1$. The latter choice is called *loner*. Because $l < r - 1$, it is beneficial to be in a group without defectors instead of being a loner. But, in dependence of the choices of the other group members, it might be better to be a loner than to participate in a public goods game. Basically, the dilemma is mitigated because r is kept constant and the group size n decreases if the number of loners increases. In populations with many loners, the condition $r < n$ may not hold, which means that cooperation payoff-dominates defection. However, if many agents choose to cooperate, n becomes greater than r and defection is profitable again. A cycle of the three alternatives is expected.

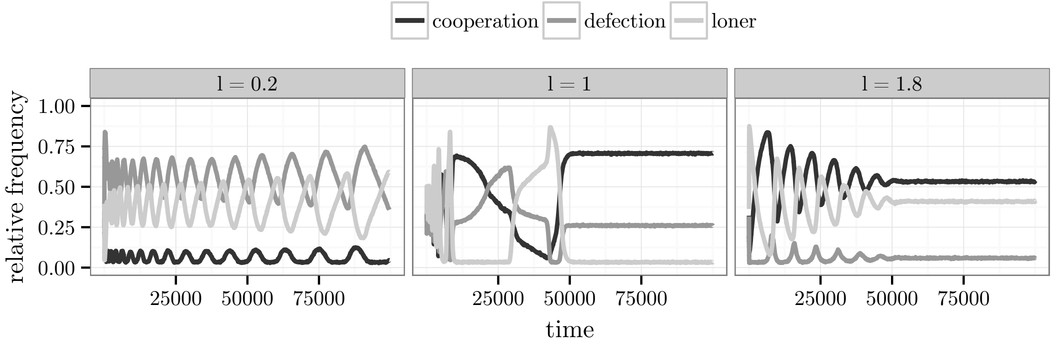


Figure 7.18: Frequencies over time in anonymous groups; $n = 5$; $r = 3$

Figure 7.18 refers to simulations in which small groups ($n = 5$) are formed randomly before every round of a public goods game. The reward r is set to a medium value ($r = 3$). Similar to the results of Hauert et al. (2002), a dynamic of continuous adjustments is visible. The agents switch between cooperation, defection, and loner. The pattern of global adjustments depends on the loner parameter l. If $l = 0.2$, there are very few cooperators, and defectors continuously change to loners and back. In case of $l = 1$, the cycle is irregular at the beginning and stable after 50 000 interactions. The last two plots of figure 7.18 indicate that a considerable level of cooperation is possible if the loner payoff is high.

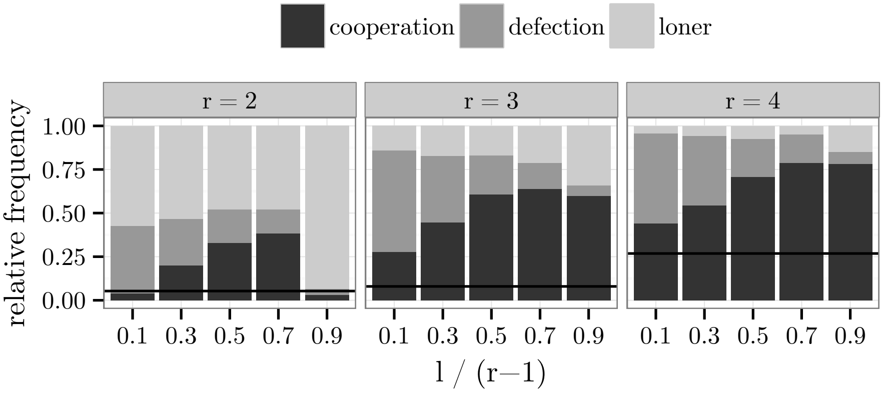


Figure 7.19: Frequencies at the 25 000th round; fixed groups; $n = 5$

In simulations with fixed groups, the rates of choice are stable (appendix B.4). Figure 7.19 shows these rates at the 25 000th round of the public goods game for different values of r and $\frac{l}{r-1}$. The horizontal lines indicate the levels of cooperation in comparable simulations without the option to be a loner. The rate of cooperation is higher if agents can choose to abstain from the interaction. Additionally, it increases with l because a high loner payoff l lowers the expected group size and raises the probability of $r > n$. This effect is less pronounced in large groups because it is less likely that $r > n$ (appendix B.4).

7.4 Conclusion

Three sociologically relevant situations were analysed in this chapter. First, the emergence of social conventions and institutions was explained by melioration and an adequate structure of interactions. In the long run, the agents' behaviour convergences to the risk-dominant outcome. If there is no risk-dominance relationship, several alternatives may coexist in restrictive structures, which means that no convention is established. Less restrictive structures and a high number of partners support the agreement on a convention.

Second, the "diffusion of responsibility" (Darley and Latané, 1968) was reproduced by agents who repeatedly interact in a volunteer's dilemma. The predictions of melioration are in line with the Nash equilibria. The larger a group, the lower the relative frequencies of volunteering. In an asymmetric version of the game, melioration implies more intuitive outcomes than the mixed Nash equilibrium.

Third, melioration constitutes another solution of the multi-person prisoner's dilemma. In combination with favourable conditions, a small or medium number of cooperators is sustained. The occasional exploration of punishment or the option to abstain from the interaction raises the level of cooperation even higher.

Finally, a short remark on the comparison of the simulation results with empirical data is made. For example, in case of the volunteer's dilemma, it was observed empirically that the individual likelihood of volunteering decreases with group size (Darley and Latané, 1968). This finding is in line with the simulations because they reproduce the mixed or the pure Nash equilibrium. Both concepts explain a negative relationship between group size and volunteering.

Nevertheless, the empirically observed decline in the relative frequency of volunteers is generally not as sharp as implied by the mixed equilibrium or the inverse of the group size (see for example Franzen, 1995, or Goeree et al., 2005). This also means that the predictions of melioration are incorrect. But this conclusion should be made with caution. There are some problems when comparing the simulation results of the previous sections with experimental studies. For instance, in Franzen (1995), the game was run only once. In this case, there is very limited learning, which mainly takes place during the instructions of the experiment.

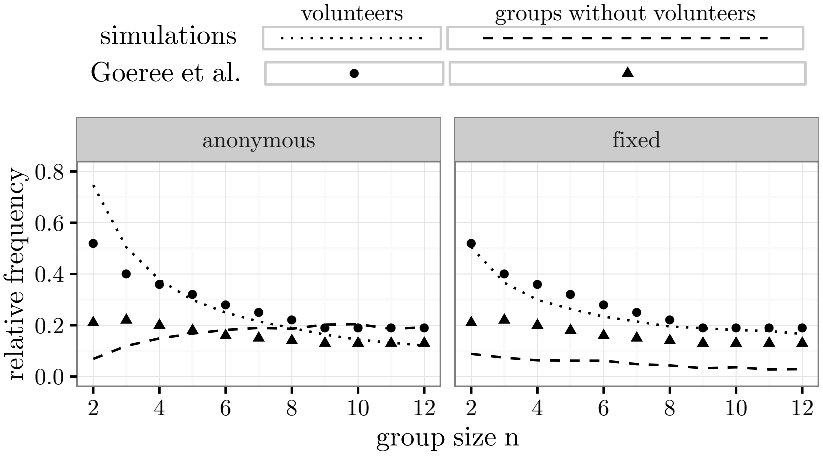


Figure 7.20: Comparison of results from simulations and experiments

In the experiments of Goeree et al. (2005), on the contrary, subjects participated in 20 consecutive rounds of the volunteer's dilemma. Some learning can be assumed. Figure 7.20 shows the relative frequencies of volunteering and of groups without volunteers for these experiments and the corresponding simulations. The results are close to each other in the fixed groups condition. However, in the experiments, the subjects interacted with randomly selected partners at each round. The predictions of the anonymous groups condition are less accurate.

A difference between the subjects of the experiments and the agents of the simulations is that the former had information about the structure of the game. It can be suspected that, even in anonymous groups, the subjects tried to coordinate their choices to a pure equilibrium, which is an obvious solution of the game. Because of the fluctuating composition of a group, this was difficult and often without success. As a result, the relative frequency of volunteering corresponded to the predictions of the pure equilibrium, but the number of groups without volunteers was slightly higher. It is presumed that, if the subjects of an experiment are not aware of the structure of the game, the findings will correspond to the predictions of the simulations.

Chapter 8

The evolution of the matching law

According to chapter 4, optimal behaviour generally deviates from the matching law in situations of repeated decision-making. From an evolutionary point of view, this finding challenges the matching law as a general theory of individual behaviour. A mechanism of decision-making that regularly leads to suboptimal behaviour should be displaced by mechanisms that converge to optimal behaviour. Hence, mechanisms that lead to the matching law, such as melioration learning, cannot be evolutionary stable (see also Houston and Sumida, 1987).

Nevertheless, there are arguments in favour of the evolutionary success of the matching law. For instance, Herrnstein (1997, p. 99) argued that matching "is a cognitive realistic approximation to maximization in many natural environments involving choices between probabilistic alternatives." This statement is in line with the results of the previous chapters. The matching law corresponds to optimal behaviour under particular circumstances (proposition 4.1) and can, therefore, be seen as an approximation to optimal behaviour.

Additionally, the matching law requires only weak assumptions about the mental abilities of an actor. With the model of melioration learning, two simple conditions were shown to be sufficient for the matching law to occur in the long run: accounting for average values and choosing greedily among the alternatives. If these processes are less cognitively demanding than any mechanism that leads to an optimal outcome, individuals who learn by melioration may have an evolutionary advantage to maximising individuals.

Even without the assumption of less cognitive requirements, the evolution of the matching law can be explained. This chapter presents a justification that builds on a theory of Ronald Heiner. In a series of papers, Heiner (1983, 1985a,b,c, 1988, 1990) forwarded an explanation of behavioural rules that frequently lead to suboptimal results. One of his main points was that suboptimal behaviour is evolutionary stable because of *uncertainties* in the actor's environment.

8.1 Uncertainties and suboptimal behaviour

Heiner (1983) stated that uncertainties in choice are sufficient conditions for the evolution of behavioural regularities and suboptimal behaviour. Uncertainties in choice exists if an actor is not always able to correctly identify a decision problem and to solve it by choosing an optimal action. Full flexibility in decision-making is required for optimal behaviour because the actor must consider any possible action as a solution. But if many uncertainties exist, less flexibility of choice might be beneficial, and the application of a simple rule of choice is possibly more successful than a complex optimal solution.

Heiner clarifies this hypothesis by the following thought experiment. Let a be an action that is optimal in some but not all situations. Furthermore, O is a fully flexible actor who considers action a as an alternative of choice. Because of uncertainties, actor O is prone to two types of mistakes: first, not choosing a in the correct situation and, second, choosing a in the wrong situation. A less flexible actor R who does not consider action a at all makes only the first type of mistake (but with certainty). Depending on various factors, such as the actors' abilities to distinguish the right from the wrong situation, the gain and loss connected to the different choices, and the likelihood of the correct situations to occur, actor R may have an advantage compared to actor O (Heiner, 1983, p. 566).

In other words, the overall performance in uncertain environments is not necessarily improved by administering an extensive set of choice alternatives (Heiner, 1983, p. 563). If the loss of a wrong action is high, it might be better not to consider this action at all. Consequently, behaviour that results from evolutionary processes is inflexible and rule-governed (Heiner, 1983, 1990). Another consequence is that, in some situations, an actor always behaves suboptimally.

According to Heiner (1985b, p. 393), there are at least two major sources of uncertainties. First, the actor's ability to process information may be limited. Second, due to a complex environment, information about the situation is possibly unreliable. The first source is disputable in an evolutionary context. Actors with unlimited abilities should have evolved. The second source of uncertainties is more plausible and is studied in the following.

The theory of Heiner (1988) describes the presence and reliability of information as crucial factors. A large amount of information allows the correct interpretation of the situation and, therefore, the choice of an optimal action. But the consideration of imperfect information can lead to worse results than ignoring it. Only if the information is sufficiently reliable, the performance is improved by letting it influence the decision-making.

In the following sections, it is tested whether the arguments of Heiner can be applied to an evolutionary explanation of melioration learning. Since melioration results in the matching law, it occasionally leads to suboptimal outcomes. The theory is retraced by simple agent-based simulations of foraging behaviour.

The simulations illustrate that the consideration of additional information is beneficial only in environments with low uncertainty. If uncertainties impede decision-making, additional information leads to additional mistakes, and this lowers the performance of an actor. As a consequence, matching behaviour is as successful as optimal behaviour and evolves by natural selection.

8.2 A model of foraging behaviour

The evolutionary success of a particular decision rule depends on the mix of problems that are faced by the decision-maker. The matching law may have been individually successful in the past because situations in which this behaviour is suboptimal are very rare. Furthermore, the outcomes between matching and optimising may only differ in situation that are unessential for the survival of a species. Hence, an evolutionary justification must start with the definition of a situation that occurred frequently in the evolution of humans and that was critical for their survival. Subsequently, the success of different forms of behaviour can be assessed, for example, by computer simulations.

The simulation-based approach to evolution is common in a research field called Artificial Life. In this field, the synthesis of life is analysed by means of computer simulations or other human-made techniques (e.g. Langton, 1995). An example is a series of studies by Seth (1999, 2001, 2007). The author simulated the behaviour of battery-driven robots. Each robot consisted of a simple neural network and three sensors that controlled its movements in a simulated environment. The environment contained energy items, which were needed to refill a robot's battery. There were two different types of energy items that varied in color and probability of refilling a battery. Additionally, the consumption of an energy item by one robot precluded its consumption by another robot.

In the simulations of Seth (1999, 2001, 2007), a genetic algorithm induced the development of robots that were optimally adapted to the environment. It was shown that the optimal behaviour is the probabilistic choice of different types of energy items. Even if one type had a higher probability of refilling than the other types, it paid to choose the latter ones from time to time because the highly refilling items were often depleted by other robots. Similar results were found by Niv et al. (2002), who used artificial bees instead of robots and a field of flowers that differed in their reward probabilities.

The examples illustrate that one of the main areas of studies in Artificial Life is the evolution of *foraging behaviour* (see also Dyer, 1995, p. 123). Even the adaptive properties of classical and instrumental conditioning have been analysed in regard to foraging (Baldassarre and Parisi, 2000). Since the search for food is critical behaviour in respect to the survival of a species, the following analysis concentrates on this subject in order to justify the evolution of the matching law and, more specifically, the evolution of melioration learning.

8.2.1 The environment

In this and the next section, a simple model of *foraging behaviour* is introduced. It is inspired by the work of Seth (2001) and implemented as an agent-based model using the NetLogo simulation framework (Wilensky, 1999) and the ql-extension (appendix A). The two main parts of any agent-based model are the *actors*, which are described in the next section, and the *environment*.

An elementary environment forms the basis of the simulations. Two types of resources, which constitute the food of the actors, are distributed randomly over the environment. The types of resources are denoted by the set $E = \{1, 2\}$. They differ in value, which is given by a positive real number $\gamma_1, \gamma_2 \in (0, \infty)$, and in color (or any other characteristic recognisable by the actors). Figure 8.1 shows the implementation of the environment in NetLogo. The resources are displayed as squares of different grey scales. White squares are places without resource.

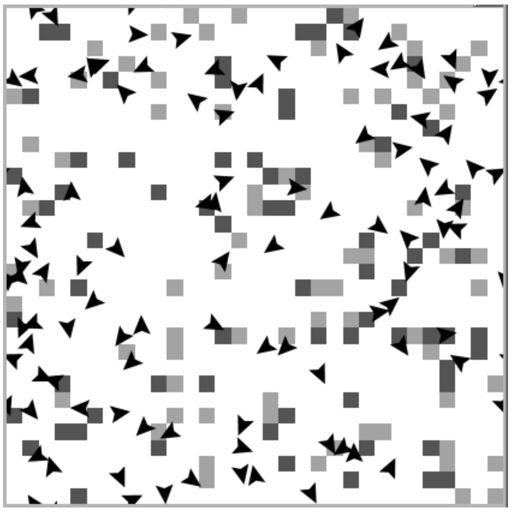


Figure 8.1: The environment of the NetLogo foraging model

In figure 8.1, the actors are represented by arrowheads. They are assumed to repeatedly choose one of the two kinds of resources. After a decision was made, an actor selects a square of the environment that contains the chosen kind and is closest to him. During the next rounds of the simulation, the actor moves towards the square. If he is the first to arrive at the square, the value of the resource is received as reward. At the same time, the resource disappears. Any other actor who has selected the same square receives no value after reaching the place. In other words, the search for resources is competitive. The consumption of a resource by one actor precludes its consumption by another one. Furthermore, the resources regrow with a fixed probability, and an actor is replaced randomly before every decision. Appendix B.5 reveals that the assumption of random replacement is essential for the following results.

In a first set of simulations, the characteristics of the environment are illustrated. A model with 1 000 actors and approximately 9 squares per actor is analysed. All actors follow the same rule, which consists of a probability vector $\boldsymbol{q} = (q_1, q_2) \in [0,1]^2$ with $q_1 + q_2 = 1$. The vector specifies the probability of choosing the first and the second kind of resource. The decision rule is fixed during a simulation run. Hence, the relative frequencies of choice approach $\boldsymbol{q}$ in the long run.

After approximately 1 000 choices, the *average values of decisions* are measured. It is distinguished between the average value $\overline{v}_1$ of choosing the first resource and the average value $\overline{v}_2$ of choosing the second resource. The overall value of choice (independent of the resource type) is denoted by v:

$$v \approx q_1 \cdot \overline{v}_1 + q_2 \cdot \overline{v}_2$$

The maximum values of $\overline{v}_1$ and $\overline{v}_2$ are γ_1 and γ_2, respectively. But, since the search for resources is competitive, the actors occasionally receive zero value, and this gradually lowers $\overline{v}_1$ and $\overline{v}_2$.

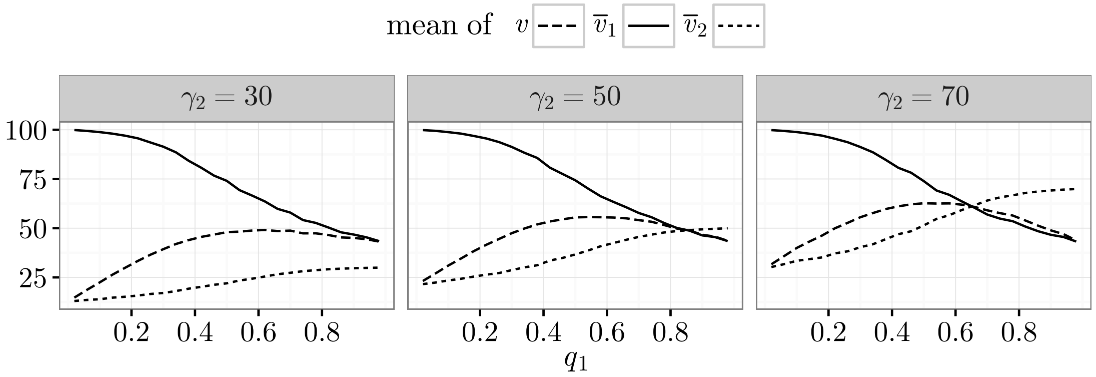


Figure 8.2: The model with 10% growth rate and $\gamma_1 = 100$

Figure 8.2 shows the means of v, $\overline{v}_1$, and $\overline{v}_2$ in regard to the whole population of actors and in dependence of the probability of choice q_1. The value of the first resource γ_1 is set to 100 and the value of the second resource γ_2 is varied ($\gamma_2 \in \{30, 50, 70\}$). Resources regrow at a rate of 10% (results from simulations with different growth rates are presented in appendix B.5). It is seen that the average values depend on the probability q_1. More specifically, both $\overline{v}_1$ and $\overline{v}_2$ decrease with the rate of consuming the respective resource. The parameter γ_2 affects the average values by raising the maximum value of $\overline{v}_2$ and by moving the optimal point of v closer to $q_1 = 0.5$.

When comparing figure 8.2 with the plots of chapter 3 or 4, it becomes clear that the foraging model is a global version of the problem of distributed choice. It is a global version because the average values of reinforcement depend not only on the individual frequencies of choice but also on the frequencies of other actors.

In every plot of figure 8.2, there is a maximum point q_1, at which the actors optimise their overall value of choice v. Even though this point is globally optimal, a single actor is able to increase his performance by deviating from this point. For example, if $\gamma_2 = 50$, the maximum occurs at $q_1 \approx 0.6$. But $\overline{v}_1$ is strictly greater than v at this point. Since the choice of a single actor only slightly affects these values, it is individually beneficial to select the first resource with certainty ($q_1 = 1$) if all other actors remain at $q_1 = 0.6$.

However, in case that all actors switch to $q_1 = 1$, the overall value declines, and the choice of the second resource becomes profitable again. Therefore, the intersection of the curves $\overline{v}_1$ and $\overline{v}_2$ might be a stable point at which no actor has an incentive to deviate from the global probability of choice. But in order to specify stable points, a model of decision-making is needed.

8.2.2 The actor models

In most simulations of Artificial Life, the actors are represented by some kind of genome that is translated into behaviour. At the beginning, a large pool of different genome specifications constitutes the population. Subsequently, the most successful specifications are selected by a genetic algorithm (e.g. Lindgren and Nordahl, 1993; Mitchell, 1995).

Similarly, the actors of the following simulations evolve by natural selection. But no genome or genetic algorithm is required. All actors follow the same decision rule. Instead of a variety of genome specifications, there is a variance in the amount of information that is used to make a decision. Thus, the simulations appoint the type of actor that employs the most beneficial amount of information.

More specifically, the melioration learning algorithm of chapter 5 is assumed. There are two types of actors. Type A ignores the state of the environment and considers only the rewards of previous decisions (algorithm 8.2.1).

Algorithm 8.2.1 Foraging behaviour of a type A actor

```
Require: exploration rate $\varepsilon \in (0,1)$, set of resource types $E$
$t \leftarrow 0$
initialise $Q_1(e) \leftarrow 0$, for all $e \in E$
initialise $K_1(e) \leftarrow 0$, for all $e \in E$
repeat
    $t \leftarrow t + 1$
    if $\varepsilon >$ random number between 0 and 1 (uniform distribution) then
        chose a random resource type $e \in E$ using a uniform distribution
    else
        choose resource type $e \in E$ greedily using the Q-values $\{Q_t(e)\}_{e \in E}$
    end if
    move gradually towards one of the closest squares with resource $e$
    if square still contains resource then
        set reward $y \leftarrow \gamma_e$
    else
        set reward $y \leftarrow 0$
    end if
    $K_{t+1}(e) \leftarrow K_t(e) + 1$
    $Q_{t+1}(e) \leftarrow Q_t(e) + \frac{1}{K_{t+1}(e)} \cdot (y - Q_t(e))$
    for all $e' \neq e$ do
        $K_{t+1}(e') \leftarrow K_t(e')$
        $Q_{t+1}(e') \leftarrow Q_t(e')$
    end for
until termination
```

Similar to algorithm 6.2.1, algorithm 8.2.1 describes stateless melioration learning with a constant exploration rate $\varepsilon \in (0,1)$. With probability ε, the actor picks a resource type randomly. Otherwise, a resource type with the currently highest Q-value is chosen. Subsequently, one of the closest squares of the environment with this resource is selected, and the actor moves towards this square during the next rounds of the simulation. If the resource is still available after reaching the square, the actor receives the value of the resource as reward.

In contrast, an actor of type B takes the state of the environment into account when making a decision. This is described in algorithm 8.2.2.

Algorithm 8.2.2 Foraging behaviour of a type B actor

```
Require: exploration rate $\varepsilon \in (0,1)$, set of resource types $E$
$t \leftarrow 0$
initialise $Q_1(s,e) \leftarrow 0$, for all $s \in \{0,1\}$, $e \in E$
initialise $K_1(s,e) \leftarrow 0$, for all $s \in \{0,1\}$, $e \in E$
repeat
    $t \leftarrow t+1$
    if squares with first resource nearby then
        $s \leftarrow 1$
    else
        $s \leftarrow 0$
    end if
    if $\varepsilon >$ random number between 0 and 1 (uniform distribution) then
        chose a random resource type $e \in E$ using a uniform distribution
    else
        choose resource type $e \in E$ greedily using the Q-values $\{Q_t(s,e)\}_{e \in E}$
    end if
    move gradually towards one of the closest squares with resource $e$
    if square still contains resource then
        set reward $y \leftarrow \gamma_r$
    else
        set reward $y \leftarrow 0$
    end if
    $K_{t+1}(s,e) \leftarrow K_t(s,e) + 1$
    $Q_{t+1}(s,e) \leftarrow Q_t(s,e) + \frac{1}{K_{t+1}(s,e)} \cdot (y - Q_t(s,e))$
    for all $s' \neq s$ and $e' \neq e$ do
        $K_{t+1}(s',e') \leftarrow K_t(s',e')$
        $Q_{t+1}(s',e') \leftarrow Q_t(s',e')$
    end for
until termination
```

The type B actor distinguishes states of the environment in regard to the local presence of resources. Since there are only two types of resources, it is sufficient to look for one of them. The state s is set to 1 if the first type is found on one of the squares that surround the actor. Depending on the particular position, $10 - 14$ of the closest squares are considered (in NetLogo: `patches with [distance myself < 2 ]`). The state s is set to 0 if no resource of the first type is close by.

8.3 Simulation results

Each simulation is run with 1 000 actors, half of which are of type A and the other half of type B. The actors explore their alternatives with probability $\varepsilon = 0.05$, the value of the first kind of resource is $\gamma_1 = 100$, and the growth rate is set to 10%. For every actor, the relative frequency of selecting the first resource is measured over 9 000 rounds after a burn-in period of 1 000 rounds. This frequency is denoted by f_1 and shown in the first plot of figure 8.3.

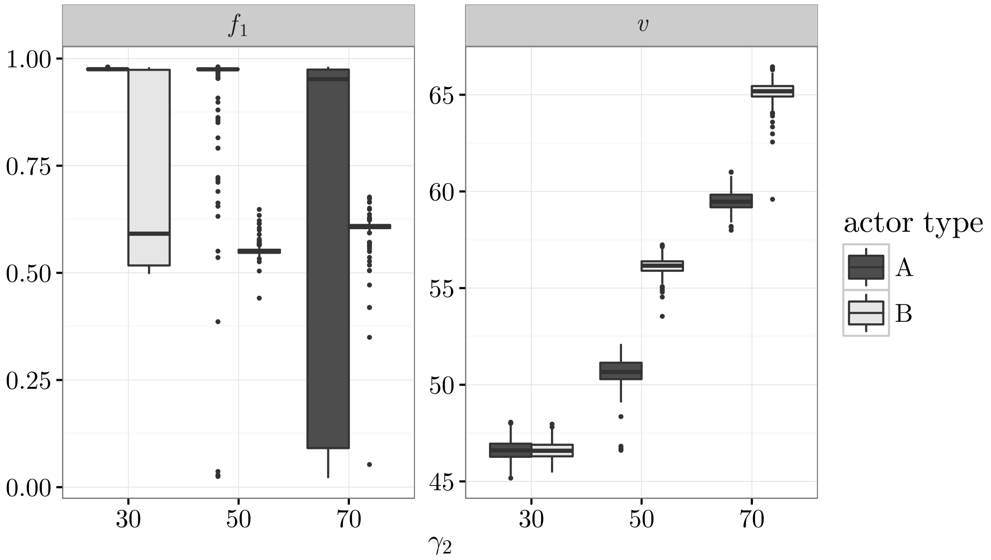


Figure 8.3: Relative frequencies and average values

The boxplots indicate that type A actors select the first resource exclusively if $\gamma_2 \in \{30, 50\}$. Actor type B chooses resource 1 and 2. In case that $\gamma_2 = 70$, the decisions of almost all actors consist of a mix of both resources. Although a high variance in frequencies persists, actors of the same type perform equally well. This is seen in the second plot of figure 8.3. It shows the average value v of all decisions during the 9 000 rounds. It is also evident that type B actors are more successful than or at least as good as type A actors. Since the former considers information about the environment, they can react to a local scarceness of the first kind of resource and switch to the other kind.

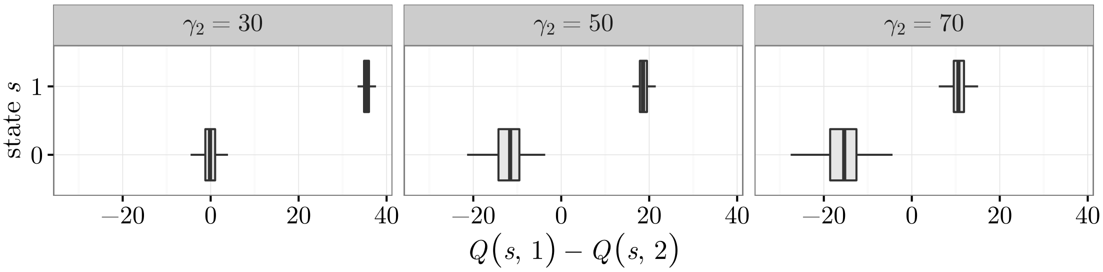


Figure 8.4: The differences in Q-values for type B actors

More specifically, type B actors learn to select the first resource in state 1 and the second resource in state 0. In figure 8.4, the differences in Q-values $Q(s, 1) - Q(s, 2)$ are plotted for both states s. In case that the first resource is not nearby ($s = 0$), this difference is less than or equal to zero. This implies the choice of the second kind of resource or a mix of both kinds. In the other state, the first resource is close-by and $Q(1, 1) > Q(1, 2)$. Hence, the actor always chooses the first resource in state $s = 1$.

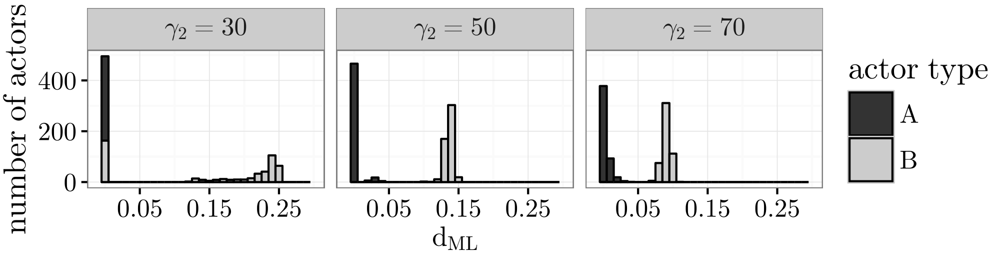


Figure 8.5: The matching law measures

Finally, it is tested whether the observed behaviour of the actors conform to the matching law. Figure 8.5 presents histograms of the matching law measures d_{ML}

(section 6.2). Only type A actors choose their resources according to the matching law. Hence, the foraging model implies a discrepancy between the matching law and optimal behaviour. Assuming that the model is a valid representation of a relevant real-world situation, this finding is inconsistent with an evolutionary explanation of the matching law.

8.3.1 Foraging with uncertainties

The results of the previous simulations are in line with the theory of Heiner (1983). Actors are better in solving a problem of decision-making if additional information about the situation is used. However, if this information is not reliable, the actors' performance should deteriorate.

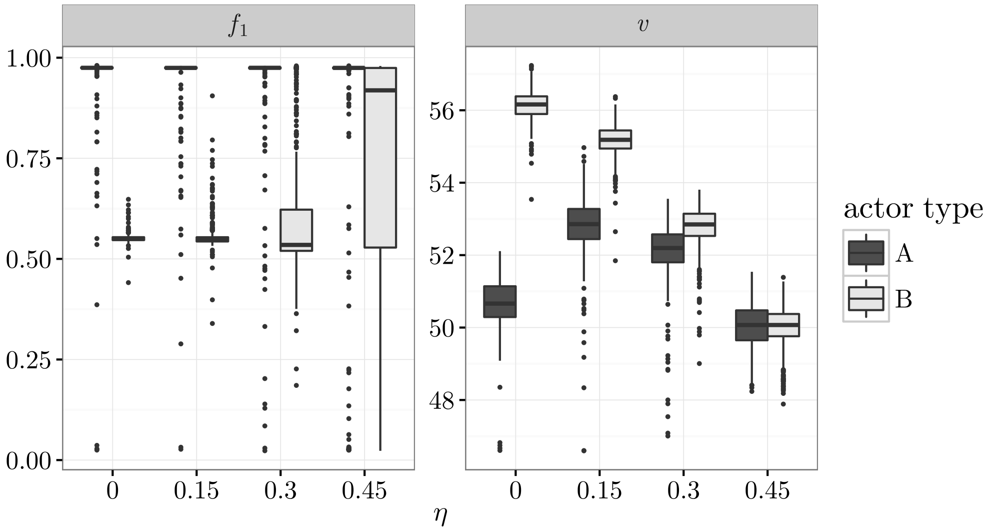


Figure 8.6: The effect of uncertainty if $\gamma_2 = 50$

Uncertainties in decision-making are introduced to the model by a parameter $\eta \in [0, 1]$ that constrains the perception of the situation. With probability η, a type B actor specifies the current state wrongly as the opposite of the actual state. The effect of different levels of uncertainty is displayed for simulations with $\gamma_2 = 50$ in figure 8.6. The discrepancy between the performance of the two actor types decreases with uncertainty. The average value of type B actors drops to the level of actor A.

Furthermore, if $\eta = 0.15$, type A is slightly more successful than in simulations without uncertainty. This is due to the faulty behaviour of type B actors. More valuable resources remain in the environment if they frequently make mistakes.

For any level of uncertainty, type A actors select the first resource exclusively. Hence, their decisions are in line with the matching law. Only in simulations with high uncertainty, the behaviour of both types of actors correspond to the matching law. This is seen in figure 8.7.

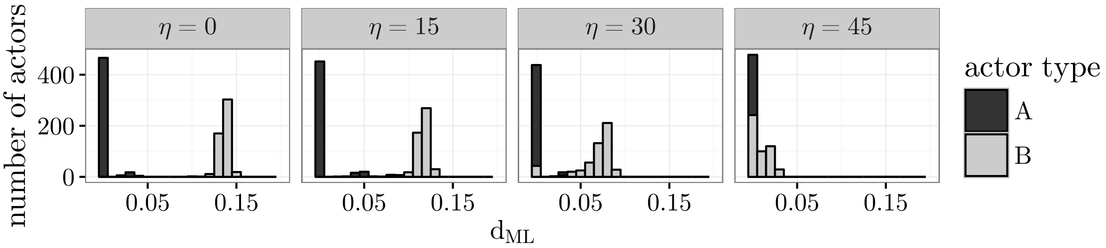


Figure 8.7: The matching law measures under uncertainty if $\gamma_2 = 50$

As a result, the matching law is evolutionary justified under the assumption of uncertainty. If the environment is sufficiently complex and the information about its state is unreliable, a simple rule that leads to the matching law is at least as successful as a more sophisticated rule that is optimal in situations without uncertainties. However, the success of an actor depends on the composition of the population. In the previous simulations, the proportion of type A actors was fixed at 50%. If the population evolves by natural selection, this proportion changes over time. In a population with many type A actors, it might be beneficial to be type B, and vice versa.

8.3.2 The evolution of actor types

The development of actor types over time is simulated by a simple evolutionary algorithm. After approximately 100 decisions, 10 of the 1 000 actors are randomly chosen and removed from the population. This process implements a natural death of the actors. Subsequently, 10 actors with an above-average performance are selected, and 10 new actors with the same type are added to the population. In the case of one new actor, the type is randomly chosen (5% mutation rate).

The parameters of the simulations are set to $\varepsilon = 0.05$, $\gamma_1 = 100$, $\gamma_2 = 50$, and a growth rate of 10%. Figure 8.8 shows the relative frequency of type A actors over time for four simulations. One time step stands for 100 decisions of the actors. It is distinguished between different levels of uncertainty η.

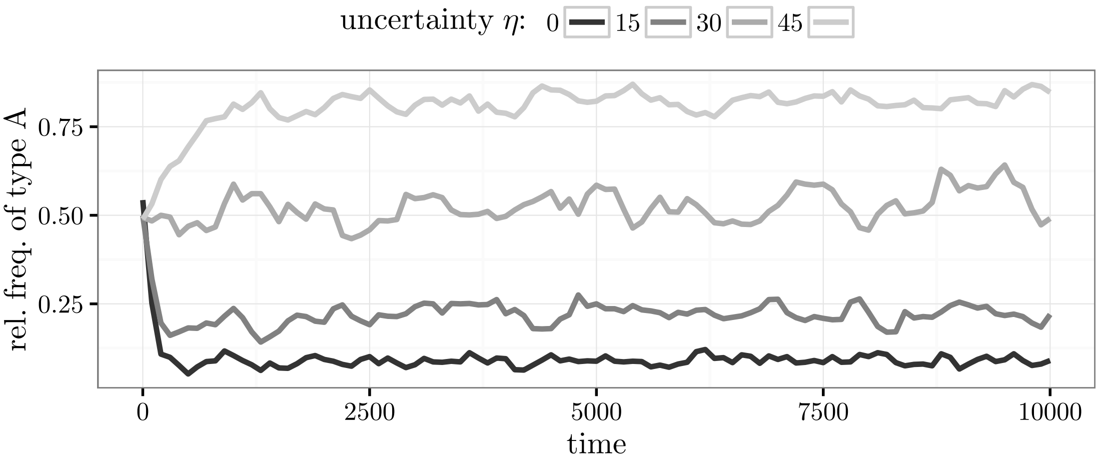


Figure 8.8: The development of type A over time

In accordance with the previous results, type A actors are displaced by type B actors in environments of low uncertainty. If $\eta = 0.3$, the performance of all actors is similar, and both types coexist with approximately the same relative frequency. In case of the highest level of uncertainty, almost all type B actors are replaced by actors of type A. In contrast to the simulations with static population, type A is superior if the composition of the population changes with the relative success of the actors.

8.4 Conclusion

The simulations demonstrated that the matching law is evolutionary stable in competitive environments with high levels of uncertainty. The results can be seen as an illustration of the theory of Heiner (1983). Simple rules of decision-making that lead to inefficient outcomes are not necessarily displaced by complex rules of optimal behaviour. This is also in line with empirical findings about the evolutionary success of economic irrationality (Tsetsos et al., 2016).

Nevertheless, the model is highly simplified. It is not claimed that, in reality, the observation and consideration of the state of the environment is irrelevant. The existence of type B actors is plausible because not all real situations depict a high level of uncertainty. Furthermore, type B behaviour is consistent with empirical findings. Experiments with humans indicated a clear disposition to the matching law if no information about the mechanism of reinforcement was given (Herrnstein et al., 1993). But with sufficient information, the subjects managed to optimise their behaviour. This variation in behaviour was found in the simulations for type B actors (figures 8.6 and 8.7).

Consequently, it is suspected that the matching law is a general regularity of choice in uncertain environments and that humans are capable of improving their outcomes if sufficient information is supplied, i.e. if uncertainties about the situation are reduced.

Similar arguments are found in other studies. For example, Flache (2002) analysed a particular social situation with rational actors, who behave optimally by definition. But the decisions were suboptimal if an actor was uncertain about the choices of the other actors. Likewise, a study of Sims et al. (2013) indicated that rational actors choose a suboptimal equilibrium if only few information about the situation is present. Sims et al. (2013, p. 139) concluded that "melioration can be reinterpreted not as irrational choice but rather as globally optimal choice under uncertainty".

Conclusion and future work

This thesis is an attempt to retrieve explanations of social phenomena from the matching law. A short overview of past research was given, and the findings were used to integrate the matching law into economic consumer theory. In contrast to the earlier work, it was emphasised that hypotheses about individual behaviour must be derived from situational properties and the preferences of the actor. Economic theory may assist in this approach.

A characteristic of the matching law is its deviation from optimal behaviour in many situations of repeated decision-making. Whereas this property was known before, it has not been formally stated for the large class of situations that was defined in proposition 4.1. Additionally, chapter 8 justified the matching law as evolutionary stable despite its suboptimal outcomes. Uncertainties in decision-making and low cognitive requirements might have benefited actors who applied the matching law instead of complex optimal solutions.

Since the matching law by itself is an insufficient behavioural assumption, the melioration learning model was introduced. For the time being, there are no general results about the convergence of melioration learning. In simple settings that can be described by Markov decision processes, melioration is guaranteed to converge to the matching law. But most social situations cannot be reduced to these models. It is still possible to analyse them by means of computer simulations.

In the previous chapters, melioration was examined in the context of various social settings. Even without strict assumptions about available information and cognitive skills, the actors were able to arrive at equilibria that are known from and justified by game theory. But this is not true in general. A guaranteed convergence to a Nash equilibrium or an optimal state requires a more advanced learning model, such as Fictitious play or Bayesian learning.

There are many open problems and ideas for future work. Most importantly, hypotheses that were derived from the matching law and melioration learning must be tested empirically. Laboratory experiments are a convenient method, but some problems exist. For example, melioration requires a long period of decision-making. Moreover, most of the existing data is not applicable because, in these experiments, information about the structure of the situation is given to the subjects. Since this information may affect the decisions but is not considered by actors who learn by melioration, new experiments have to be conducted.

Observational data may also be appropriate. On online platforms, real decisions are made repeatedly by a large set of actors. This data is, in some cases, easily accessible. However, a problem emerges in regard to the definition of the situation. The preferences of the actors must be estimated from the data. Furthermore, the statistical methods of past studies are not applicable because only one data point per actor is usually available. Thus, new methods must be developed for testing the matching law with observational data.

Besides empirical research, various theoretical extensions of the melioration learning model are possible. For instance, the melioration algorithm was mostly reduced to *stateless* learning. If different states of the environment are considered, actors are able to distinguish, for instance, between different partners. Thereby, past behaviour or the reputation of a partner is taken into account.

Additionally, information about the structure of the situation or choices of other actors can be directly included into the decision-making process. The actors may acquire beliefs about the situation that help with the coordination of behaviour and the achievement of optimal outcomes. While the basic ideas of behavioural psychology and operant conditioning neglect anything that goes beyond the own decisions and discriminative states, there are more advanced models of multi-agent reinforcement learning, which, for example, include the estimation of Q-values for a joint action space (Nowé et al., 2012, p. 455).

Finally, future research should apply the melioration model to further situations of social interactions. First, simulations with other two-person games, especially games with existing empirical results, might be performed. Second, the n-way interactions were based on very simple coordination games and on rarely seen network structures. The assumption of more realistic games and different network structures is possible. Third, enhancements can be made in regard to the multi-person prisoner's dilemma. More particularly, in Hauert et al. (2007), the options of punishment and loner are analysed in combination. Also different versions of the public goods game and other mechanisms of punishment have been investigated in the past (e.g. Hardin, 1982; Oliver et al., 1985; Heckathorn, 1988). Fourth, melioration learning may be applied to spatial games (Nowak and May, 1992) or social exchange networks (Cook and Yamagishi, 1992).

Appendix A

Simulation software

Two software tools were developed in support of the analyses of the previous chapters. The software and sample files are available at

`https://github.com/JZschache/NetLogo-ql`

and

`https://github.com/JZschache/NetLogo-games`

The source code of both tools is found at

`https://github.com/JZschache/NetLogo-Extensions`

The tools were built on the NetLogo programming environment (Wilensky, 1999). More specifically, the following software packages were used:

- NetLogo 5.2, `https://ccl.northwestern.edu/netlogo/`
- Java 7, `http://openjdk.java.net`
- Scala 2.9.3, `http://www.scala-lang.org`
- Scala STM 0.5, `https://nbronson.github.io/scala-stm/`
- Akka 2.0.5, `http://akka.io`
- typesafe/config 1.2.0, `https://github.com/typesafehub/config`
- Colt 1.2.0, `https://dst.lbl.gov/ACSSoftware/colt/`
- Gamut 1.0.1, `http://gamut.stanford.edu`

The next section deals with the *ql-extension*. It is the core of all simulations of the previous chapters, for it implements melioration learning and other learning algorithms. The software also provides an infrastructure that simplifies the research process. While the particular advantages of this tool were listed in section 6.1, the following section gives a comprehensive description of the usage and the architecture of the ql-extension.

In section A.2, the *games-extension* is presented. It facilitates the definition of two-person games as well as the calculation of optimal outcomes and the Nash equilibrium in NetLogo.

A.1 The ql-extension

Since the ql-extension is an extension of NetLogo, the latter must be installed first. The installation was tested for NetLogo 5.2.1. Afterwards, a directory named `ql` should be created in the `extensions` subdirectory of the `NetLogo` installation (see also `http://ccl.northwestern.edu/netlogo/docs/extensions.html`). All files from

`https://github.com/JZschache/NetLogo-ql/tree/master/extensions/ql`

should be downloaded and moved to the newly created directory `extensions/ql`. For example:

```
git clone https://github.com/JZschache/NetLogo-ql.git
mv NetLogo-ql/extensions/ql path-to-netlogo/extensions
```

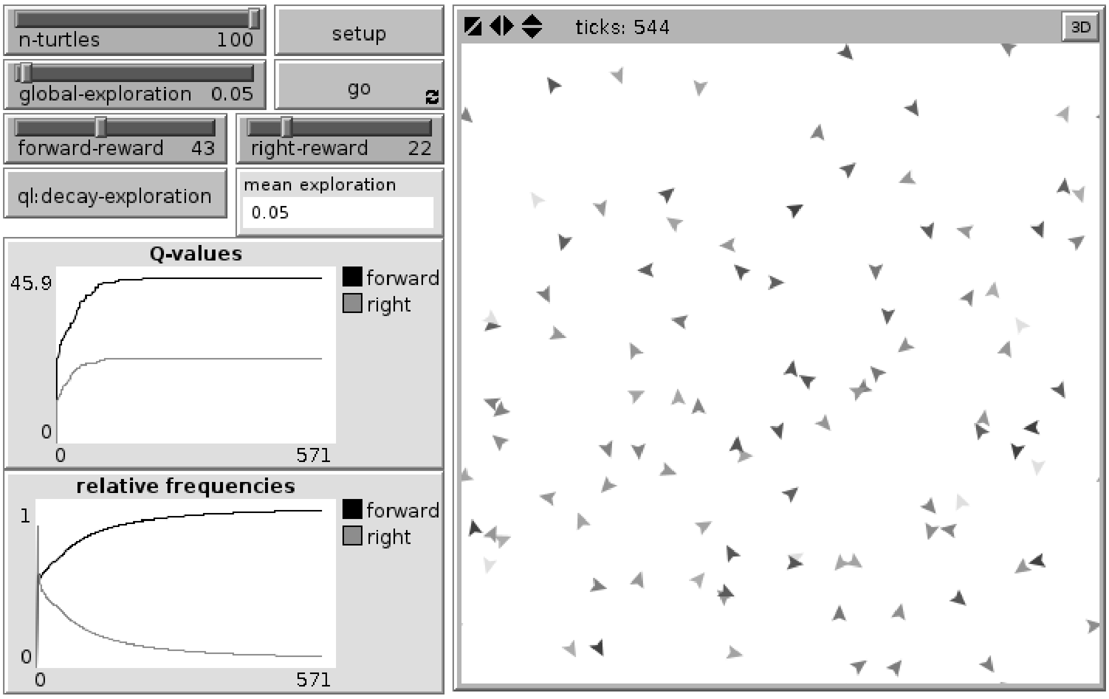
After starting NetLogo, a sample model from `NetLogo-ql/models` can be loaded.

A.1.1 A first example

Amongst other things, the ql-extension enables the simulation of agents who make decisions by melioration learning. As explicated in chapters 6 and 7, this algorithm can be applied in situations of repeated decision-making. The set of choice alternatives should be relatively stable. A simple example is given in the NetLogo model of listing A.1 and figure A.1 (downloadable as `NetLogo-ql/models/basic.nlogo`).

```
extensions[ql]

turtles-own[exploration-rate exploration-method
            alternatives q-values frequencies rel-freqs]

to setup
  clear-all
  create-turtles n-turtles [
    setxy random-xcor random-ycor
    set exploration-rate global-exploration
    set exploration-method "epsilon-greedy"
    set alternatives [ 0 1 ]
    set rel-freqs [0 0]
  ]
  ql:init turtles
  reset-ticks
end

to go
  ask turtles [
    let action ql:one-of [ 0 1 ]
    ifelse (action = 0) [
      fd 1
      ql:set-reward action (random-normal forward-reward 1)
    ] [
      right 90
      ql:set-reward action (random-normal right-reward 1)
    ]

    let total-freq sum frequencies
    if (total-freq > 0) [
      set rel-freqs map [? / total-freq] frequencies
    ]
  ]
  tick
end
```

Listing A.1: NetLogo code of first example



n-turtles 100

global-exploration 0.05

forward-reward 43

right-reward 22

mean exploration 0.05

ticks: 544

Figure A.1: NetLogo interface of first example

The agents of this model (the turtles) learn Q-values of two alternatives: “move one step forward” (alternative 0) and “turn to the right” (alternative 1). The reward of either alternative is drawn from a normal distribution with mean `forward-reward` or `right-reward` and standard deviation one. Besides these means, also the number of turtles (`n-turtles`) and the initial exploration rate (`global-exploration`) are set by the NetLogo interface (see figure A.1).

After including the line `extensions[ql]` at the beginning of the NetLogo code, the ql-extension is ready for use. Figure A.1 and listing A.1 contain a number of special commands and reporters, which are identified by the prefix ’`ql:`’. In the `setup` procedure of listing A.1, turtles are created and randomly distributed over the NetLogo world. The ql-extension is initialised by `ql:init` and an agentset (a turtleset or a patchset). If any of the variables `exploration-rate`, `exploration-method`, or `alternatives` are specified before `ql:init` is called, these values are used for the agents. Otherwise, default values are employed (0.05, “epsilon-greedy”, and empty list). The default values as well as the names of the variables are defined in the configuration file `application.conf`.

The *list of alternatives* must be a list of integers. The *exploration rate* is a positive number. Note that this rate cannot be interactively changed during the simulation (as it is usually possible in NetLogo). With `ql:decay-exploration`, a decay of this rate can be started at any point during the simulation. Afterwards, the initial exploration rate is divided by the logarithm of the number of choices (it starts counting the choices with the call of `ql:decay-exploration`).

Three *methods of exploration* are currently implemented:

- "epsilon-greedy" denotes the implementation of algorithm 5.2.1, but with exploration decreasing at logarithmic speed if `ql:decrease-exploration` is called: $\varepsilon_s \leftarrow \frac{\varepsilon}{\log\left(2+\sum_{j\in E} K_t(s,j)\right)}$.
- "softmax" refers to Boltzmann exploration as described in section B.1. The temperature is given by the exploration rate and decreases logarithmically if `ql:decrease-exploration` is called.
- "Roth-Erev" stands for the model of algorithm 6.2.2. The exploration rate decreases logarithmically if `ql:decrease-exploration` is called.

As stated in sections 5.2.2 and 6.2, the learning algorithms require the agents to use and modify Q-values when making decisions. The current state of the Q-values are accessed via the agent variable `q-values`, which is a list of numbers. This list is automatically updated during the simulation if defined by `turtles-own` (or `patches-own`). Besides the Q-values, also the names of the alternatives (`alternatives`), the frequencies of choice (`frequencies`), and the exploration rate (`exploration-rate`) are continuously updated. The names of the variables can be changed in the configuration file (`application.conf`). The relative frequencies `rel-freqs` are not part of the ql-extension and must be implemented separately (see listing A.1).

The decision of an agent and the assignment of a reward are controlled by `ql:one-of` and `ql:set-reward`:

- `ql:one-of` takes a list of alternatives (integers) and returns one of them by employing the specified exploration method.
- `ql:set-reward` maps a reward to a decision.

Furthermore, the *state* of the environment is considered by the agents (see section 5.2). The agent variable `state` keeps track of the state if it is defined by `turtles-own` (or `patches-own`). The state is an integer and can be set for each agent similar to the other variables (e.g. `alternatives`) before calling `ql:init`. Otherwise, the default (0) is used. A change in state is executed by calling `ql:set-reward-and-state` instead of `ql:set-reward` and appending a third parameter (the new state). Since the agent distinguishes between the states, the list of `alternatives` becomes a list of pairs of integers after `ql:init` was called. The first element of the pair indicates the state and the second element the alternative.

A.1.2 Parallelising the simulation

By building on the Akka framework, the ql-extension is able to parallelise the simulation and utilise multiple cores (Wyatt, 2013). Akka is written in Java and Scala. Since concurrency in Java is based on threads, also Akka uses threads. But the difficulties of data sharing and synchronisation are handled by a message-passing architecture. More concretely, Akka requires the implementation of "Akka actors" that run independently and share data by sending messages to each other (for a general introduction to the differences between thread-based and message-passing parallel programming, see Rauber and Rünger, 2013).

In the basic example of the previous section, the simulation is already parallelised. First, the NetLogo threads are not used for the ql-extension, which means that the latter runs independently of the former. Second, the learning and decision-making of the agents take place simultaneously because the ql-extension runs on multiple threads. The number of threads is controlled by the configuration file (`application.conf`: `akka.actor.default-dispatcher`; see also `http://doc.akka.io/docs/akka/2.0.5/general/configuration.html`).

Nevertheless, many parts of the simulation are executed by NetLogo, which does not parallelise naturally. This is a major bottleneck of the simulations because the ql-extension must wait for NetLogo to finish its calculations. The ql-extension solves this problem by the operation of multiple concurrently running instances of NetLogo. The deployment of multiple NetLogo instances is enabled by setting `enable-parallel-mode` to `true` (`application.conf`).

Given the current implementation, certain conditions must hold for the parallel mode to work:

1. It must be possible to assemble the agents into several groups. This may happen once at the beginning of the simulation (static groups) or repeatedly at each round (dynamic groups).

2. The situation must permit the rewards to be calculated for each group separately at a given point in time. Only the decisions of the group members and global variables of NetLogo can be used for this calculation.

This impedes the usage of the parallel mode, for instance, in the foraging model of chapter 8. The reward of an agent cannot be calculated without considering all other agents and the current state of the environment. Hence, all agents must be member of the same group, and no parallelisation is possible.

For a better understanding of the parallel mode, the architecture of the software is explained in the next section. It can be skipped if only the usage of the simulation is relevant. Section A.1.4 contains an example that makes use of the parallel mode.

A.1.3 The architecture of the ql-extension

The architecture of the ql-extension is illustrated by the class diagram of figure A.2. It clarifies the connection between the extension and the NetLogo package `org.nlogo`. It also explains how concurrency is implemented by "Akka actors". First, each NetLogo agent (a turtle or a patch) is linked to an "Akka actor". This is realised by the `QLAgent` class, which constitutes the counterpart of a NetLogo agent in the ql-extension. It is characterised by an exploration rate, a list of `QValues`, and a decision-making algorithm ("epsilon-greedy", "softmax", or "Roth-Erev"). A `QValue` instance is created for each alternative and specifies the current Q-value. The decision-making algorithm returns an element of a list of alternatives (a list of integers). It uses the exploration rate and the `QValues`.

Agents are grouped together by the class `NLGroup`. This is a subclass of `org.nlogo.api.ExtensionObject`, which makes it accessible within NetLogo code. It consists of NetLogo agents and the corresponding `QLAgents`.

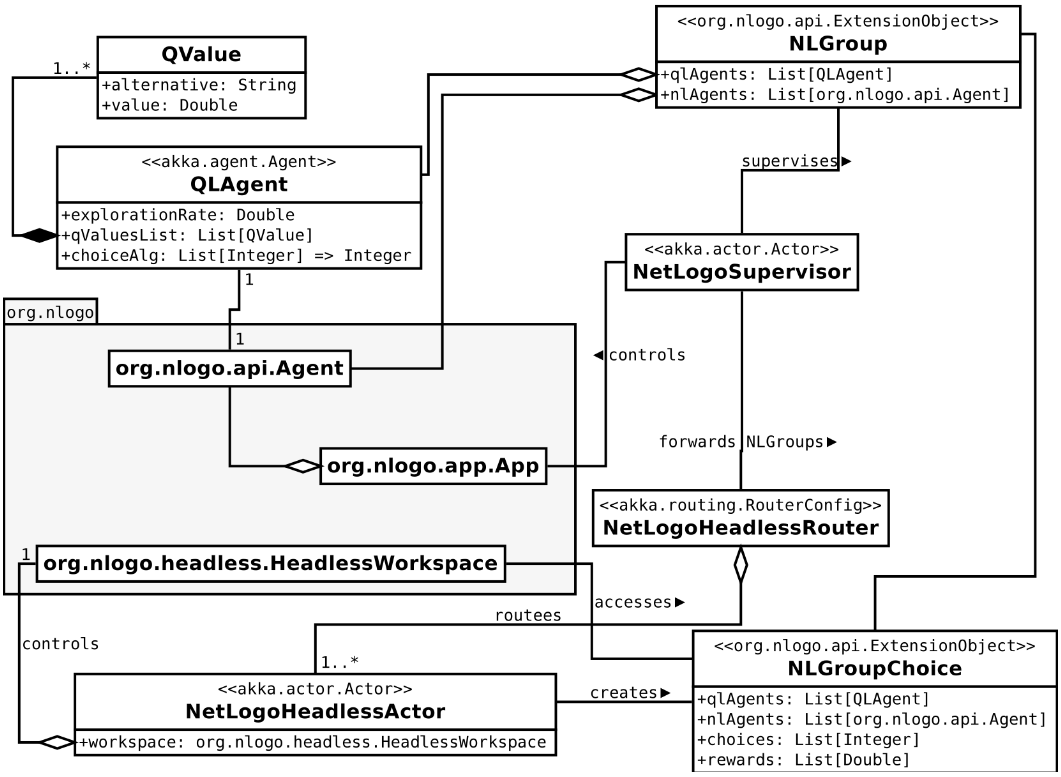


Figure A.2: Class diagram of the ql-extension

The main "Akka actor" of the extension is the `NetLogoSupervisor`. There is only one instance of this class. The `NetLogoSupervisor` has mutliple tasks. For example, it supervises all `NLGroups` and continuously triggers the choices of the agents. The speed of the repeated trigger is regulated by the corresponding slider of the NetLogo interface. When triggering the choice of the agents, the list of all `NLGroups` is forwarded to the `NetLogoHeadlessRouter`. Depending on the number of `NetLogoHeadlessActors`, the router splits this list in multiple parts. Afterwards, the `NetLogoHeadlessActors` handle the choices of the agents, and the `NetLogoSupervisor` is free to do other things.

More specifically, the `NetLogoSupervisor` also controls the main NetLogo instance (`org.nlogo.app.App`). On the one hand, it repeatedly calls the command `update` after all agents have received a reward. This command can be used to set a new `tick` and, hence, to update the NetLogo interface. On the other hand, the `NetLogoSupervisor` invokes the main NetLogo instance if the groups change during the simulation (section A.1.7).

When initialising the NetLogoSupervisor by ql:init, several *headless* workspaces of NetLogo are started in the background. *Headless* means that no graphical user interface is deployed. The number of headless workspaces is specified in the configuration file (application.conf). A separate “Akka actor” (the NetLogoHeadlessActor) controls each headless NetLogo instance. This actor continuously receives a list of NLGroups from the NetLogoSupervisor (via the NetLogoHeadlessRouter) .

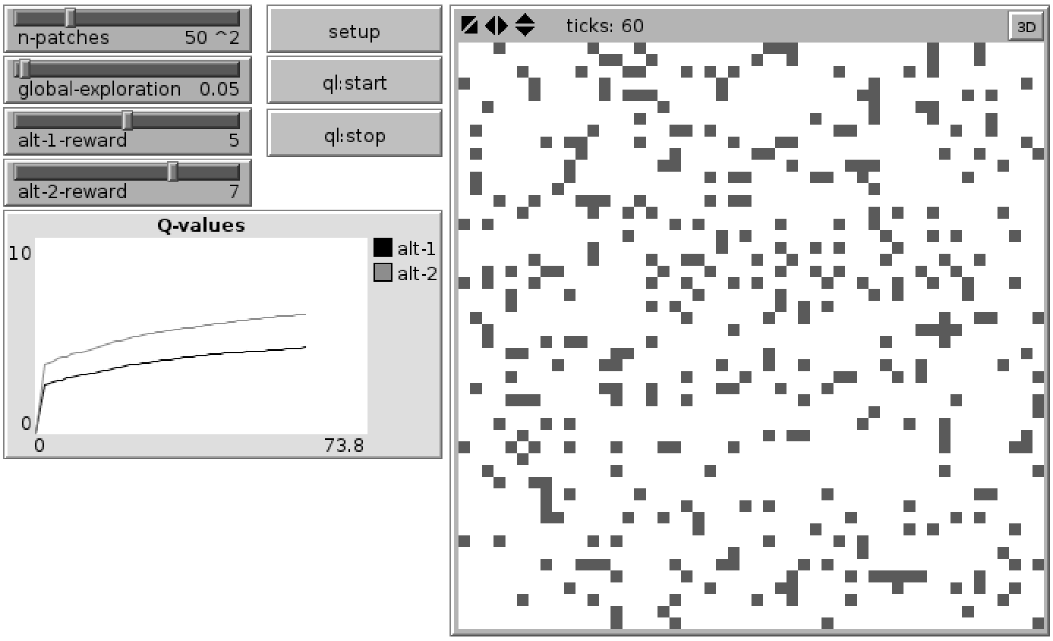
The headless NetLogo workspaces and the NetLogoHeadlessActors were added to the ql-extension in order to improve the performance. The interaction with NetLogo is the main bottleneck of the ql-extension. But repeatedly invoking NetLogo is necessary because the reward function, which maps the agents’ choices to rewards, should be specified in the NetLogo model and not within the ql-extension. Therefore, multiple instances of NetLogo are run in parallel. Their only task is to repeatedly calculate the rewards of several groups of agents.

The performance of repeatedly calling the reward function is optimised by compiling this function only once. This is problematic because the NetLogo extensions API does currently not support the passing of arguments to a compiled function (see https://github.com/NetLogo/NetLogo/issues/413). A solution was mentioned by Seth Tisue in the corresponding discussion[1] and is implemented in the ql-extension. Each NetLogoHeadlessActor is identified by a unique number. This number is forwarded to the reward function when it is called by the NetLogoHeadlessActor. The reward function calls ql:get-group-list with the identifying number and receives a list of NLGroupChoices. Besides the agents, an NLGroupChoice also contains a list of the agents’ decisions. The agents and their choices are accessed by the reporters ql:get-agents and ql:get-decisions. The rewards are set to an NLGroupChoice by ql:set-rewards. The reward function can also be used to update the (NetLogo) agents directly, e.g. by moving the agents within the NetLogo world or by setting variables. Since the agents are passed from the main NetLogo instance, the changes take effect in this instance as well. Finally, the reward function must return a new list of NLGroupChoices that correspond to the received list but with the rewards attributes set.

[1] https://groups.google.com/forum/#!msg/netlogo-devel/8oDmCRERDlQ/0IDZm015eNwJ

A.1.4 The two-armed bandit problem

In figure A.3 and listing A.2, a second NetLogo model is given (downloadable as `NetLogo-ql/models/n-armed-bandit.nlogo`). It makes use of the parallel mode and implements the two-armed bandit problem. An agent, who is represented by a patch, chooses repeatedly between two alternatives. The reward of either choice is drawn from a normal distribution with mean `alt-1-reward` or `alt-2-reward` and standard deviation one. Different colors (white or grey) indicate the last decision of an agent.



n-patches 50 ^2
global-exploration 0.05
alt-1-reward 5
alt-2-reward 7
ticks: 60

Figure A.3: NetLogo interface of the 2-armed bandit problem

The setup function of listing A.2 is similar to the one of the first example in section A.1.1. The `exploration-rate` and `exploration-method` are set for each agent, and the ql-extension is initialised by `ql:init`. Since there is no interaction between the agents, each group consists of a single agent. In line 13 of listing A.2, a list of groups is created, one group for each agent. At the same time, a list of alternatives is defined. The group structure is handed over to the extension in line 14. It is a static structure because the groups do not change during the simulation. In section A.1.7, it is demonstrated how to implement a dynamic group structure.

```
extensions[ql]
patches-own[ exploration-rate exploration-method q-values ]

to setup
  clear-all
  set-patch-size 400 / n-patches
  resize-world 0 (n-patches - 1) 0 (n-patches - 1)
  ask patches [
    set exploration-rate global-exploration
    set exploration-method ”epsilon-greedy”
  ]
  ql:init patches
  let groups [ql:create-group (list (list self [0 1]))] of patches
  ql:set-group-structure groups
  reset-ticks
end

to-report get-rewards [ headless-id ]
  let group-list ql:get-group-list headless-id
  report map [reward ?] group-list
end

to-report reward [group]
  let agent first ql:get-agents group
  let decision first ql:get-decisions group
  ifelse decision = 0 [
    ask agent [set pcolor blue]
    report ql:set-rewards group (list random-normal alt-1-reward 1)
  ] [
    ask agent [set pcolor red]
    report ql:set-rewards group (list random-normal alt-2-reward 1)
  ]
end

to update
  tick
end
```

Listing A.2: NetLogo code of the 2-armed bandit problem

The simulation is started and stopped by `ql:start` and `ql:stop`. After starting the simulation, two functions are called repeatedly by the ql-extension and, hence, must be implemented in the NetLogo model. By default, the functions are named `get-rewards` and `update`. The names of the functions can be changed in the file `application.conf`. The first function is used to calculated the rewards of a group of agents. It comes with exactly one parameter (`headless-id`). The second function is executed repeatedly after every agent has received a reward. In listing A.2, a new tick is set, which updates the NetLogo interface.

The following list describes some commands of the ql-extension in detail:

- `ql:create-group` is a reporter that creates a group from a list of pairs. Each pair is a list with two elements: first, an agent and, second, a list of integers (the alternatives). An object of type `NLGroup` is returned.

- `ql:set-group-structure` takes a list of objects of type `NLGroup` as parameter. It sets a static group structure.

- `ql:start` or `ql:stop` starts or stops the simulation.

- `ql:get-group-list` can only be called from the reward function and must forward the `headless-id`. It returns a list of objects of type `NLGroupChoice`.

- `ql:get-agents` returns the list of NetLogo agents (turtles or patches) that are held by a `NLGroupChoice`.

- `ql:get-decisions` returns the list of decisions that are held by a `NLGroupChoice`. The indices of the decisions correspond to the indices of the agents that are held by the `NLGroupChoice` such that the decision at index i belongs to the agent at index i.

- `ql:set-rewards` sets a list of rewards for the decisions that are held by a `NLGroupChoice`. It returns a copy of the `NLGroupChoice` with the rewards attribute set. The indices of the rewards must correspond to the indices of the agents that are held by the `NLGroupChoice` such that the reward at index i belongs to the agent at index i.

A.1.5 Parameter sweeps

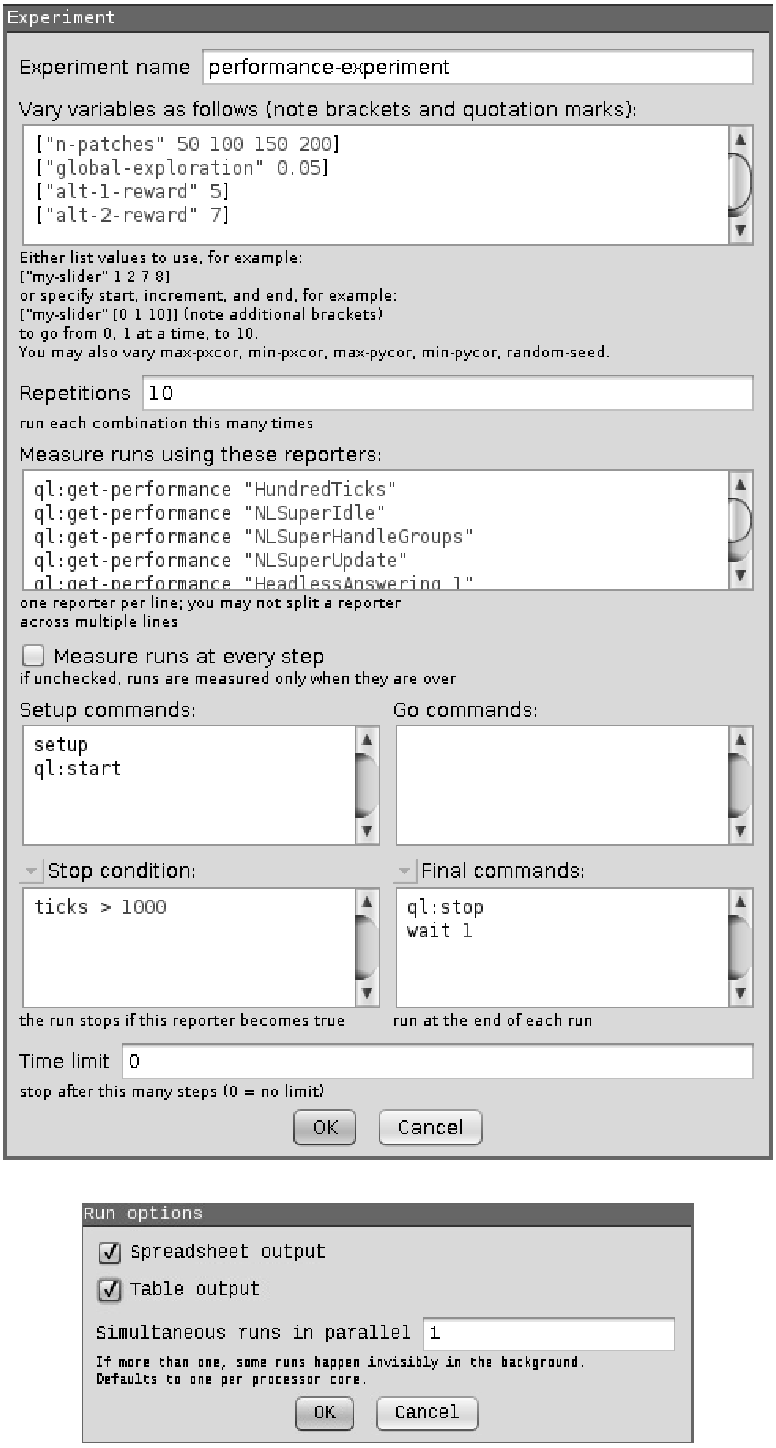
The ql-extension supports parameter sweeps with the BehaviourSpace of NetLogo (see `http://ccl.northwestern.edu/netlogo/docs/behaviorspace.html`). The setup slightly differs from the usual proceeding (see figure A.4). First, there are no “Go commands”. Instead, `ql:start` is called as “Setup command”. Second, `ql:stop` is added to “Final commands”. Since the BehaviourSpace recreates the agents instantly even if the ql-extension has not finished yet, an error occurs once in a while. By calling `wait 1` after `ql:stop`, this error is prevented. If the BehaviourSpace waits for one second, the ql-extension is usually ready for new agents. Third, a “Stop condition” must be present because the time limit does not work. Finally, only one experiment can run simultaneously (see lower window of figure A.4). The parallelisation is already built into the ql-extension.

This setup works as long as the measures are run at the end of the simulation. If the option “measure runs at every step” is enabled, a “Go command” that forces NetLogo to wait for the next tick must be added. For example:

```
to wait-for-tick
  set nextTick nextTick + 1
  while [ticks < nextTick] [
    random 100
    ; do useful stuff
    ; e.g. update globals here
  ]
end
```

It is also possible to run the experiments from command line (`http://ccl.northwestern.edu/netlogo/docs/behaviorspace.html#advanced`). As already stated, the number of threads is limited to one when using the ql-extension:

```
java -Xmx1024m -Dfile.encoding=UTF-8 -cp NetLogo.jar org.nlogo.
    headless.Main --model ql-model.nlogo --experiment performance-
    experiment --threads 1
```



Experiment

Experiment name: performance-experiment

Vary variables as follows (note brackets and quotation marks):

```
["n-patches" 50 100 150 200]
["global-exploration" 0.05]
["alt-1-reward" 5]
["alt-2-reward" 7]
```

Either list values to use, for example:
["my-slider" 1 2 7 8]
or specify start, increment, and end, for example:
["my-slider" [0 1 10]] (note additional brackets)
to go from 0, 1 at a time, to 10.
You may also vary max-pxcor, min-pxcor, max-pycor, min-pycor, random-seed.

Repetitions: 10

run each combination this many times

Measure runs using these reporters:

```
ql:get-performance "HundredTicks"
ql:get-performance "NLSuperIdle"
ql:get-performance "NLSuperHandleGroups"
ql:get-performance "NLSuperUpdate"
ql:get-performance "HeadlessAnswering 1"
```

one reporter per line; you may not split a reporter
across multiple lines

□ Measure runs at every step

if unchecked, runs are measured only when they are over

Setup commands:

```
setup
ql:start
```

Go commands:

Stop condition:

```
ticks > 1000
```

the run stops if this reporter becomes true

Final commands:

```
ql:stop
wait 1
```

run at the end of each run

Time limit: 0

stop after this many steps (0 = no limit)

Run options

☑ Spreadsheet output

☑ Table output

Simultaneous runs in parallel: 1

If more than one, some runs happen invisibly in the background.
Defaults to one per processor core.

Figure A.4: NetLogo interfaces of an experiment

A.1.6 Performance

An advantage of the ql-extension is the enhanced performance of the simulations, which is achieved by concurrency. The performance of a simulation can be measured by monitoring the `NetLogoSupervisor` and the `NetLogoHeadlessActors`. Different measures are included in the ql-extension and obtained by the reporter `ql:get-performance`. The reporter takes one of the following string parameters and returns the time in milliseconds:

- "HundredTicks" - the average time that is needed for 100 ticks.
- "NLSuperIdle" - the average time of the `NetLogoSupervisor` being idle, which means that it waits for the `NetLogoHeadlessActors` to finish the reward calculations.
- "NLSuperHandleGroups" - the average time of the `NetLogoSupervisor` forwarding the `NLGroups` to the `NetLogoHeadlessRouter` (this becomes relevant in case of a dynamic group structure because the primary NetLogo instance needs to be invoked).
- "NLSuperUpdate" - the average time of the `NetLogoSupervisor` executing the `update` procedure.
- "HeadlessIdle 1" - the average time of the first `NetLogoHeadlessActor` being idle, which means that it waits for the `NetLogoSupervisor` to forward `NLGroups` or for the NetLogo headless workspaces to calculate the rewards.
- "HeadlessHandleGroups 1" - the average time of the first `NetLogoHeadless-Actor` initiating the agents to make a decision.
- "HeadlessHandleChoices 1" - the average time of the first `NetLogoHeadless-Actor` handling the agents' decisions (calling the NetLogo headless workspace).
- "HeadlessAnswering 1" - the average time of the first NetLogo headless workspace waiting for the first `NetLogoHeadlessActor` to forward the list of `NLGroupChoices`.

By changing the number of the last four parameters, the corresponding values of the other `NetLogoHeadlessActors` are obtained. The usage of the performance measures must be enabled in the configuration file (`application.conf`).

In the following, it is shown that the number of `NetLogoHeadlessActors` is a relevant factor in regard to the performance of the simulation. A performance leak exists because the `NetLogoSupervisor` must wait for the `NetLogoHeadlessActors` to finish their calculations. This means that the performance of a simulation is optimised by increasing the number of concurrently working headless actors. With many of them, the `NetLogoSupervisor` is rarely idle, and the simulation runs faster. This should be evident by the performances measure "NLSuperIdle" and "HundredTicks". In case of the two-armed bandit simulation (section A.1.4), both measures are pictured in figure A.5 for different numbers of patches and different numbers of concurrently working `NetLogoHeadlessActors`.

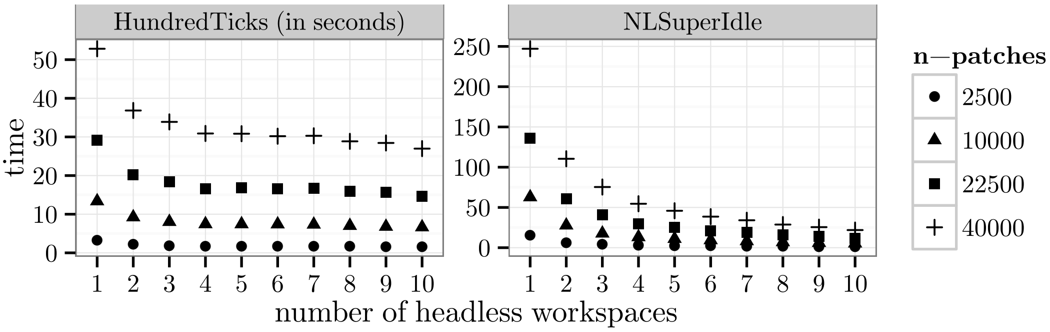


Figure A.5: Performance of the two-armed bandit simulation

Figure A.5 reveals that the performance of the simulation increases with the number of `NetLogoHeadlessActors`. Comparing the simulations with one and ten `NetLogoHeadlessActors`, the former needs twice as long as the latter for 100 ticks. Moreover, the second graph illustrates that there is not much room for improvement. The `NetLogoSupervisor` is almost never idle if ten `NetLogoHeadlessActors` work concurrently.

Furthermore, the measures "NLSuperHandleGroups", "HeadlessHandleChoices", and "HeadlessAnswering" are very close to zero and independent of the number of patches or `NetLogoHeadlessActors`. The time that the `NetLogoSupervisor` needs to execute the `update` procedure ("NLSuperUpdate") increases with the number of patches. The remaining two measures ("HeadlessHandleGroups" and "HeadlessIdle") show similar developments as the total performance (see figure A.6). Since the list of `NLGroups` is distributed equally among the headless actors, the average waiting time as well as the average working time of the NetLogo-HeadlessActors decreases with their number.

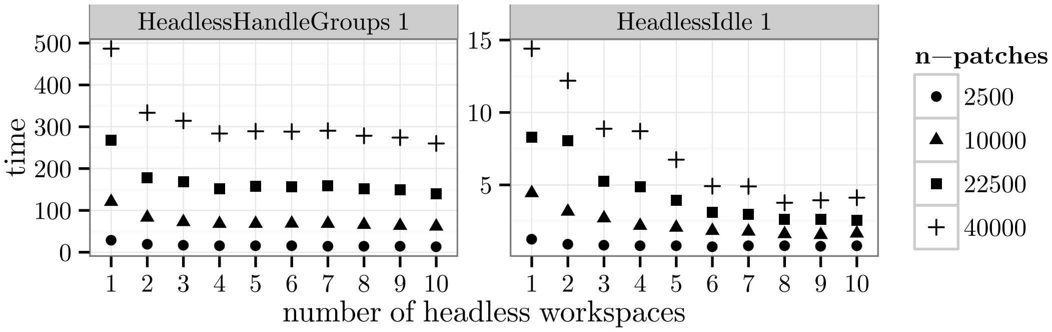


Figure A.6: Performance of the first NetLogoHeadlessActor

The optimal number of NetLogoHeadlessActors depends on the disposable computational resources. Every NetLogo headless workspace is started on a new thread. If ten NetLogoHeadlessActors are used, NetLogo occupies at least ten threads. These threads must block from time to time when waiting for the ql-extension. It is, therefore, beneficial to have at least the same number of threads available for the ql-extension. The minimum and maximum number of threads for the ql-extension are set in the configuration file (application.conf). Further experiments have shown that performance is best if the maximum number of threads is slightly above the number of NetLogoHeadlessActors (for example, maximal 12 threads if 10 headless actors are employed). The optimal number of threads and NetLogoHeadlessActors should be evaluated empirically given the maximum number of agents.

A.1.7 Dynamic group structures

In the simulations of section 7.1, two-person games were played with multiple partners. The simulations were run for 1 000 rounds. If the group structure had been fixed, agents with 2 partners would have had 2 000 choices and agents with 4 partners 4 000 choices during one simulation. Since the number of choices affects the outcome variables, such as Q-values and relative frequencies, simulations with different numbers of partners are not comparable.

One solution is that each agent has one choice per round regardless of the number of partners. In simulations with multiple partners, this requires a different group structure at each round. As stated in section A.1.3, the `NetLogoSupervisor` is responsible for updating the group structure. If `ql:set-group-structure` is not called after initialising the ql-extension, the `NetLogoSupervisor` executes the procedure `get-group` before every new round of decision-making. This procedure must, hence, be implemented in the NetLogo model (the name can be changed in the file `application.conf`). An example is given in listing A.3.

```
globals [ group-structure ]
turtles-own [ exploration-rate exploration-method ]

to setup
  clear-all
  create-turtles 100 [
    set exploration-rate 0.5
    set exploration-method ”epsilon-greedy”
  ]
  ql:init turtles
  set group-structure []
  let i 0
  while [i < (count turtles)] [
    ask turtle i [
      let anotherTurtle turtle ((i + 1) mod (count turtles))
      let group ql:create-group (list
        (list self (n-values 2 [ ? ]))
        (list anotherTurtle (n-values 2 [ ? ])))
      set group-structure lput group group-structure
    ]
    set i i + 1
  ]
end

to-report get-groups
  report n-of 50 group-structure
end
```

Listing A.3: NetLogo code of a dynamic group structure

Similar to the "Small Worlds" model, which is available from the NetLogo commons (`http://ccl.northwestern.edu/netlogo/models/SmallWorlds`), the procedure of listing A.3 creates 100 turtles that are embedded in a perfect 1-lattice (see also section 7.1). After the global variable `group-structure` has been created, the command `ql:set-group-structure` is not called. Instead, the function `get-groups` is implemented.

The `NetLogoSupervisor` expects the function `get-groups` to return a list of `NLGroups`. These groups are either created in this function or in the setup. If created in the setup, `ql:init` must be called before the groups are created (see listing A.3). In the example, the function `get-groups` is used to randomly select 50 groups at each round.

In the present implementation, not every agent has exactly one choice at any given round. Some agents may have two choices, and some agents may have no choice. But, on average, an agent has one choice per round. A more accurate procedure is possible. But `get-groups` is one of the "performance bottlenecks" of the simulation. It should, therefore, be implemented as simple as possible.

A.2 The games-extension

The games-extension provides a convenient way to define normal-form game-theoretic situations. Optimal points and Nash equilibria are calculated and returned to NetLogo in a well-arranged form. The games-extension is installed by creating a directory named `games` in the `extensions` subdirectory of the `NetLogo` program. All files from

`https://github.com/JZschache/NetLogo-games/tree/master/extensions/games`

have to be downloaded and moved to the newly created directory `extensions/games`. For example:

```
git clone https://github.com/JZschache/NetLogo-games.git
mv NetLogo-games/extensions/games path-to-netlogo/extensions
```

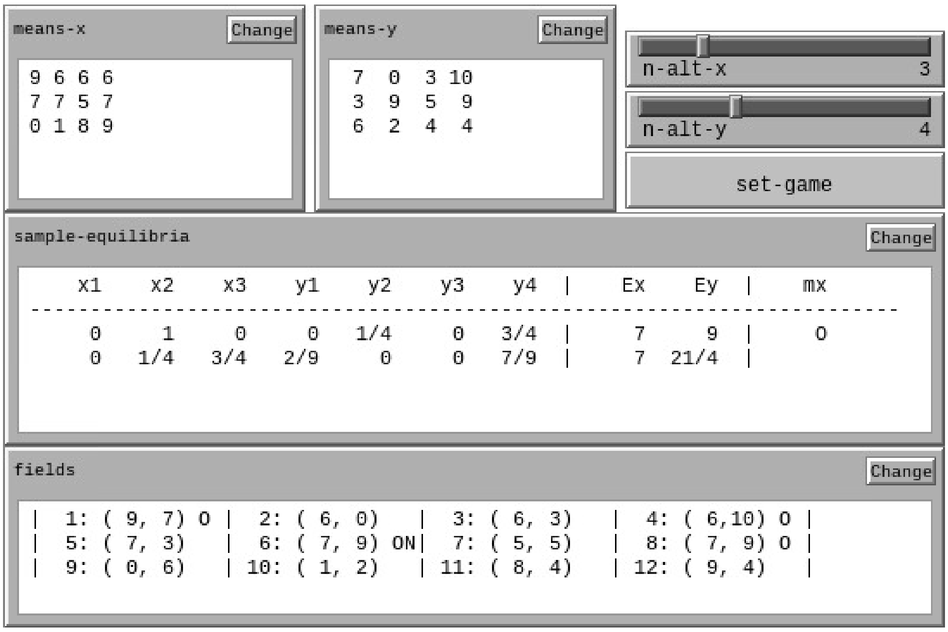
If the games-extension is used in combination with the ql-extension, the jars `games.jar` and `gamut.jar` must be added to the variable `additional-jars` in the file `extensions/ql/application.conf`:

```
netlogo {
  ...
  parallel {
    ...
    # all additional jars that must be loaded by NetLogo
    additional-jars = ["extensions/games/games.jar",
                       "extensions/games/gamut.jar"]
    ...
  }
}
```

This holds true for every extension that is used with the ql-extension. In the next two sections, features of the games-extension are presented, and some details about the calculation of Nash equilibria and optima are given.

A.2.1 Defining two-person games

Given the games-extension, a two-person game can be defined manually or by a predefined name. The first way is demonstrated with the help of figure A.7.



means-x

```
9 6 6 6
7 7 5 7
0 1 8 9
```

means-y

```
 7  0  3 10
 3  9  5  9
 6  2  4  4
```

n-alt-x 3

n-alt-y 4

sample-equilibria

```
    x1    x2    x3    y1    y2    y3    y4  |    Ex    Ey  |    mx
------------------------------------------------------------------------
     0     1     0     0   1/4     0   3/4  |     7     9  |     O
     0   1/4   3/4   2/9     0     0   7/9  |     7  21/4  |
```

fields

```
|  1: ( 9, 7) O |  2: ( 6, 0)   |  3: ( 6, 3)   |  4: ( 6,10) O |
|  5: ( 7, 3)   |  6: ( 7, 9) ON|  7: ( 5, 5)   |  8: ( 7, 9) O |
|  9: ( 0, 6)   | 10: ( 1, 2)   | 11: ( 8, 4)   | 12: ( 9, 4)   |
```

Figure A.7: NetLogo interface of the games-extension

In figure A.7, two NetLogo input fields named `means-x` and `means-y` are seen. Each field contains the mean rewards for player `x` or player `y`, respectively, given the choices of both players. Player `x` is the row-player in both fields. In order to create a game from the two input fields, two *game-matrices* must be created by the reporter `games:matrix-from-row-list`. This function takes a list of lists of numbers as parameter, which is, for example, created by the following reporter:

```
 1 to-report read-means-matrix [ nr ]
 2   let row-list []
 3   let temp means-x
 4   if (nr = 2) [set temp means-y]
 5   while [temp != ""] [
 6     let line-break position "\n" temp
 7     ifelse line-break = false [
 8       set row-list lput temp row-list
 9       set temp ""
10     ] [
11       set row-list lput (substring temp 0 line-break) row-list
12       set temp substring temp (line-break + 1) (length temp)
13     ]
14   ]
15   report (map [read-from-string (word "[ " ? " ]")] row-string-list)
16 end
```

From line 2 until line 14, a list of strings is created and saved to the local variable `row-list`. Each string is a row of the input field. Afterwards, Netlogo's reporter `read-from-string` is deployed to get the required list of lists of numbers.

Using the reporter of the previous listing, a game is defined by the commands `games:matrix-from-row-list` and `games:two-persons-game`:

```
to set-game
  let m1 games:matrix-from-row-list (read-means-matrix 1)
  let m2 games:matrix-from-row-list (read-means-matrix 2)
  let game games:two-persons-game m1 m2
end
```

The second way of creating a two-person game requires only a name and, occasionally, the numbers of alternatives for both players:

```
let game games:two-persons-gamut-game game-name n-alt-x n-alt-y
```

The reporter `games:two-persons-gamut-game` is based on the Gamut library (`http://gamut.stanford.edu`). Gamut makes over thirty games, which are commonly found in the economic literature, available (for details see the documentation: `http://gamut.stanford.edu/userdoc.pdf`). The games-extension currently supports the following parameters as name of a game:

- "BattleOfTheSexes"
- "Chicken"
- "CollaborationGame"
- "CoordinationGame"
- "DispersionGame" (considers first number of alternatives)
- "GrabTheDollar" (considers first number of alternatives)
- "GuessTwoThirdsAve" (considers first number of alternatives)
- "HawkAndDove"
- "MajorityVoting" (considers first number of alternatives)
- "MatchingPennies"
- "PrisonersDilemma"
- "RandomGame" (considers both numbers of alternatives)
- "RandomZeroSum" (considers both numbers of alternatives)
- "RockPaperScissors"
- "ShapleysGame"

It should be noted that some of these names do not generate the commonly expected game. For example, the structure of a "HawkAndDove" game resembles a prisoner's dilemma instead of a game of chicken.

By default, the minimum and maximum payoff is set to zero and ten, respectively. The values are specified in the configuration file `application.conf`.

Some of the games take the numbers of alternatives into account. They are given as additional parameters to `games:two-persons-gamut-game`. Only the "RandomGame" and the "RandomZeroSum" do not require that these numbers match. The other games consider only the first of the two parameters. Finally, the games-extension uses integers for the reward matrices. The values are scaled by changing the parameter `int-mult` in `application.conf` (see also `http://gamut.stanford.edu/userdoc.pdf`, p. 3).

After a game has been defined via Gamut, the input fields and sliders of the NetLogo interface can be updated as demonstrated in the following listing:

```
1  to-report write-means-matrix [ matrix ]
2    let strings games:matrix-as-pretty-strings matrix
3    let result ""
4    foreach strings [
5      set result (word result (reduce [(word ?1 " " ?2 )] ?) "\n")
6    ]
7    report result
8  end
9
10 to set-game
11   let game games:two-persons-gamut-game game-name n-alt-x n-alt-y
12   let m1 games:game-matrix game 1
13   let m2 games:game-matrix game 2
14   let m-strings games:matrix-as-pretty-strings m1 "   "
15   set n-alt-x length m-strings
16   set n-alt-y length first m-strings
17   set means-x write-means-matrix m1
18   set means-y write-means-matrix m2
19   set sample-equilibria games:get-solutions-string game "   "
20   set fields games:get-fields-string game "   "
21 end
```

The two game-matrices are obtained by `games:game-matrix` (lines 12 and 13). The first matrix is used to update the numbers of alternatives `n-alt-x` and `n-alt-y`. Since these values are not directly available, the matrix is converted into a list of lists of strings by `games:matrix-as-pretty-strings`. The strings are "pretty" because it is accounted for differences in length of the numbers by inserting offsets. In the example, two spaces (`"  "`) are inserted before every number with only one character. Consequently, the entries of each column are displayed right-aligned in `means-x` or `means-y`. The reporter `write-means-matrix` maps the list of lists of strings into one string. The NetLogo globals `sample-equilibria` and `fields` (figure A.7) are updated similarly by special commands of the games-extension. Some available commands are described in the following list:

- `games:matrix-transpose` takes a games-matrix as parameter and returns the transpose of this matrix. This reporter assists when defining symmetric games. The input matrix must be quadratic.
- `games:get-reward` returns an entry of a games-matrix. Therefore, three parameters are required: the matrix, a row index, and a column index.
- `games:get-solutions-string` can be used to update a NetLogo input field (see listing above and figure A.7). It prints (strictly) mixed Nash equilibria (if some are found). It also prints the expected reward of each player and indicates, by an `O` in the last column, whether a solution is (Pareto) optimal compared to the other (pure and mixed) solutions. Similar to `games:matrix-as-pretty-strings`, the second parameter is used to adjust the alignment.
- `games:get-fields-string` can be used to update a NetLogo input field (see listing above and figure A.7). It prints a joint payoff matrix. Each field of the matrix contains an index and the mean rewards as specified by the game. It also indicates the pure Nash equilibria (`N`) and pure (Pareto) optima (`O`). Similar to `games:matrix-as-pretty-strings`, the second parameter is used to adjust the alignment.
- `games:pure-solutions` returns a list of boolean values, one for each field of the joint payoff matrix (as given by `games:get-fields-string`). The boolean value indicates whether this field is pure Nash equilibria.
- `games:pure-optima` returns a list of boolean values, one for each field of the joint payoff matrix (as given by `games:get-fields-string`). The boolean value indicates whether this field is (Pareto) optimal compared to the other (pure and mixed) solutions.

A.2.2 The calculation of Nash equilibria and optima

While pure Nash equilibria are easily identified, the search for a Nash equilibrium of a normal-form game is, in general, computationally intensive (see e.g. Shoham and Leyton-Brown, 2009, ch. 4). More specifically, the problem of finding a Nash equilibrium of a general-sum finite game with two players is PPAD-complete (Daskalakis et al., 2009). For the class of PPAD-complete problems, it is known that at least one solution (Nash equilibrium) exists. But, to the best of the current knowledge, there exists no algorithm that is guaranteed to find this solution in polynomial time (e.g. Papadimitriou, 2014, p. 15884). Instead, the computation time of known algorithms increases exponentially with the number of alternatives.

Nevertheless, existing algorithms run efficiently in practice (e.g. Codenotti et al., 2008). One of the better known (but not the fastest) one (Shoham and Leyton-Brown, 2009, p. 91) is the Lemke-Howard algorithm (Lemke and Howson, 1964). This algorithm is implemented in the games-extension (as given by Codenotti et al., 2008). Even though the Lemke-Howard algorithm necessarily finds a Nash equilibrium, it is generally not able to find all equilibria (Shoham and Leyton-Brown, 2009, p. 98). The implementation of the games-extension tries to find multiple equilibria by starting the algorithm with every possible variable that can be part of the solution (see the pseudocode in Shoham and Leyton-Brown, 2009, p. 96). This step is repeated for every solution that has already been calculated. Since not all Nash equilibria are found, the input field of the NetLogo interface was named `sample-equilibria`.

Furthermore, the problem of stating whether a Nash equilibrium is also Pareto optimal is NP-hard (Shoham and Leyton-Brown, 2009, p. 102). In other words, this problem is “as hard as it can get”. The difficulty of this problem stems from the necessity to search an infinite space of possible outcomes. With a finite set of outcomes, the search for the optimal ones can be completed in polynomial time (and, on average, even in linear time, Godfrey et al., 2007). Consequently, the games-extension inspects only the pure and mixed Nash equilibria that are found directly or by the Lemke-Howard algorithm. The labelling of an outcome by an `O` must, hence, be understood relatively to the outcomes that are shown.

Appendix B

Sensitivity analysis

In this appendix, the simulation results of the previous chapters are tested for robustness. More specifically, it is checked whether the results are sensitive to small changes in the parameters. It is focussed on the melioration learning model. The Roth-Erev model is not tested for robustness.

B.1 Melioration learning in two-person games

In chapter 6, various two-person games were analysed. Results of simulations with different values of the exploration rate are shown in the following. It is also checked whether a decay in the exploration rate changes the outcome. In this case, algorithm 5.2.1 is used, and the exploration rate decreases at logarithmic speed:

$$\varepsilon_s \leftarrow \frac{\varepsilon}{\log\left(2 + \sum_{j \in E} K_t(s, j)\right)}.$$

This ensures that all alternatives are chosen sufficiently often during a simulation.

Additionally, the softmax/Boltzmann exploration method is tested. This method has one parameter $\tau \in (0, \infty)$, which is called *temperature*. The probability $q_t(e)$ of choosing alternative $e \in E$ at time $t \in \mathbb{N}$ is given by

$$q_t(e) = \frac{e^{\frac{Q_t(e)}{\tau}}}{\sum_{j \in E} e^{\frac{Q_t(j)}{\tau}}}.$$

If τ is large, the probabilities of all alternatives are approximately equal. With a decreasing τ, alternatives with high Q-value are preferred. In the limit to zero, softmax approaches the greedy selection of alternatives (Sutton and Barto, 1998, pp. 30-31). In the simulations, this limit is achieved by decreasing the temperature logarithmically (see also Singh et al., 2000, p. 303).

B.1.1 Games with dominant alternatives

In section 6.3, the following prisoner's dilemma was studied:

	A	B
A	(5,5)	(2,10)
B	(10,2)	(0,0)

The simulations of this game are repeated for different rewards of the outcome (A, A), for different exploration rates, and with decay in exploration rate. The results are presented in figure B.1.

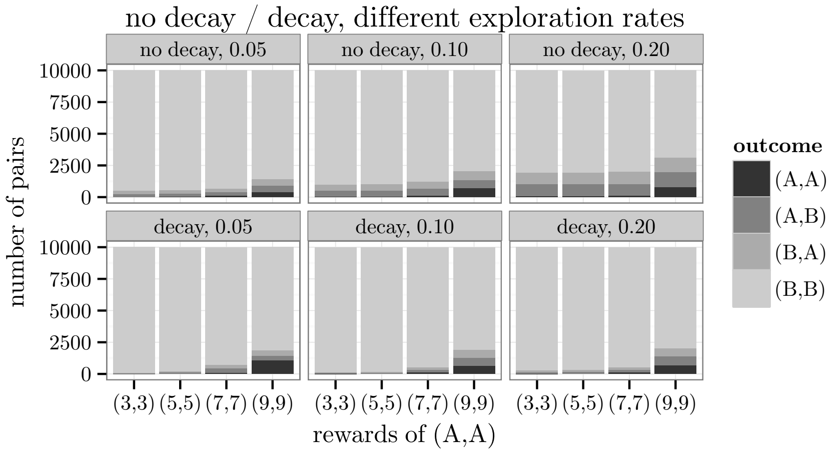


Figure B.1: Sensitivity analysis of the prisoner's dilemma

The frequency of the outcome (A, A) slightly increases with its expected reward. Since the outcomes (A, B) and (B, A) appear only because of exploration, they occur less often if decay is enabled. Generally, the dominant alternatives are chosen most of the time. This remains true in the case of softmax exploration, unless the temperature is unrealistically high and without decay (figure B.2).

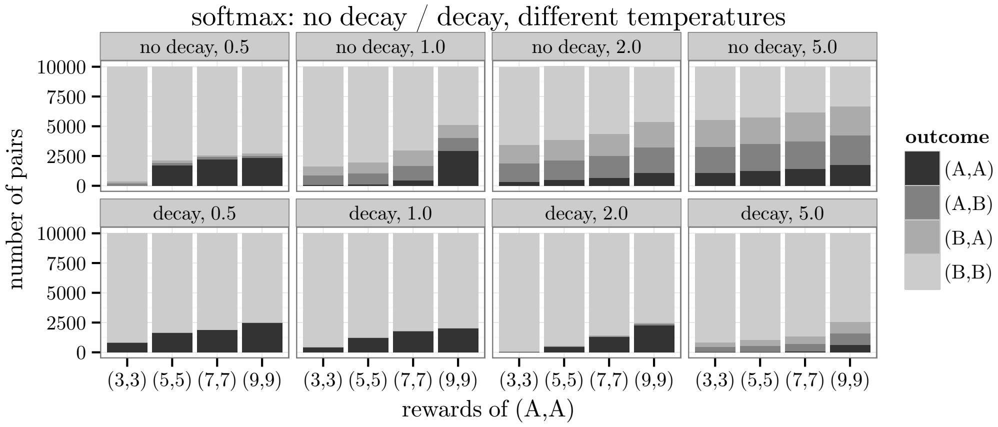


Figure B.2: Sensitivity analysis of the prisoner's dilemma: softmax

B.1.2 Games with multiple pure equilibria

The first game of section 6.3.2 was defined by the following reward matrix:

$$\begin{array}{c|cc} & A & B \\ \hline A & (10,10) & (0,0) \\ B & (0,0) & (8,8) \end{array}$$

Figure B.3 should be compared to figure 6.5. It shows the relationship between the rewards of outcome (B, B) and the frequency distribution of pairs of agents.

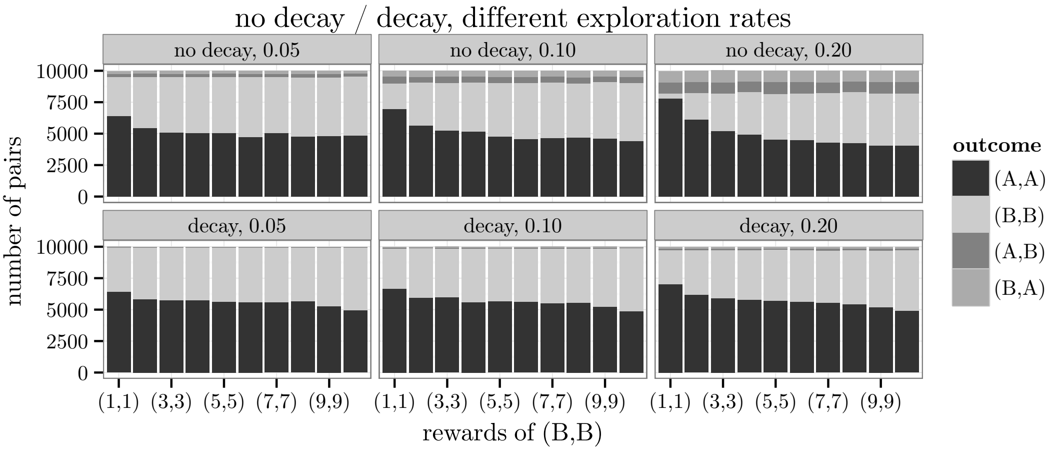


Figure B.3: Sensitivity analysis of the coordination game

A higher exploration rate comes with higher frequencies of non-equilibria and with an increased slope of the relationship between rewards and frequencies. In case of a decaying exploration rate, there is no visible effect of different exploration rates, and the relationship between rewards and frequencies diminishes.

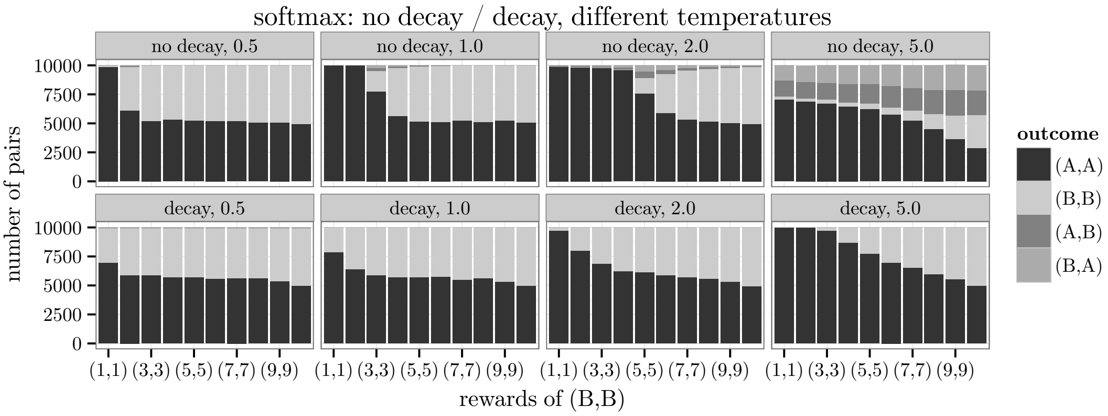


Figure B.4: Sensitivity analysis of the coordination game: softmax

If the agents select alternatives by softmax, the results are similar (figure B.4). There is a positive correlation between the rewards of (B, B) and its frequency. If the temperature is very high ($\tau = 5.0$) and there is a decay in temperature, the agents are most likely to find the optimal Nash equilibrium (A, A). But without the decay, also non-equilibria are chosen frequently.

Figure B.5 contains the results of a sensitivity analysis in regard to the game of chicken of figure 6.7:

	A	B
A	(5,5)	(2,10)
B	(10,2)	(0,0)

Different exploration rates or a decay in this rate only marginally affect the results of section 6.3.2. However, in case of decay and a high reward of (A, A), several agents end up in (A, A), which is an optimum but no equilibrium. The same tendency is found if the agents use softmax as selection rule. Even with medium rewards of (A, A), this outcome is frequently observed. Since (B, B) occurs as well, a similarity between softmax and the Roth-Erev model is indicated (see figure 6.7).

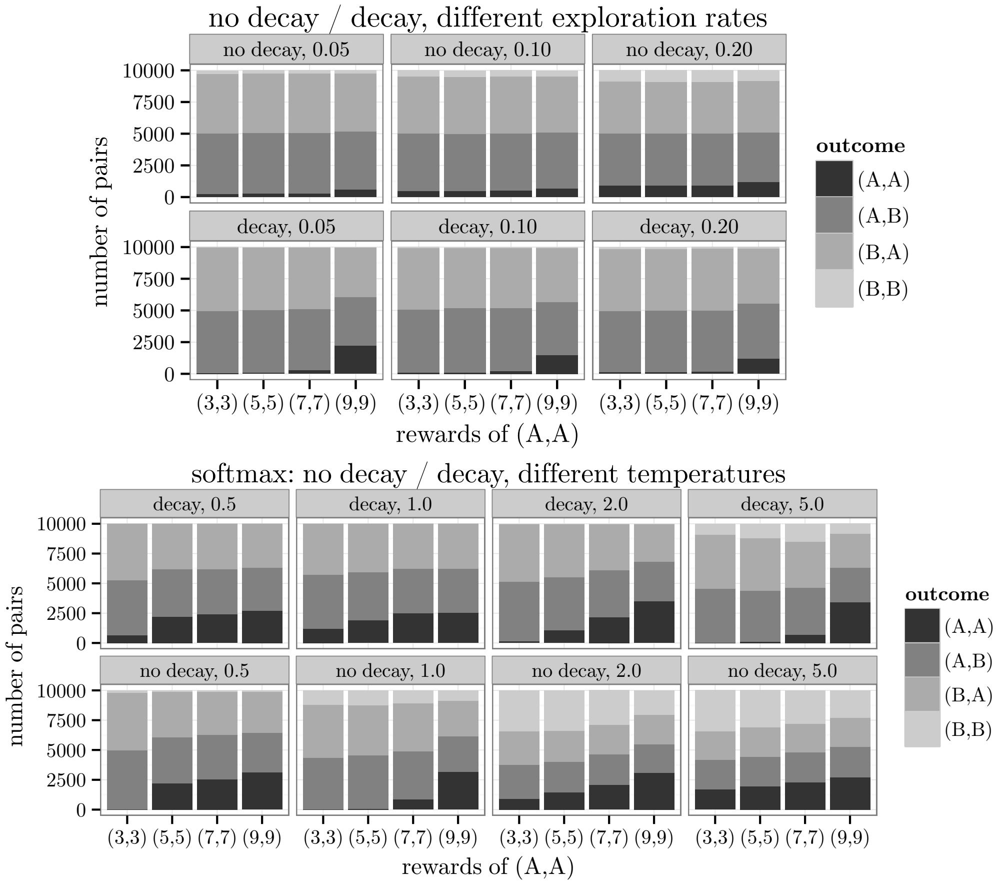


Figure B.5: Sensitivity analysis of the game of chicken

B.1.3 Games with a single Nash equilibrium

In games without a pure Nash equilibrium, a sufficiently high exploration rate is needed for the agents to approach the matching law. This is pictured in figure B.6 for the game "matching pennies" (figure 6.9) and the inspection game (figure 6.11):

		y	
		A	B
x	A	(10,0)	(0,10)
	B	(0,10)	(10,0)

matching pennies

		y	
		A	B
x	A	(8,15)	(15,10)
	B	(10,5)	(10,10)

inspection game

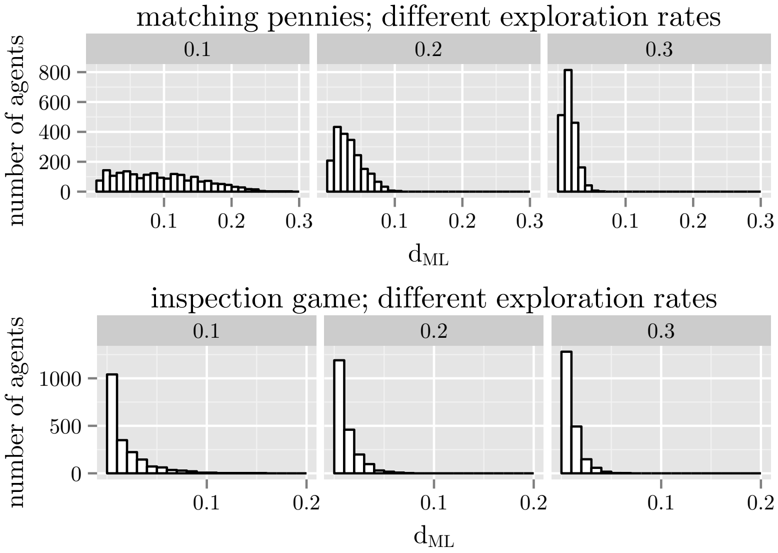


Figure B.6: Matching law measures of "matching pennies" and inspection game

The same effect is visible in figure B.7, which shows frequency distributions over the relative frequencies of alternative A. A high exploration rate implies a low variance in relative frequencies. A low variance means that the behaviour has already converged to the mixed Nash equilibrium.

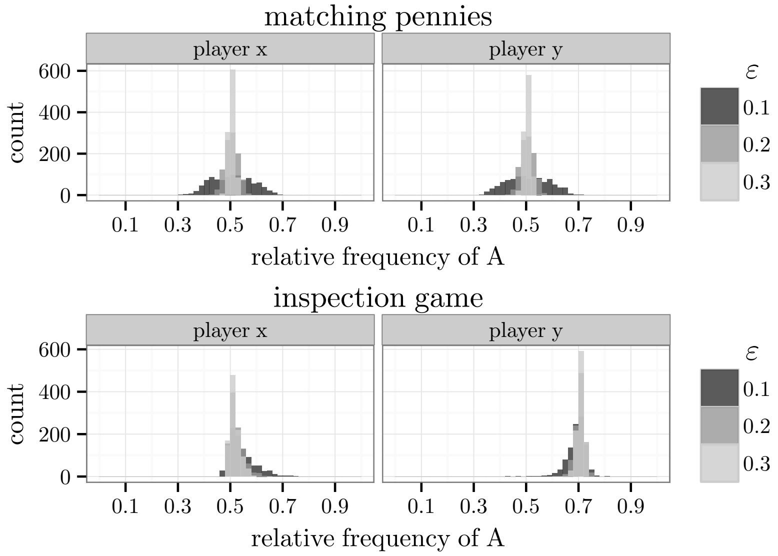


Figure B.7: Relative frequencies of "matching pennies" and inspection game

Since a sufficiently high exploration rate is beneficial in games without pure Nash equilibria, a decay in exploration rate generally prevents the convergence to the mixed Nash equilibrium.

Finally, figure B.8 provides proof that neither a high exploration rate nor a decay in exploration can help the players of Shapley's game to approach a stable point. The behaviour does not converge, regardless of the level of exploration.

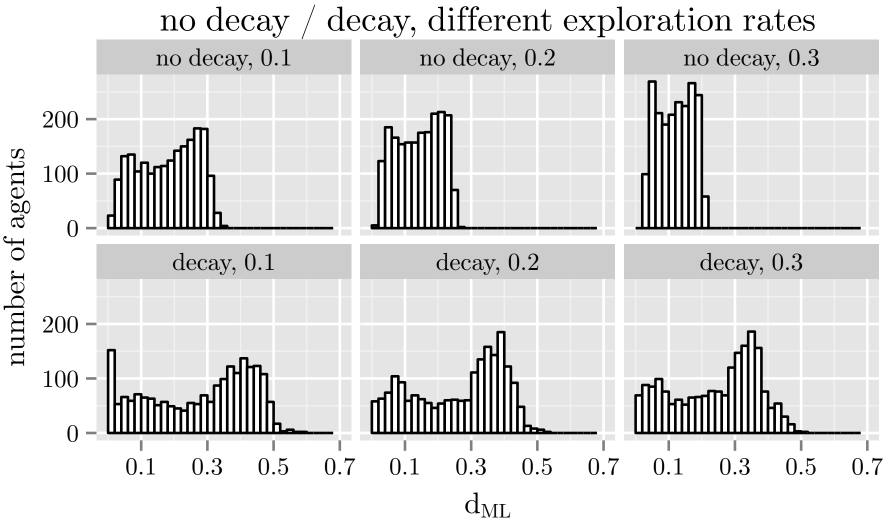


Figure B.8: Sensitivity analysis of Shapley's game

B.2 Melioration learning in n-way games

It was shown in section 7.1 that agents who learn by melioration are able to coordinate their choices in n-way games. It was stated that this depends on the reward b of the second equilibrium (B, B) (figure 7.2). In the simulations of figure B.9, this result is tested for robustness by assuming different exploration rates $\varepsilon \in \{0.05, 0.1, 0.2\}$. The histograms are reduced to three intervals: $[0, 0.1]$, $(0.1, 0.9]$, and $(0.9, 1]$. Similar to figure 7.2, the agents' ability to coordinate their choices depends on the reward b. The establishment of a convention is impeded by a high exploration rate because there is a higher variance in decisions. Also the effect of a larger number of connections or a higher level of randomness (figures 7.4 and 7.5) is less pronounced if exploration is high (not shown here).

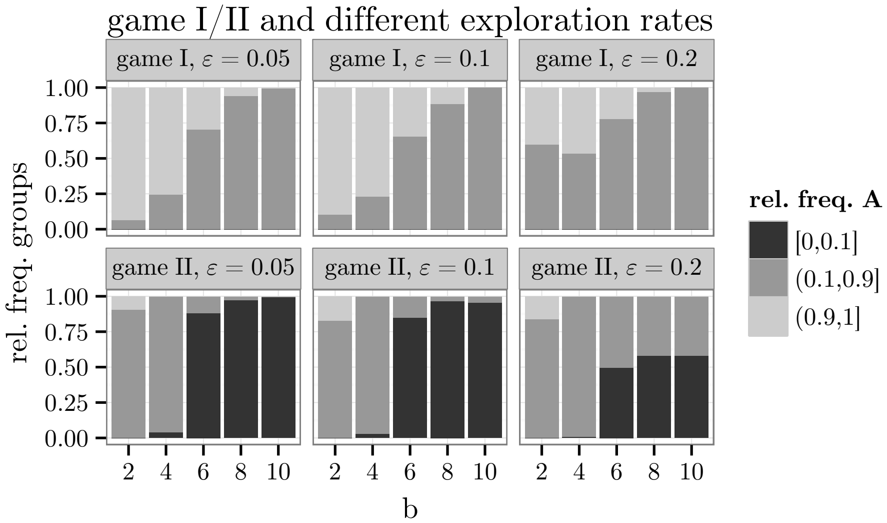


Figure B.9: Networks with $\overline{d} = 2$, $\beta = 0$, and $n = 50$

In the simulations of section 7.1, every group consisted of 50 agents. A smaller or greater group size n does not change the results (figure B.10). This in line with a statement of Young (1998, pp. 101-102): the speed of convergence to a risk-dominant equilibrium is independent of the number of vertices if the network is *close knit* to a certain degree. The networks of the simulations are polygons and, therefore, satisfy this condition (Young, 1998, p. 101).

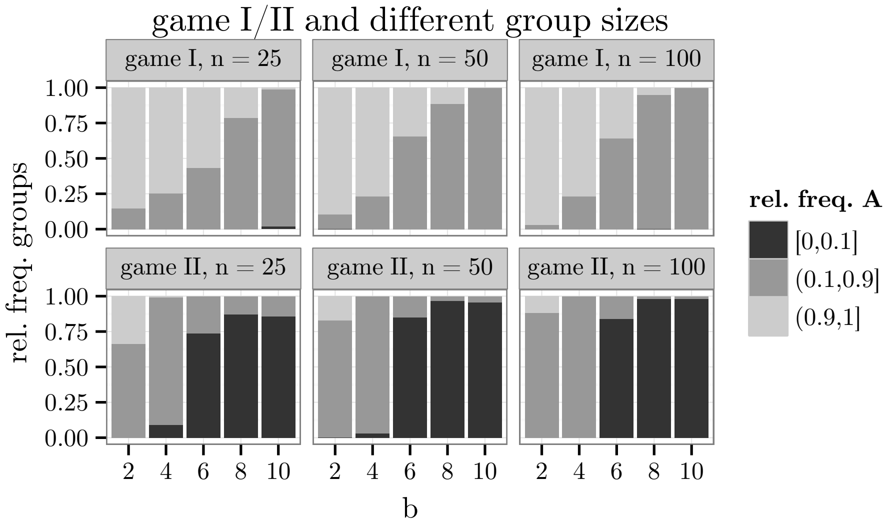


Figure B.10: Networks with $\overline{d} = 2$, $\beta = 0$, and $\varepsilon = 0.1$

However, in games without risk-dominant equilibrium, the group size is relevant. The relationship between average number of connections $\overline{d}$ and distribution of choices is pictured in figure B.11 for different group sizes n. Only the extreme case of game I with two equally appealing equilibria ($b = 10$) is shown. There is no significant effect in game II with $b = 4$. As stated in section 7.1, a high number of neighbours supports the agreement on a single alternative. But this depends on the size of the group. The larger the group, the more connections between the agents are needed for the coordination of decisions.

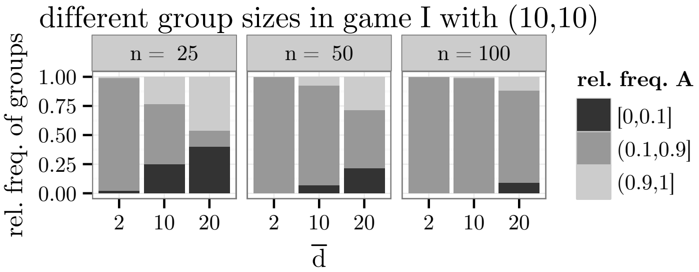


Figure B.11: Networks with $\beta = 0$ and $\varepsilon = 0.1$

Furthermore, in case of $\beta > 0$ and small $\overline{d}$, a network may be disconnected. This means that some agents cannot be reached by other agents via a sequence of edges. It is seen in figure B.12 that no convention is established in game I with $b \in \{8, 10\}$ and $\overline{d} = 2$. In comparison to figure 7.5, the relationship between the randomness β and the distribution of choices is stronger for $\overline{d} = 20$ than for $\overline{d} = 10$. This corresponds to the result of section 7.1: a large number of connections or a high level of randomness supports the establishment of a convention.

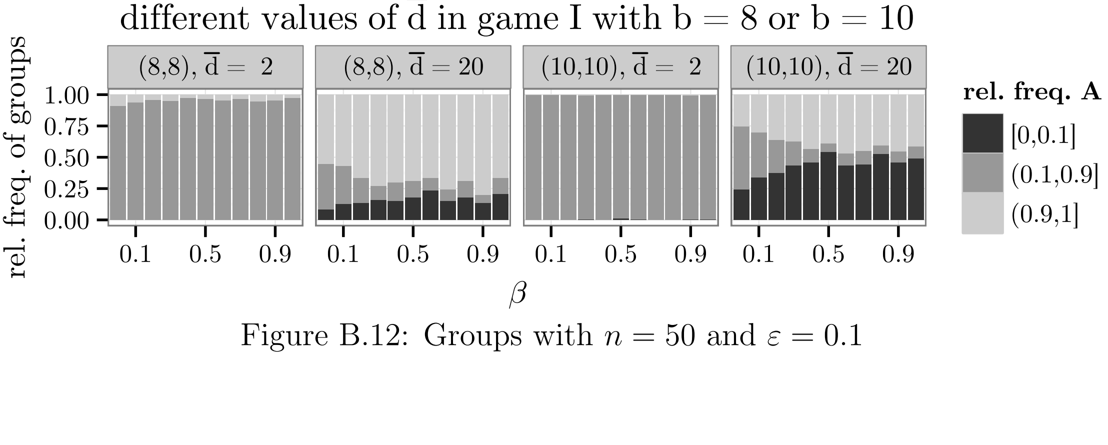


Figure B.12: Groups with $n = 50$ and $\varepsilon = 0.1$

B.3 Melioration and the volunteer's dilemma

In section 7.2, the volunteer's dilemma was analysed. Figure B.13 shows that, in simulations with fixed groups, the difference between simulation results and the inverse of the group size increases with the exploration rate.

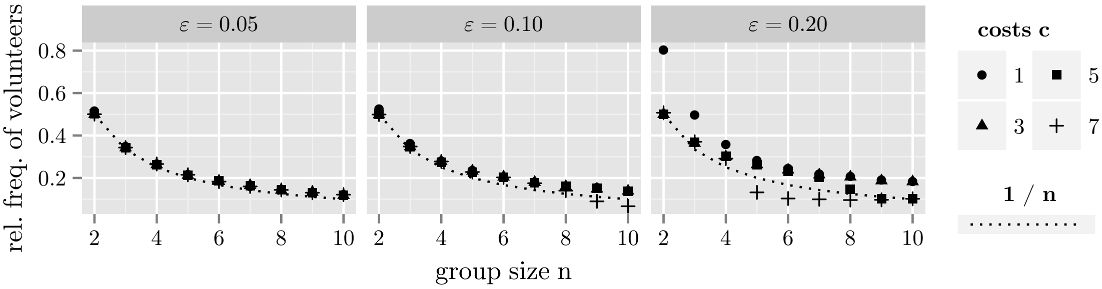


Figure B.13: The effect of group size in fixed groups

In reference to figure 7.6, additional data is presented in figure B.14. It contains histograms of the individual relative frequencies of volunteering. Agents are partitioned into three classes: agents who mainly volunteer (the interval $(0.9, 1]$), agents who almost never volunteer (the interval $[0, 0.1]$), and agents who occasionally volunteer (the interval $(0.1, 0.9]$). The plots indicate that, except for the case of high costs of volunteering, most agents either always volunteer or are idle during the whole simulation.

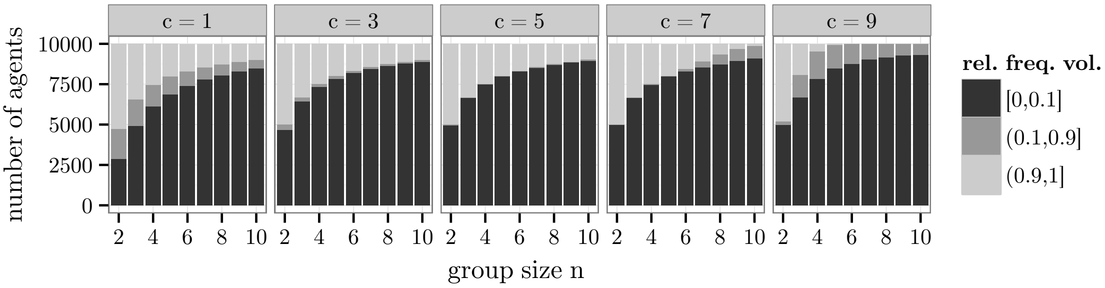


Figure B.14: Histograms of individual frequencies; $\varepsilon = 0.1$

Further findings of the sensitivity analysis are omitted because they show that the results of section 7.2 are robust in regard to different exploration rates.

B.4 Melioration and the prisoner's dilemma

In section 7.3, a relationship between group size and incentive of cooperation $\frac{r}{n}$, on the one hand, and the rate of cooperation, on the other hand, was observed.

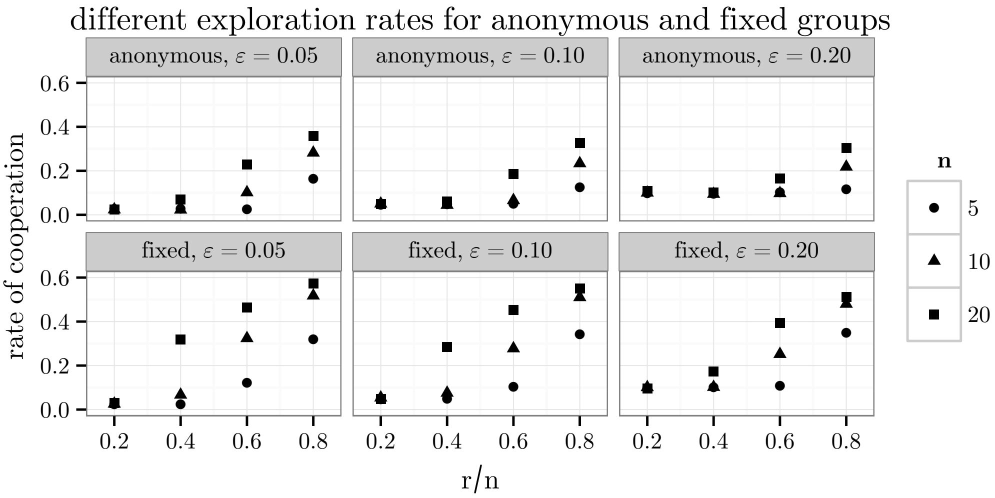


Figure B.15: The rate of cooperation in the public goods game

Figures B.15 and B.16 show that this result is also found in simulations with different exploration rates and in simulations with a decay in exploration.

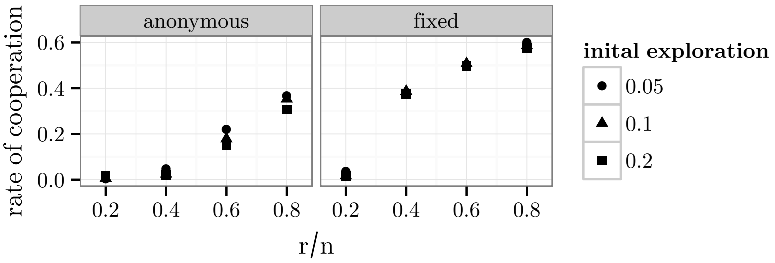


Figure B.16: The rate of cooperation with decay in exploration; $n = 20$.

In the simulations with punishment, a penalty of $s = 1$ was assumed. Figure B.17 displays the relative frequencies of cooperation if $s \in \{0.1, 0.5, 1\}$. The frequencies of cooperation refer to agents who cooperate, including those who carry out the punishments. The costs of punishment are specified in relation to the penalty as c/s. The horizontal lines mark the levels of cooperation without punishment. It is evident from figure B.17 that the rate of cooperation is low if the penalty s is small. This finding confirms the hypothesis that high levels of cooperation are due to accidental punishments of fellow group members.

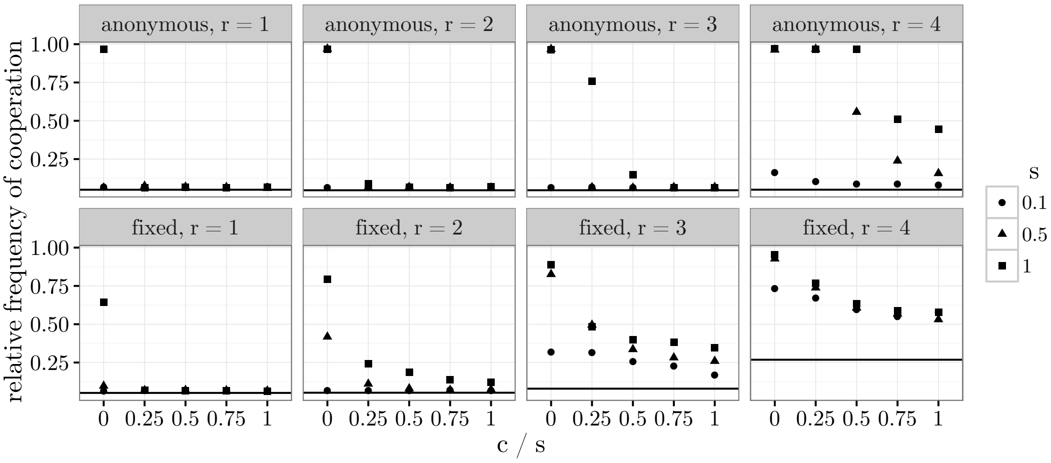


Figure B.17: Frequencies at the 25 000th round with punishment; $n = 5$

The effect of punishment is intensified in large groups (figure B.18). The larger the group, the higher the number of accidental punishments. This decreases the Q-value of defection. In many conditions, this results in a cooperation rate of 0.93. The agents deviate from cooperation only because of the exploration rate.

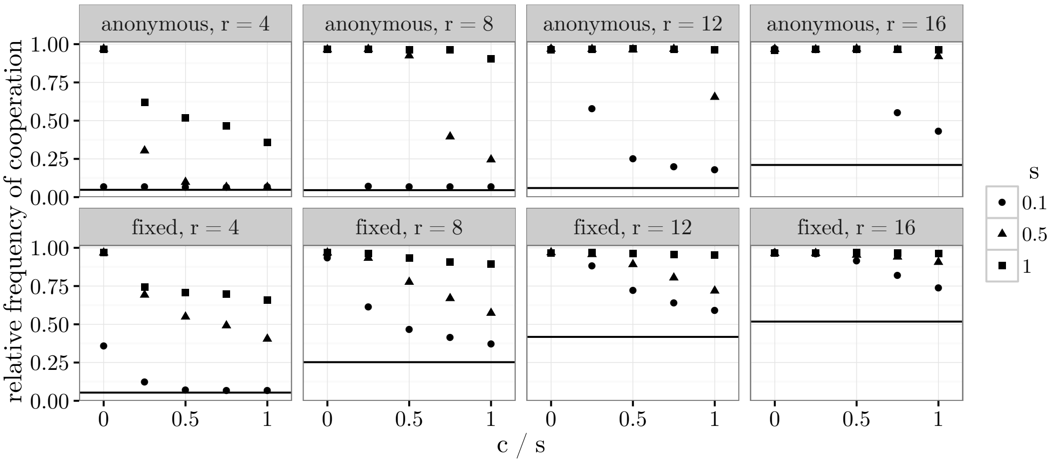


Figure B.18: Frequencies at the 25 000th round with punishment; $n = 20$

In the last part of section 7.3, the option to abstain from an interaction was analysed. First, simulations were run with anonymous groups. Similar to figure 7.18, figure B.19 shows the development of the frequencies of choice. There is no visible pattern, but medium levels of cooperation are achieved. In large groups, the dynamic is more stable than in small groups.

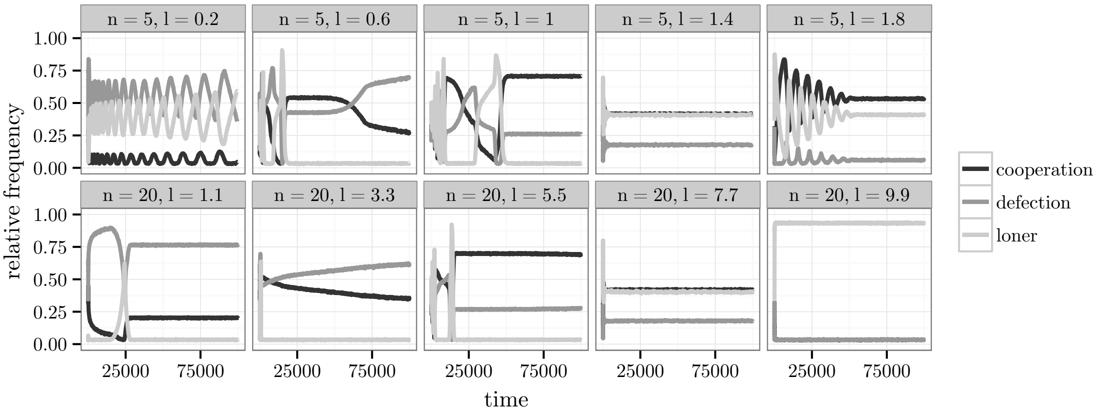


Figure B.19: Frequencies over time in anonymous groups; $\frac{r}{n} = 0.6$

In simulations with fixed groups, the rates of choice are stable over time (figure B.20). The level of cooperation increases with the loner payment l. In the simulation with $n = 20$ and $l = 9.9$, this payment is too high, and all agents decide to be a loner. The maximum reward of cooperation is $r = 12$. Because of the exploration rate, this reward is never reached and it pays off to be a loner.

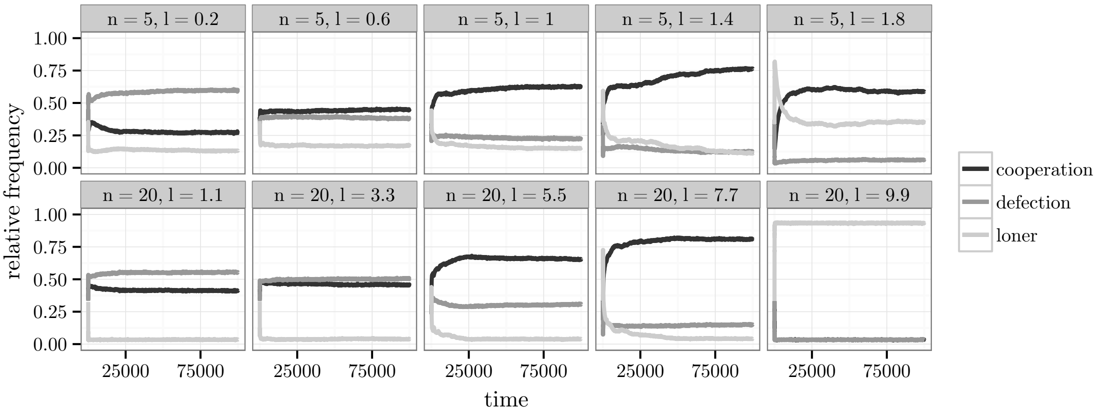


Figure B.20: Frequencies over time in fixed groups; $\frac{r}{n} = 0.6$

Finally, figure B.21 parallels the results of figure 7.19 but with $n = 20$. The effect of the loner option is less pronounced in large groups. On the one hand, this is due to the higher level of cooperation in simulations without this option (horizontal lines). On the other hand, it is less likely that $r > n$ if n is large.

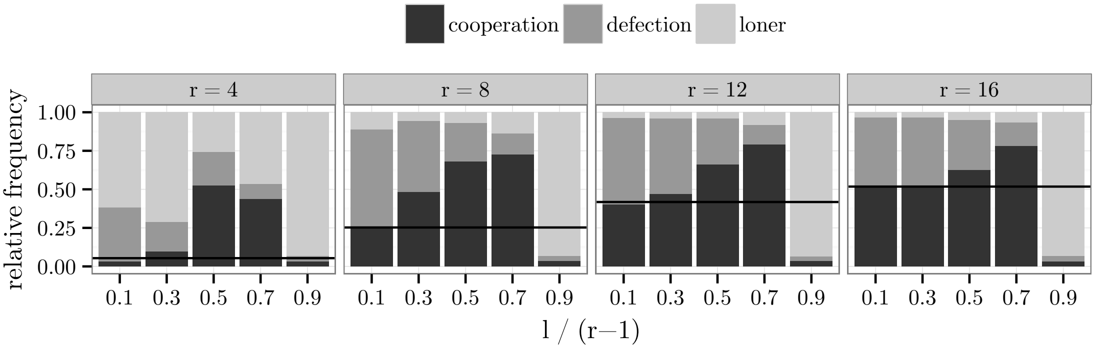


Figure B.21: Frequencies at the 25 000th round in fixed groups; $n = 20$

B.5 The foraging model

In this section, a sensitivity analysis of the findings of chapter 8 is presented.

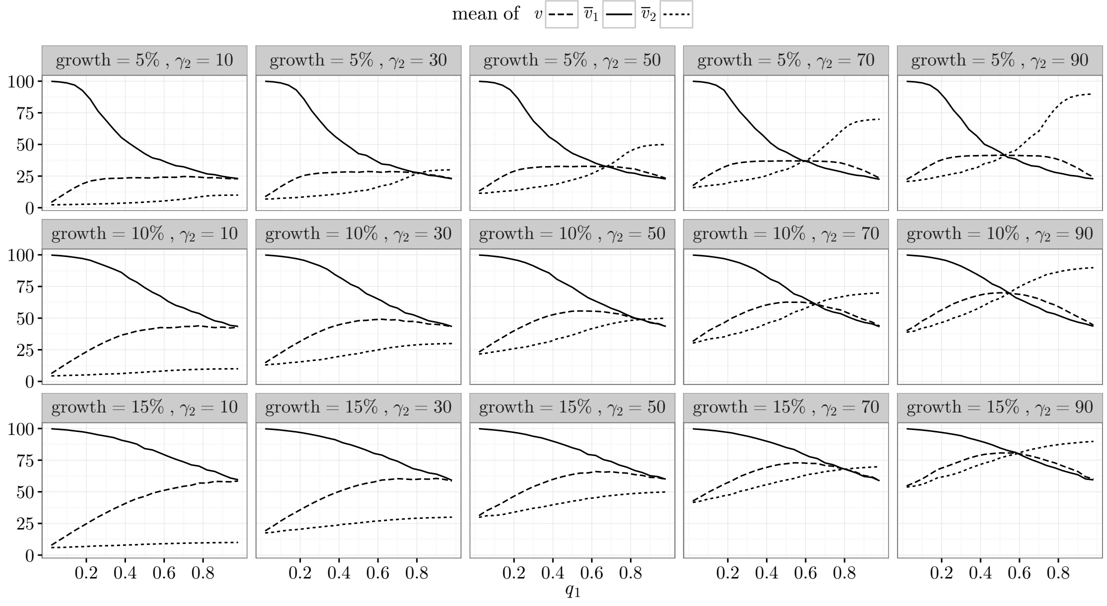


Figure B.22: The foraging model with different growth rates and values of γ_2

The plots of figure B.22 extend the results of section 8.2.1 in regard to different growth rates and values of γ_2. All of the 1 000 actors follow the same rule $\boldsymbol{q} = (q_1, q_2)$, which specifies the probability of choosing the first and the second type of resource. The value of the first resource γ_1 is set to 100. The optimal point q_1^* of v is closer to 0.5 if the value of the second resource is high. The growth rate affects the optimal value of v and the shape of the curves. In case that all actors use the optimal rule q_1^*, the performance of a single actor is usually improved by deviating from this rule. Since $\overline{v}_1(q_1^*) > v(q_1^*)$ in many situations, the actor obtains a higher reward by choosing the first resource exclusively.

The following simulations replicate the simulations of section 8.3. Half of the actors are of type A and the other half of type B. The value of the first kind of resource is $\gamma_1 = 100$, and the growth rate is set to 10%. Figure B.23 contains the means of v, which are measured over a period of 5 000 time steps. It is distinguished between the two actor types, the level of uncertainty η, the value of the second kind of resource γ_2, and a fourth parameter, which is called the *waiting rate* $\psi \in [0, 1]$.

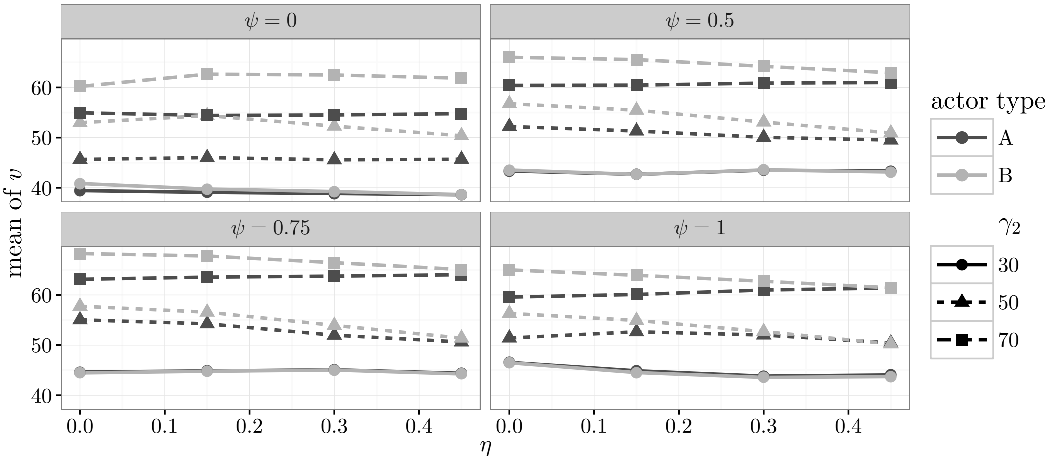


Figure B.23: The effect of uncertainty η for different γ_2 and ψ

In the simulations of section 8.3, the waiting rate was set to 1 (last plot of figure B.23). It is seen that the effect of uncertainty is present for every value of γ_2. Type B actors perform as well as type A actors if a high level of uncertainty distorts the perception of the state. In case that $\gamma_2 = 30$, all actors choose the first resource exclusively and achieve the same rewards.

A waiting rate of 1 is equivalent to the random replacement of the actors before every decision. If $\psi < 1$, the actors wait a random number of rounds before making the next decision. The number of rounds follows a geometric distribution with parameter $1 - \psi$. It is evident from figure B.23 that the effect of uncertainty is less pronounced if $\psi < 1$.

Moreover, in case that the actors do not wait ($\psi = 0$), uncertainty has no significant effect on success. This is due to a complex interaction between the actors. In all simulations, the density of actors is high. Therefore, it happens frequently that two actors choose the same resource. Since the actors learn to follow similar rules of decision-making, this also means that two actors often follow the same path over several rounds. In this way, the actors impede each other. It would be better to wait a random number of rounds or to occasionally switch to another rule of decision-making.

Type B actors are able to change their rules of decision-making. They maintain two sets of Q-values, one for each state of the environment. One of the sets may advice to always choose the first resource and the other set to select the second resource. Independent of the actual state of the environment, it is beneficial to use two different kinds of selection mechanism and switch randomly between them. Hence, a high level of uncertainty is not harmful for type B actors if $\psi = 0$.

List of Symbols

$(0,1)$	$(0,1) = \{r \in \mathbb{R} \mid 0 < r < 1\}$
$(0,\infty)$	$(0,\infty) = \{r \in \mathbb{R} \mid r > 0\}$
$(p_j)_{j\in E}$	if $E = \{e_1, e_2, \dots, e_m\}$, $m \in \mathbb{N}$, then $(p_j)_{j\in E} = (p_{e_1}, p_{e_2}, \dots, p_{e_m})$
$[0,1]$	$[0,1] = \{r \in \mathbb{R} \mid 0 \leq r \leq 1\}$
$[0,\infty)$	$[0,\infty) = \{r \in \mathbb{R} \mid r \geq 0\}$
$\lvert S\rvert$	if S is a set, $\lvert S\rvert$ gives the cardinality of S
$\mathbb{N}$	the set of all strictly positive integers $\{1, 2, 3, \dots\}$
$\mathbb{R}$	the set of real numbers
$\mathbf{1}_{\{X=x\}}$	if X is a random variable: $\mathbf{1}_{\{X=x\}} = \begin{cases} 1 & \text{if } X = x, \\ 0 & \text{else} \end{cases}$
$\mathcal{P}$	$\mathcal{P} = \Big\{(p_{e_1}, p_{e_2}, \dots, p_{e_m}) \in [0,1]^E \;\mid\; \sum_{j\in E} p_j = 1\Big\}$
$\mathcal{X}$	$\mathcal{X} = [0,\infty)^E$
$\succsim$	$\succsim \subseteq \mathcal{X} \times \mathcal{X}$
$\{v_j\}_{j\in E}$	if $E = \{e_1, e_2, \dots, e_m\}$, $m \in \mathbb{N}$, then $\{v_j\}_{j\in E} = \{v_{e_1}, v_{e_2}, \dots, v_{e_m}\}$
$Pr(\cdot)$	the underlying probability measure
S^E	if $E = \{e_1, e_2, \dots, e_m\}$, $m \in \mathbb{N}$, then $S^E = \{(s_{e_1}, s_{e_2}, \dots, s_{e_m}) \mid s_j \in S \text{ for all } j \in E\}$

References

Abdallah, S. and V. Lesser (2006). Learning the task allocation game. In *AAMAS '06 Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, New York, NY, USA, pp. 850–857. ACM.

Alferink, L. A., T. S. Critchfield, and J. L. Hitt (2009). Generality of the matching law as a descriptor of shot selection in basketball. *Journal of Applied Behavior Analysis 42*(3), 595–608.

Alonso, E. and E. Mondragón (2006). Associative learning for reinforcement learning: Where animal learning and machine learning meet. In E. Alonso and Z. Guessoum (Eds.), *Proceedings of the Fifth Symposium on Adaptive Agents and Multi-Agent Systems*, Paris, France, pp. 87–99.

Antonides, G. and S. Maital (2002). Effects of feedback and educational training on maximization in choice tasks: Experimental-game evidence. *The Journal of Socio-Economics 31*(2), 155–165.

Arrow, K. J., H. B. Chenery, B. S. Minhas, and R. M. Solow (1961). Capital-labor substitution and economic efficiency. *The Review of Economics and Statistics 43*(3), 225–250.

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

Axelrod, R. (1987). The evolution of strategies in the iterated Prisoner's dilemma. In L. Davis (Ed.), *Genetic Algorithms and Simulated Annealing*, pp. 32–41. Morgan Kaufman Publishers.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. *Complexity 3*(2), 16–22.

Bacotti, A. V. (1977). Matching under concurrent fixed-ratio variable-interval schedules of food presentation. *Journal of the Experimental Analysis of Behavior 25*(1), 171–182.

Baldassarre, G. and D. Parisi (2000). Classical and instrumental conditioning: From laboratory phenomena to integrated mechanisms for adaptation. In R. Pfeifer (Ed.), *From animals to animats 6: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behaviour*, Complex adaptive systems, Cambridge, Mass. MIT Press.

Barron, G. and E. Yechiam (2002). Private e-mail requests and the diffusion of responsibility. *Computers in Human Behavior 18*(5), 507–520.

Barto, A. G., R. S. Sutton, and C. J. C. H. Watkins (1990). Learning and sequential decision making. In M. Gabriel and J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, Cambridge, Mass, pp. 539–602. MIT Press.

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior 22*(1), 231–242.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior 32*(2), 269–281.

Baum, W. M. (1981). Optimization and the matching law as accounts of instrumental behaviour. *Journal of the Experimental Analysis of Behavior 36*(3), 387–403.

Baum, W. M. and J. A. Nevin (1981). Maximization theory: Some empirical problems. *Behavioral and Brain Sciences 4*(3), 389–390.

Belinsky, R., F. González, and J. Stahl (2004). Optimal behavior and concurrent variable interval schedules. *Journal of Mathematical Psychology 48*(4), 247–262.

Belinsky, R., F. González, and J. Stahl (2005). Optimal behavior and concurrent variable ratio-variable interval schedules. *Journal of Mathematical Psychology 49*(4), 339–353.

Bellman, R. E. (1957). A Markov decision process. *Journal of Mathematics and Mechanics 6*(5), 679–684.

Bendor, J. (1987). In good times and bad: Reciprocity in an uncertain world. *American Journal of Political Science 31*(3), 531–558.

Bendor, J., D. Diermeier, and M. Ting (2007). Comment: Adaptive models in sociology and the problem of empirical content. *American Journal of Sociology 112*(5), 1534–1545.

Berger, R. and R. Hammer (2007). Die doppelte Kontingenz von Elfmeterschüssen. Eine empirische Analyse. *Soziale Welt 58*(4), 397–418.

Berninghaus, S. K. and U. Schwalbe (1996). Conventions, local interaction, and automata networks. *Journal of Evolutionary Economics 6*(3), 297–312.

Bianchi, F. and F. Squazzoni (2015). Agent-based models in sociology. *WIREs Computational Statistics 7*(4), 284–306.

Borrero, J. C., S. S. Crisolo, Q. Tu, W. A. Rieland, N. A. Ross, M. T. Francisco, and K. Y. Yamamoto (2007). An application of the matching law to social dynamics. *Journal of Applied Behavior Analysis 40*(4), 589–601.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences 100*(6), 3531–3535.

Boyd, R. and P. J. Richerson (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology 13*(3), 171–195.

Brandt, H., C. Hauert, and K. Sigmund (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London. Series B: Biological Sciences 270*(1519), 1099–1104.

Braun, N. and T. Gautschi (2011). *Rational-Choice-Theorie.* Weinheim und Basel: Beltz Juventa.

Brenner, T. (2006). Agent learning representation: Advice on modelling economic learning. In L. Tesfatsion and K. L. Judd (Eds.), *Handbook of Computational Economics. Agent-based Computational Economics*, Volume 2. North-Holland.

Brenner, T. and U. Witt (2003). Melioration learning in games with constant and frequency-dependent pay-offs. *Journal of Economic Behavior & Organization 50*(4), 429–448.

Burgess, R. L. and D. Bushell (Eds.) (1969). *Behavioral Sociology. The Experimental Analysis of Social Processes.* New York and London: Columbia University Press.

Bush, R. R. and F. Mosteller (1964). *Stochastic Models for Learning* (2nd ed.). New York: Wiley.

Buskens, V., R. Corten, and J. Weesie (2008). Consent or conflict: Coevolution of coordination and networks. *Journal of Peace Research 45*(2), 205 – 222.

Buskens, V. and C. Snijders (2015). Effects of network characteristics on reaching the payoff-dominant equilibrium in coordination games: A simulation study. *Dynamic Games and Applications*, 1–18. DOI 10.1007/s13235-015-0144-4.

Camerer, C. and T.-H. Ho (1999). Experience-weighted attraction learning in normal form games. *Econometrica 67*(4), 827–874.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princton, New Jersey: Princeton University Press.

Caron, P.-O. (2013a). On applying the matching law to between-subject data. *Animal Behaviour 85*(4), 857–860.

Caron, P.-O. (2013b). On the empirical status of the matching law: Comment on McDowell (2013). *Psychological Bulletin 139*(5), 1029–1031.

Caron, P.-O. (2015). Matching without learning. *Adaptive Behavior 23*(4), 227 – 233.

Chiappori, P.-A., S. Levitt, and T. Groseclose (2002). Testing mixed-strategy equilibrium when players are heterogeneous: The case of penalty kicks in soccer. *The American Economic Review 92*(4), 1138–1151.

Claus, C. and C. Boutilier (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI '98 Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pp. 746–752.

Codenotti, B., S. D. Rossi, and M. Pagan (2008, Nov). An experimental analysis of lemke-howson algorithm. `arXiv:0811.3247`.

Coleman, J. S. (1990). *Foundations of Social Theory.* Cambridge, Mass., and London, England: The Belknap Press of Harvard University Press.

Collett, D. (1999). *Modelling binary data.* Boca Raton: Chapman and Hall / CRC.

Conger, R. and P. Killeen (1974). Use of concurrent operants in small group research: A demonstration. *The Pacific Sociological Review 17*(4), 399–416.

Cook, K. S. and T. Yamagishi (1992). Power in exchange networks: a power-dependence formulation. *Social Networks 14*, 245–265.

Corrado, G. S., L. P. Sugrue, H. S. Seung, and W. T. Newsome (2005). Linear-nonlinear-Poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior 84*(3), 581–617.

Darley, J. M. and B. Latané (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology 8*(4), 377–383.

Daskalakis, C., P. W. Goldberg, and C. H. Papadimitriou (2009). The complexity of computing a Nash equilibria. *SIAM Journal on Computing 39*(1), 195 – 259.

de Villiers, P. A. and R. J. Herrnstein (1976). Toward a law of response strength. *Psychological Bulletin 83*(6), 1131–1153.

de Vos, H., R. Smaniotto, and D. A. Elsas (2001). Reciprocal altruism under conditions of partner selection. *Rationality and Society 13*(2), 139–183.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution 29*(4), 605 – 610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game. Theory and experimental evidence. *International Journal of Game Theory 22*(1), 75–85.

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review 67*(3), 297 – 308.

Dorris, M. C. and P. W. Glimcher (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron 44*(2), 365–378.

Dyer, M. G. (1995). Towards synthesizing artificial neural networks that exhibit cooperative intelligent behavior: Some open issues in Artificial Life. In C. G. Langton (Ed.), *Artificial Life. An Overview*, pp. 111–134. The MIT Press.

Emerson, R. M. (1969). Operant psychology and exchange theory. In R. L. Burgess and D. Bushell (Eds.), *Behavioral Sociology. The Experimental Analysis of Social Processes*, Chapter 17, pp. 379–405. New York: Columbia University Press.

Emerson, R. M. (1972a). Exchange theory, part i: A psychological basis for social exchange. In J. Berger, M. Zelditch, and B. Anderson (Eds.), *Sociological Theories in Progress*, Volume 2, Chapter 3, pp. 38–57. Boston: Houghton Mifflin Company.

Emerson, R. M. (1972b). Exchange theory, part II: Exchange relations and network structures. In J. Berger, M. Zelditch, and B. Anderson (Eds.), *Sociological Theories in Progress*, Volume 2, Chapter 4, pp. 58–87. Boston: Houghton Mifflin Company.

Erev, I. and A. E. Roth (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review 88*(4), 848–881.

Fantino, E., N. Squires, N. Delbrück, and C. Peterson (1972). Choice behavior and the accessibility of the reinforcer. *Journal of the Experimental Analysis of Behavior 18*(1), 35–43.

Feltovich, N. (2000). Reinforcement-based vs. belief-based learning models in experimental asymmetric-information games. *Econometrica 68*(3), 605–641.

Fishburn, P. C. (1970). *Utility theory for decision making.* New York: Wiley.

Flache, A. (2002). The rational weakness of strong ties: Failure of group solidarity in a highly cohesive group of rational agents. *The Journal of Mathematical Sociology 26*(3), 189–216.

Flache, A. and M. W. Macy (2002). Stochastic collusion and the power law of learning: A general reinforcement learning model of cooperation. *Journal of Conflict Resolution 46*(5), 629.

Foster, D. P. and R. V. Vohra (1993). A randomization rule for selecting forecasts. *Operations Research 41*(4), 704–709.

Frank, R. H. (2011). *The Darwin economy. Liberty, competition, and the common good.* Princeton and Oxford: Princeton University Press.

Franzen, A. (1995). Group size and one-shot collective action. *Rationality and Society 7*(2), 183–200.

Friedman, D., D. W. Massaro, S. N. Kitzis, and M. M. Cohen (1995). A comparison of learning models. *Journal of Mathematical Psychology 39*(2), 164–178.

Fudenberg, D. and D. K. Levine (1998). *The Theory of Learning in Games.* Cambridge, Massachusetts: The MIT Press.

Gallistel, C. R., T. A. Mark, A. P. King, and P. E. Latham (2001). The rat approximates an ideal detector of changes in rates or reward: Implications for the law of effect. *Journal of Experimental Psychology: Animal Bahvior Processes 27*(4), 354–372.

Gilbert, N. (2008). *Agent-based Models*. Number 153 in Quantitative Applications in the Social Sciences. Los Angeles: SAGE Publications.

Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton and Oxford: Princeton University Press.

Godfrey, P., R. Shipley, and J. Gryz (2007). Algorithms and aanalyse for maximal vector computation. *The VLDB Journal 16*, 5–28.

Goeree, J. K., C. A. Holt, and A. K. Moore (2005). An experimental examination of the volunteer's dilemma. `http://people.virginia.edu/~cah2k/vg_paper.pdf`.

Gomes, E. R. and R. Kowalczyk (2009). Dynamic analysis of multiagent Q-learning with ϵ-greedy exploration. In A. P. Danyluk, L. Bottou, and M. L. Littman (Eds.), *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montréal, Canada*, pp. 369–376. ACM.

Gotts, N. M., J. G. Polhill, and A. N. R. Law (2003). Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review 19*(1), 3–92.

Gray, L. N., W. I. Griffith, M. H. von Broembsen, and M. J. Sullivan (1982). Social matching over multiple reinforcement domains: An explanation of local exchange imbalance. *Social Forces 61*(1), 156–182.

Gray, L. N., M. C. Stafford, and I. Tallman (1991). Rewards and punishments in complex human choices. *Social Psychology Quarterly 54*(1), 318–329.

Gray, L. N. and I. Tallman (1984). A satisfaction balance model of decision making and choice behavior. *Social Psychology Quarterly 47*(2), 146–159.

Gray, L. N. G. and M. H. von Broembsen (1976). On the generalizability of the law of effect: Social psychological measurement of group structure and process. *Sociometry 39*(3), 175–183.

Green, L. and D. E. Freed (1993). The substitutability of reinforcers. *Journal of the Experimental Analysis of Behavior 60*(1), 141–158.

Gureckis, T. M. and B. C. Love (2009). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology 53*(3), 180–193.

Hamblin, R. L. (1977). Behavior and reinforcement: A generalization of the matching law. In R. L. Hamblin and J. H. Kunkel (Eds.), *Behavioral Theory in Sociology. Essays in Honor of George C. Homans*, pp. 469–502. New Brunswick, N.J.: transaction Books.

Hamblin, R. L. (1979). Behavioral choice and social reinforcement: Step function versus matching. *Social Forces 57*(4), 1141–1156.

Hamblin, R. L. and J. H. Kunkel (Eds.) (1977). *Behavioral Theory in Sociology. Essays in Honor of George C. Homans*. New Brunswick, N.J.: transaction Books.

Hardin, R. (1982). *Collective action*. Baltimore, London: John Hopkins University Press.

Harsanyi, J. C. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge, UK: Cambridge University Press.

Harsanyi, J. C. and R. Selten (1992). *A general theory of equilibrium selection in games* (2nd ed.). Cambridge, Mass.: The MIT Press.

Hart, D. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*(5), 1127–1150.

Hauert, C. et al. (2002). Volunteering as red queen mechanism for cooperation in public goods games. *Science 296*(5570), 1129–1132.

Hauert, C. et al. (2007). Via freedom to coercion: The emergence of costly punishment. *Science 316*(5833), 1905–1907.

Hauser, K. (2012, February). B553 Lecture 7: Constrained Optimization, Lagrange Multipliers, and KKT Conditions. `http://homes.soic.indiana.edu/classes/spring2012/csci/b553-hauserk/constrained_optimization.pdf`.

Heckathorn, D. D. (1988). Collective Sanctions and the Creation of Prisoner's Dilemma Norms. *American Journal of Sociology 94*(3), 535–562.

Hedström, P. and P. Bearman (2009). What is analytical sociology all about? An introductory essay. In P. Hedström and P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology*, Chapter 1, pp. 3–24. Oxford, England: Oxford University Press.

Heiner, R. A. (1983). The origin of predictable behavior. *The American Economic Review 73*(4), 560–595.

Heiner, R. A. (1985a). Experimental economics: Comment. *The American Economic Review 75*(1), 260–263.

Heiner, R. A. (1985b). Origin of predictable behavior: Further modeling and applications. *The American Economic Review 75*(2), 391–396.

Heiner, R. A. (1985c). Predictable behavior: Reply. *The American Economic Review 75*(3), 579–585.

Heiner, R. A. (1988). The necessity of imperfect decisions. *Journal of Economic Behavior and Organization 10*(1), 29 – 55.

Heiner, R. A. (1990). Rule-goverened behavior in evolution and human society. *Constitutional Political Economy 1*(1), 19–46.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior 4*(3), 267–272.

Herrnstein, R. J. (1969). Method and theory in the study of avoidance. *Psychological Review 76*(1), 49–69.

Herrnstein, R. J. (1970). On the law of effect. *Journal of the Experimental Analysis of Behavior 13*(2), 243–266.

Herrnstein, R. J. (1971). Quantitative hedonism. *Journal of Pychiatric Research 8*(3), 399–412.

Herrnstein, R. J. (1974). Formal properties of the matching law. *Journal of the Experimental Analysis of Behavior 21*(1), 159–164.

Herrnstein, R. J. (1979). Derivatives of matching. *Psychological Review 86*(5), 486–495.

Herrnstein, R. J. (1981). A first law for behavioural analysis. *Behavioral and Brain Sciences 4*(3), 392–395.

Herrnstein, R. J. (1982). Melioration as behavioral dynamism. In M. L. Commons, R. J. Herrnstein, and H. Rachlin (Eds.), *Quantitative Analysis of Behavior, Vol. II: Matching and Maximizing Accounts*, pp. 433 – 458. Cambridge, Mass.: Ballinger Publishing Company.

Herrnstein, R. J. (1990a). Behavior, reinforcement and utility. *Psychological Science 1*(4), 217–224.

Herrnstein, R. J. (1990b). Rational choice theory: Necessary but not sufficient. *American Psychologist 45*(3), 356–367.

Herrnstein, R. J. (1997). *The Matching Law. Papers in Psychology and Economics.* Cambridge, Mass. & London, England: Harvard University Press.

Herrnstein, R. J., G. F. Loewenstein, D. Prelec, and W. Vaughan (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making 6*(3), 149–185.

Herrnstein, R. J. and D. Prelec (1992). A theory of addiction. In G. Loewenstein and J. Elster (Eds.), *Choice over Time*, pp. 331 – 361. New York: Russell Sage Press.

Herrnstein, R. J. and D. Prelec (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives 5*(3), 137–156.

Herrnstein, R. J. and W. Vaughan (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits to Action: The Allocation of Individual Behaviour*, Chapter 5, pp. 143–175. New York: Academic Press.

Hester, T. and P. Stone (2012). Learning and using models. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 4, pp. 111–141. Berlin and Heidelberg: Springer.

Homans, G. C. (1961). *Social behavior. Its Elementary Forms.* London: Routledge & Kegan Paul.

Homans, G. C. (1974). *Social behavior. Its Elementary Forms* (2. rev. ed.). New York: Harcourt Brace Jovanich, Inc.

Houston, A. I. and J. McNamara (1981). How to maximize reward rate on two variable-interval paradigms. *Journal of the Experimental Analysis of Behavior 35*(3), 367–396.

Houston, A. I. and B. H. Sumida (1987). Learning rules, matching and frequency dependence. *Journal of Theoretical Biology 126*(3), 289–308.

Hursh, S. R. (1978). The economics of daily consumption controlling food- and water-reinforced repsonding. *Journal of the Experimental Analysis of Behavior 29*(3), 475–491.

Izquierdo, L. R., S. S. Izquierdo, N. M. Gotts, and J. G. Polhill (2007). Transient and asymptotic dynamics of reinforcement learning in games. *Games and Economic Behavior 61*(2), 259–276.

Jaakkola, T., M. I. Jordan, and S. P. Singh (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation 6*(6), 1185–1201.

Jaffray, J.-Y. (1974). On the extension of additive utilities to infinite sets. *Journal of Mathematical Psychology 11*(4), 431–452.

Judson, D. H. and C. Duran-Aydintug (1991). A test of the satisfaction-balance decision model using direct numeric estimation. *Social Forces 70*(2), 475–494.

Kandori, M., G. J. Mailath, and R. Rob (1993). Learning, mutation, and long run equilibria in games. *Econometrica 61*(1), 29–56.

Kangas, B. D., M. S. Berry, R. N. Cassidy, J. Dallery, M. Vaidya, and T. D. Hackenberg (2009). Concurrent performance in a three-alternative choice situation: Response allocation in a Rock/Paper/Scissors game. *Behavioural Processes 82*(2), 164–172.

Kianercy, A. and A. Galstyan (2012). Dynamics of Boltzmann Q-Learning in two-player two-action games. *Physical Review E 85*(4), 041145.

Krantz, D. H., R. D. Luce, P. Suppes, and A. Tversky (1971). *Foundations of Measurement. Volume 1: Additive and Polynomial Representations.* New York: Academic Press.

Kunkel, J. H. (1975). *Behavior, Social Problems, and Change. A Social Learning Approach.* Englewood Cliffs, N.J.: Prentice Hall, Inc.

Langton, C. G. (Ed.) (1995). *Artificial Life. An Overview.* The MIT Press.

Lemke, C. E. and J. T. Howson (1964). Equilibrium points of bimatrix games. *Journal of the Society for Industrial and Applied Mathematics 12*(2), 413–423.

Lindgren, K. and M. G. Nordahl (1993). Cooperation and community structure in artificial ecosystems. *Artificial Life 1*(1-2), 15–37.

Loève, M. (1978). *Probability Theory II* (4th ed.). New York, Heidelberg, and Berlin: Springer.

Loewenstein, Y. (2010). Synaptic theory of replicator-like melioration. *Frontiers in Computational Neuroscience 4*, 17.

Loewenstein, Y., D. Prelec, and H. S. Seung (2009). Operant matching as a Nash equilibrium of an intertemporal game. *Neural Computation 21*(10), 2755–2773.

Loewenstein, Y. and H. S. Seung (2006). Operant matching is a generic outcome of synaptic plasticity based on the covariance betwen reward and neural activity. *Proceedings of the National Academy of Sciences 103*(41), 15224–15229.

Lytinen, S. L. and S. F. Railsback (2012). The evolution of agent-based simulation platforms: A review of NetLogo 5.0 and ReLogo. In *Proceedings of the fourth international symposium on agent-based modeling and simulation.*

Macy, M. and M. Tsvetkova (2015). The signal importance of noise. *Sociological Methods & Research 44*(2), 306–328.

Macy, M. W. (1991). Learning to cooperate: Stochastic and tacit collusion in social exchange. *American Journal of Sociology 97*(3), 808–843.

Macy, M. W. and A. Flache (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences 99*(Suppl 3), 7229.

Macy, M. W. and A. Flache (2009). Social dynamics from the bottom up. Agent-based models of social interaction. In P. Hedström and P. Bearman (Eds.), *The Oxford Handbook of Analytical Sociology*, Chapter 11, pp. 245–268. Oxford, England: Oxford University Press.

Macy, M. W. and J. Skvoretz (1998). The evolution of trust and cooperation between strangers: A computational model. *American Sociological Review 63*(5), 638–660.

March, J. G. (1991). Exploration and exploitation in organization learning. *Organization Science 2*(1), 71–87.

Maynard Smith, J. (1976). Evolution and the theory of games. *American Scientist 64*(1), 41–45.

Mazur, J. E. (1981). Optimization theory fails to predict performance of pigeons in a two-response situation. *Science 214*(4522), 823–825.

Mazur, J. E. (2001). *Learning and Behavior* (5th ed.). Upper Saddle River, N. J.: Prentice Hall.

Mazur, J. E. and R. J. Herrnstein (1988). On the functions relating delay, reinforcer value, and behavior. *Behavioral and Brain Sciences 11*(4), 690–691.

McDowell, J. J. (1988). Matching theory in natural human environments. *The Behavior Analyst 11*(2), 95–109.

McDowell, J. J. (2004). A computational model of selection by consequences. *Journal of the Experimental Analysis of Behavior 81*(3), 297–317.

McDowell, J. J. (2005). On the classic and modern theories of matching. *Journal of the Experimental Analysis of Behavior 84*(1), 111–127.

McDowell, J. J. (2013a). On the theoretical and empirical status of the matching law and matching theory. *Psychological Bulletin 139*(5), 1000–1028.

McDowell, J. J. (2013b). A quantitative evolutionary theory of adaptive behavior dynamics. *Psychological Review 120*(4), 731–750.

McDowell, J. J. and M. L. Caron (2007). Undermatching is an emergent property of selection by consequences. *Behavioural Processes 75*(2), 97–106.

McDowell, J. J., M. L. Caron, S. Kulubekova, and J. P. Berg (2008). A compuational theory of selection by consequences applied to concurrent schedules. *Journal of the Experimental Analysis of Behavior 90*(3), 387–403.

McDowell, J. J. and A. Popa (2010). Toward a mechanism of adaptive behavior: Evolutionary dynamics and matching theory statics. *Journal of the Experimental Analysis of Behavior 94*(2), 241–260.

Michaels, J. W. and D. S. Green (1978). Behavioral sociology: Emergent forms and issues. *The American Sociologist 13*(1), 23–29.

Mitchell, M. (1995). Genetic algorithm and artificial life. In C. G. Langton (Ed.), *Artificial Life. An Overview*, pp. 267–289. The MIT Press.

Molm, L. D. (1981). The legitimacy of behavioral theory as a sociological perspective. *The American Sociologist 16*(3), 153–165.

Molm, L. D. (2006). The social exchange framework. In P. J. Burke (Ed.), *Contemporary Social Psychological Theories*, pp. 24–45. Stanford, California: Stanford University Press.

Molm, L. D. and J. A. Wiggins (1979). A behavioral analysis of the dynamics of social exchange in the dyad. *Social Forces 57*(4), 1157–1179.

Nakamura, Y. (2002). Additive utilities on densely ordered sets. *Journal of Mathematical Psychology 46*(5), 515–530.

Nash, J. F. (1951). Non-cooperative games. *The Annals of Mathematics 54*(2), 286–295.

Neth, H., C. R. Sims, and W. D. Gray (2005). Melioration despite more information: The role of feedback frequency in stable suboptimal performance. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, pp. 357–361.

Niv, Y., D. Joel, I. Meilijson, and E. Ruppin (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior 10*(1), 5–24.

Nowak, M. A. and R. M. May (1992). Evolutionary games and spatial chaos. *Nature 359*(6398), 826 – 829.

Nowé, A., P. Vrancx, and Y.-M. D. Hauwere (2012). Game theory and multi-agent reinforcement learning. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, pp. 441–470. Berlin and Heidelberg: Springer.

Ochs, J. (1995). Games with unique, mixed strategy equilibria: An experimental study. *Games and Economic Behavior 10*(1), 202–217.

Oliver, P., G. Marwell, and R. Teixeira (1985). A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action. *American Journal of Sociology 91*(3), 522–556.

Opp, K.-D. (1972). *Verhaltenstheoretische Soziologie. Eine neue soziologische Forschungsrichtung*. Reinbek bei Hamburg: Rowohlt.

Orbell, J. M. and R. M. Dawes (1993). Social welfare, cooperators' advantage, and the option of not playing the game. *American Sociological Review 58*(6), 787–800.

Palacios-Huerta, I. (2003). Professionals play minimax. *The Review of Economic Studies 70*(2), 395–415.

Papadimitriou, C. (2014). Algorithms, complexity, and the sciences. *Proceedings of the National Academy of Sciences 111*(45), 15881 – 15887.

Pavlov, I. P. (1927). *Conditional reflexes: An investigation of the physiological activity of the cerebral cortex.* Oxford, England: Oxford University Press.

Pierce, W. D. and W. F. Epling (1983). Choice, matching, and human behavior. A review of the literature. *The Behavior Analyst 6*(1), 57–76.

Poling, A., T. L. Edwards, and M. Weeden (2011). The matching law. *The Psychological Record 61*(2), 313–322.

Rachlin, H. (1971). On the tautology of the matching law. *Journal of the Experimental Analysis of Behavior 15*(2), 249–251.

Rachlin, H. (2000). *The Science of self-control.* Cambridge, Mass. & London, England: Harvard University Press.

Rachlin, H., R. C. Battalio, J. H. Kagel, and L. Green (1981). Maximization theory in behavioral psychology. *Behavioral and Brain Sciences 4*(3), 371–417.

Rachlin, H., L. Green, J. H. Kagel, and R. C. Battalio (1976). Economic demand theory and psychological studies of choice. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, Volume 10, pp. 129–154. New York: Academic Press.

Rachlin, H., J. H. Kagel, and R. C. Battalio (1980). Substitutability in time allocation. *Psychological Review 87*(4), 355–374.

Rachlin, H. and D. I. Laibson (1997). Introduction. In H. Rachlin and D. I. Laibson (Eds.), *The Matching Law. Papers in Psychology and Economics*, pp. 1–10. Cambridge, Mass. & London, England: Harvard University Press.

Railsback, S. F., S. L. Lytinen, and S. K. Jackson (2006). Agent-based simulation platforms: Review and development recommendations. *Simulation 82*(9), 609–623.

Rauber, T. and G. Rünger (2013). *Parallel Programming* (2nd ed.). Springer.

Rauhut, H. (2009). Higher punishment, less control? Experimental evidence on the inspection game. *Rationality and Society 21*(3), 359–392.

Reed, D. D. and B. A. Kaplan (2011). The matching law: A tutorial for practitioners. *Behavior Analysis in Practice 4*(2), 15–24.

Ringen, J. (1999). Radical behaviorism: B. F. Skinner's philosophy of science. In W. O'Donohue and R. Kitchener (Eds.), *Handbook of Behaviorism*, Chapter 6, pp. 159–178. San Diego: Academic Press.

Roth, A. E. and I. Erev (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behaviour 8*(1), 164–212.

Saito, T. (2011). How do we get Cobb-Douglas and Leontief functions from CES function: A lecture note on discrete and continuum differentiated object models. *Journal of Industrial Organization Education 6*(1), (Article 2). Available at SSRN: `http://ssrn.com/abstract=1567152`.

Sakai, Y. and T. Fukai (2008a). The actor-critic learning is behind the matching law: Matching versus optimal behaviors. *Neural Computation 20*(1), 227–251.

Sakai, Y. and T. Fukai (2008b). When does reward maximization lead to matching law? *PLoS ONE 3*(11), e3795.

Sakai, Y., H. Okamoto, and T. Fukai (2006). Computational algorithms and neuronal network models underlying decision processes. *Neural Networks 19*(8), 1091 – 1105.

Sandholm, T. W. and R. H. Crites (1995). On multiagent Q-learning in a semicompetitive domain. In G. Weiß and S. Sen (Eds.), *Adaptation and Learning in Multiagent Systems, IJCAI'95 Workshop, Montréal, Canada, August 1995, Proceedings*, Lecture Notes in Artificial Intelligence 1042, pp. 191–205. Springer.

Savastano, H. I. and E. Fantino (1994). Human choice in concurrent ratio-interval schedules of reinforcement. *Journal of the Experimental Analysis of Behavior 61*(3), 453–463.

Schelling, T. C. (1960). *The strategy of conflict.* Cambridge: Harvard University Press.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory 4*(1), 25–55.

Seth, A. K. (1999). Evolving behavioral choice: An investigation into Herrnstein's matching law. In D. Floreano, J.-D. Nicoud, and F. Mondada (Eds.), *Proceedings of the Fifth European Conference on Artificial Life*, pp. 225 – 236. Springer.

Seth, A. K. (2001). Modeling group foraging: Individual suboptimality, interference, and a kind of matching. *Adaptive Behavior 9*(2), 67–89.

Seth, A. K. (2002). Competitive foraging, decision making, and the ecological rationality of the matching law. In *Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior on From Animals to Animats*, Cambridge, MA, USA, pp. 359–369. MIT Press.

Seth, A. K. (2007). The ecology of action selection: Insights from artificial life. *Philosophical Transactions of the Royal Society B 362*, 1545–1558.

Shoham, Y. and K. Leyton-Brown (2009). *Multiagent Systems. Algorithmic, Game-Theoretic, and Logical Foundations.* New York: Cambridge University Press.

Shteingart, H. and Y. Loewenstein (2014). Reinforcement learning and human behavior. *Current Opinion in Neurobiology 25*, 93–98.

Shull, R. L. and S. S. Pliskoff (1967). Changeover delay and concurrent performances: Some effects on relative performance measures. *Journal of the Experimental Analysis of Behavior 10*(6), 517–527.

Sigmund, K., C. Hauert, and M. A. Nowak (2001). Reward and punishment. *Proceedings of the National Academy of Sciences 98*(19), 10757–10762.

Simen, P. and J. D. Cohen (2009). Explicit melioration by a neural diffusion model. *Brain research 1299*, 95 – 117.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics 69*(1), 99–118.

Sims, C. R., H. Neth, R. A. Jacobs, and W. D. Gray (2013). Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review 120*(1), 139–154.

Singh, S. P., T. Jaakkola, M. L. Littman, and C. Szepesvári (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning 38*(3), 287–308.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Konorski and Miller. *The Journal of General Psychology 16*(1), 272–279.

Skinner, B. F. (1953). *Science and Human Behavior.* New York: The Free Press.

Skyrms, B. (2010). *Signals. Evolution, Learning & Information.* New York: Oxford University Press.

Skyrms, B. and R. Pemantle (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences 97*(16), 9340–9346.

Slonim, R. and A. E. Roth (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica 66*(3), 569–596.

Soltani, A. and X.-J. Wang (2006). A biophysically based neural model of matching law behavior: Melioration by stochastic synapsis. *The Journal of Neuroscience 26*(14), 3731 – 3744.

Spaan, M. T. J. (2012). Partially observable Markov decision processes. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 12, pp. 387–414. Berlin and Heidelberg: Springer.

Staddon, J. E. R. (2001). *Adaptive Dynamics. The Theoretical Analysis of Behavior.* Cambridge, Mass. & London, England: The MIT Press.

Staddon, J. E. R. and D. T. Cerutti (2003). Operant conditioning. *Annual Review of Psychology 54*(1), 115–144.

Staddon, J. E. R. and S. Motheral (1978). On matching and maximizing in operant choice experiments. *Psychological Review 85*(5), 436–444.

Stafford, M. C., L. N. Gray, B. A. Menke, and D. A. Ward (1986). Modeling the deterrent effects of punishment. *Social Psychology Quarterly 49*(4), 338–347.

Sugrue, L. P., G. S. Corrado, and W. T. Newsome (2004). Matching behavior and the representation of value in the parietal cortex. *Science 304*(5678), 1782–1787.

Sunahara, D. F. and W. D. Pierce (1982). The matching law and bias in a social exchange involving choice between alternatives. *The Canadian Journal of Sociology 7*(2), 145–166.

Sutton, R. S. and A. G. Barto (1990). Time-derivative models of Pavlonian reinforcement. In M. Gabriel and J. Moore (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, Cambridge, Mass, pp. 497–537. MIT Press.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning. An Introduction.* Cambridge, Massachusetts, and London, England: The MIT Press.

Tallman, I. and L. N. Gray (1990). Choices, decisions, and problem-solving. *Annual Review of Sociology 16*, 405–433.

Thorndike, E. L. (1932). *The Fundamentals of Learning.* New York: Teachers College, Columbia University.

Thuijsman, F., B. Peleg, M. Amitai, and A. Shmida (1995). Automata, matching and foraging behavior of bees. *Journal of theoretical Biology 175*(3), 305–316.

Tsebelis, G. (1990). Penalty has no impact on crime: A game-theoretic analysis. *Rationality and Society 2*(3), 255–286.

Tsetsos, K., R. Moran, J. Moreland, N. Chater, M. Usher, and C. Summerfield (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences 113*(11), 3102–3107.

Tunney, R. J. and D. R. Shanks (2002). A re-examination of melioration and rational-choice. *Journal of Behavioral Decision Making 15*(4), 291–311.

Udehn, L. (2002). The Changing Face of Methodological Individualism. *Annual Review of Sociology 28*, 479–507.

van Hasselt, H. (2012). Reinforcement learning in continuous state and action spaces. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 7, pp. 207–251. Berlin and Heidelberg: Springer.

van Otterlo, M. and M. Wiering (2012). Reinforcement learning and markov decision processes. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 1, pp. 3–42. Berlin and Heidelberg: Springer.

Vaughan, W. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior 36*(2), 141–149.

Vaughan, W. (1985). Choice: A local analysis. *Journal of the Experimental Analysis of Behavior 43*(3), 383–405.

Vaughan, W. and R. J. Herrnstein (1987). Stability, melioration, and natural selection. In L. Green and J. H. Kagel (Eds.), *Advances in Behavioral Economics*, Volume 1, pp. 185–215. Norwood, N.J.: Ablex.

Veksler, V. D., C. W. Myers, and K. A. Gluck (2014). SAwSu: An integrated model of associative and reinforcement learning. *Cognitive Science 38*(3), 580–598.

Vlassis, N., M. Ghavamzadeh, S. Mannor, and P. Poupart (2012). Bayesian reinforcement learning. In M. Wiering and M. van Otterlo (Eds.), *Reinforcement Learning. State-of-the-Art*, Chapter 11, pp. 359–386. Berlin and Heidelberg: Springer.

Vollmer, T. R. and J. Bourret (2000). An application of the matching law to evaluate the allocation of two- and three-point shots by college basketball players. *Journal of Applied Behavior Analysis 33*(2), 137–150.

Watkins, C. J. C. H. (1989). Learning from delayed rewards. Ph. D. thesis, University of Cambridge, Englandvaughan.

Watkins, C. J. C. H. and P. Dayan (1992). Q-learning. *Machine Learning 8*(3-4), 279–292.

Watts, D. J. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton and Oxford: Princeton University Press.

Wiering, M. and M. van Otterlo (Eds.) (2012). *Reinforcement Learning. State-of-the-Art*. Berlin and Heidelberg: Springer.

Wilensky, U. (1999). Netlogo. `http://ccl.northwestern.edu/netlogo/`. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Wunder, M., M. Littman, and M. Babes (2010). Classes of multiagent Q-learning dynamics with ϵ-greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel*, pp. 1167–1174.

Wyatt, D. (2013). *Akka concurrency*. artima.

Yechiam, E., I. Erev, V. Yehene, and D. Gopher (2003). Melioration and the transition from touch-typing training to everyday use. *Human Factors: The Journal of the Human Factors and Ergonomics Society 45*(4), 671–684.

Young, H. P. (1993). The evolution of conventions. *Econometrica 61*(1), 57–84.

Young, H. P. (1998). *Individual Strategy and Social Structure*. Princeton and Oxford: Princeton University Press.

Young, H. P. (2004). *Strategic Learning and its Limits*. New York: Oxford University Press.