Contents

1 Game-theoretic models of learning and their properties **2**
1.1 Why does learning matter? . . . 2
1.2 Classification . . . 4
1.3 Basic models of rational learning in games . . . 6
1.3.1 Notation and definitions . . . 6
1.3.2 Optimal play against Nature: the model of a multi-armed bandit 7
1.3.3 Fictitious play, the first classical algorithm . . . 8
1.3.4 Definitions of convergence: which criterion is the right one? . . . 11
1.3.5 Bayesian learning . . . 14
1.4 Basic models of bounded-learning in games . . . 19
1.4.1 Calibration . . . 19
1.4.2 Directional learning . . . 20
1.4.3 Reinforcement learning . . . 21
1.4.4 Weighted Fictitious Play . . . 24
1.4.5 Experience Weighted Attraction Learning (EWA) . . . 25
1.5 Sophisticated learning . . . 29
1.5.1 Why we need another class of models? . . . 29
1.5.2 Examples and limitations . . . 30
1.6 Chapter summary . . . 34
1.6.1 Bayesian algorithm play against nature . . . 35

Bibliography **36**

Chapter 1

Game-theoretic models of learning and their properties

1.1 Why does learning matter?

The thesis is devoted to the development of one of the most significant approaches to the analysis of economic behavior in game theory: learning models. While it is not currently the most popular approach in game theory, it is as old as the game theory itself[1] steadily develops to this day and focuses on the important aspect of strategic decision-making that other approaches tend to discount.

Since learning is a very polysemantic word in sciences from game theory to machine learning to psychology, we will restrict our attention to microeconomic models of individual learning.

They typically describe a dynamic decision-making process in a repeated game (in a game-theoretic sense) where the opponents use a set of decision-making principles that allow them to adapt their strategies to their strategic environment. Naturally, this adaptation requires some feedback from the previous play history. By strategic environment here we mean that the opponents face not the "market" or "nature", but a set (usually small, we'll discuss here only two-player games) of opponents that adapt to players as well.

This approach addresses game theoretic issues that otherwise are usually overlooked. First, the behavior of the participants does not necessarily immediately fall into an equilibrium state in observations or experiments on any dynamic (in particular, repeated) game. Therefore special, "tatonnement" models of convergence of behavior are required. Second, from a purely theoretic point of view, there may be many equilibria in games themselves, which poses the natural problem of choosing the "best" of them (Van Damme, 1991).

In particular, the criterion for comparing equilibria may be the probability that players will converge exactly to that equilibria from an arbitrary initial state. To characterize such equilibria, we need to understand the dynamics of the participants' interaction and the rules by which they choose their decisions. All of this requires an explicit model of

[1]The earliest formal models date back to the 1950s (e.g. (Brown, 1951), (Bush and Mosteller, 1955)), and informal analysis can already be seen in Augustin Cournot in 1838 (see (Cournot, 1960)).

learning and game dynamics for predicting the outcomes of social interaction and finding the other possible states which could be achieved under certain conditions.

In modern non-cooperative game theory, equilibrium analysis consists of finding the fixed point of best response functions (Nash equilibria) or, in the case of their multiplicity, such fixed points which have special "desirable" properties[2]. While the existence, uniqueness, and properties of equilibrium in a particular economic model may present considerable technical difficulties for economic theory, it is often implicitly assumed that for agents in a real economic situation this equilibrium is not only obvious but it is a result of common knowledge (a consequence of the fact that both the game and the rationality of the players are common knowledge). The naive justification for this implicit assumption is that if there is only one equilibrium, "sooner or later" the players will come to it, and if there are many, then to the "sufficient" of these equilibria (why refinement is necessary). Explicit modeling of learning in games creates an opportunity to test this thesis and is an alternative to the axiomatic search for refined equilibria. That is why a significant part of the literature on learning in games is devoted to the analysis of convergence of various learning rules to equilibrium outcomes and comparing the effectiveness of different models (learning algorithms) in terms of these predictions.

Another reason to model the learning process itself is the limited computational capabilities of real players. These limitations cannot be neglected even for games with complete information (such as chess) and for players capable of trying a significant number of continuations of any game position (such as modern supercomputers). One obvious possible solution is to observe the behavior of human players, isolate the basic regularities of their play, map it into heuristics and suboptimal criteria, and finally transfer the latter into formal learning rules. This approach, in a sense, violates the assumptions about the "rationality" of the player, but instead, with minimal assumptions, it allows imitation of "observable behavior" with some approximation. In the case of laboratory experiments, an additional test is possible: is the model that the researcher builds consistent with how the participants themselves motivate their actions? Not least of all the interest in such models and related to empirical (first of all, experimental) studies. They convincingly show that, in a number of cases, real people do not behave in the way predicted by classical solutions such as Nash equilibrium. Thus, in experiments with the "beauty contest" games (Nagel, 1995) and "ultimatum" (Güth et al., 1982) people regularly and significantly deviate from a single equilibrium even when the game is repeated a sufficiently large number of times and the stimuli in the game constitute a meaningful part of their income. Individual learning models are characterized by the ability of players to update their behavior from round to round because each round must account for strategic uncertainty and new information to help resolve it.

The first chapter covers several issues in sequence. The section 1.2 and 1.3 focus on why the literature abounds in a large number of models instead of focusing around a limited standard set. In particular, the section 1.2 introduces several classifications in use, the elements of which make up the structure of chapter one. Next, the section 1.3 is devoted to the class of rational models, and section 1.4 to boundedly rational models with their properties and behavior. The necessity of the existence of sophisticated boundedly rational models is introduced in the subsection 1.3.5 on Bayesian learning, and section

[2]This is called "refinement" of equilibria (Van Damme, 1991)

1.5 demonstrates examples of such models and presents the class of models proposed by the author.

1.2 Classification

There is no single and well-established approach to classify different learning models [3]((Marimon, 1996), (Nachbar, 2020), (Fudenberg and Levine, 2009), (Fudenberg and Levine, 2016)), so the process of classification is itself an additional exposition tool to understand and present different results.

First important division - whether the algorithms are deterministic or stochastic. In both cases, we may have "best" [4] response, but there are differences. The deterministic player must choose a single currently best response. A stochastic player may deviate, being "currently the best action" implies only a higher probability of playing this action, not certainty. Naturally, we can easily add randomness to the deterministic algorithm and, conversely, make a previously stochastic algorithm always play the most currently preferred action without randomization. While one might think that therefore the distinction is not important, we will see that it is the most important quality for the equilibrium convergence results.

A distinction that is important is between "rational" algorithms – those that use all currently available information and in some sense optimally respond to it and those that due to internal or external restrictions cannot respond optimally and therefore are "bounded rational" (Marimon, 1996). Generally "rational" algorithms are deterministic, but there are some exceptions. With some caveats, we can give as examples of rational learning "multi-armed bandit", Bayesian learning, and fictitious play with no memory boundaries (we will explain these models in greater details below).

With regards to bounded rationality, this model class can be further split by what is importantly bounded in the model. Typical examples here are the computational resources of the player and the length of memory, as well as uncertainty in the goals or purposes of the opponent. Chess gives us a vivid example of computational limitations: real players, including machines, cannot calculate all the possible continuations and find the optimal move with certainty in every given position. Chess is still playable and learnable because for a learning model the raw number of moves can be more or less important depending on the heuristics. Heuristics allow us to calculate only a small number of moves after the most promising continuations, thus decreasing the number of calculations required. We illustrate the general principles of such models with an example of a standard reinforcement learning model.

Bounds of memory capacity are the simplest to implement in an algorithm – just assume that players "forget" everything that happened a long time ago, and their strategy depends only on the results of the few recent rounds. This assumption can be justified in a non-stationary environment because as the environment changes with time, recent events follow the actual environment better than older events. In particular, we can interpret classic psychological finding – Ebbinghaus forgetting curve ((Ebbinghaus, 1885/1974) via

[3] we will interchangeably call them "algorithms" as well

[4] the best response is the strategy which produces the most favorable outcome for a player, taking other players' strategies as given (Fudenberg and Tirole, 1991) p. 29

Table 1.1: Navigator by model classification

Learning rule	**Response principle**	**Beliefs/reinforcement**[5]	**Model parameters**	**Original formulation of model**
Cournot dynamics , subsection 1.3.3	Deterministic, response to previous move	Belief based	-	Cournot (1838)
fictitious play, subsection 1.3.3	Deterministic, response to empirical frequencies	Belief based	The memory attrition parameter	Brown (1951)
Bayessian learning, subsection 1.3.5	Deterministic, response to empirical frequencies	Belief based	depends on model specifications	Ramsey (1926)
A model of a multi-armed bandit, subsection 1.3.2	Deterministic, response to empirical frequencies	Belief based	Discount factor, the utility of experimentation	Robbins (1952)
Calibration, subsection 1.4.1	Stochastic, response to empirical frequencies	Reinforcement based	-	Foster, Vohra (1998)
Directional learning, subsection 1.4.2	Stochastic, response to reward	Both	Depends on the of the model	Selten, Stoecker (1986)
Reinforcement learning, subsection 1.4.3	Stochastic, response to reward	Reinforcement based	Memory attrition parameter, cut-off parameter, parameter of local experimentation	Roth, Erev (1995)
Experience-weighted attraction Learning (EWA), subsection 1.4.5	Stochastic, response to reward	Both	Discounting, strength of experience, hypothetical payoff weight, attraction sensitivity, "shape" of the previous attraction	Camerer, Ho (1999)

(Murre and Dros, 2015)) as an adaptive mechanism that discounts rare random experience and reinforces common and important experience. Among examples of such models are also bounded-memory fictitious play and Cournot best response models.

Goal uncertainty, for example, means that the player does not know the payoff matrix (whether the opponent values outcome A above or below the outcome B). It is similar to the uncertainty of the opponent's reaction to the player's actions when the player does not know, how the future opponent's actions depend the current player's actions (for example whether the opponent will cooperate in response to cooperation).

Further, we can introduce the following distinction: in one case player treats the opponent's actions as a "sample" that is essentially random and does not account for the actions of the player, in another player can affect what information about the opponent and own future actions will be revealed. The latter category includes such models as Experimental Weighted Attraction (EWA).

After a short description, we will present the models in the following order: we will begin with rational models, first against a passive opponent, the consecutively more and more complex models against an active opponent, culminating in Bayesian learning, its advantages, and disadvantages. In particular, a number of results on the impossibility of convergence to Nash equilibrium are presented and, as a consequence, appearing of other theoretical criteria (e.g. Hannan consistency) is discussed. Then we will discuss bounded rational rules: calibration, directional learning, reinforcement learning, Experience-weighted attraction, and I-SAW. Both sections begin from a statistical, not

[5]These concepts represent two different ways to process information about events happening in-game, we discuss them in detail in the next section

entirely game-theoretic basis to underline the particularities of game-theoretic learning. Finally in the third section, we going to cover sophisticated learning topics, such as strategic learning and strategic teaching. We discuss there why standard theoretical criteria are not quite suitable for analyzing such topics and what criteria can be used to develop a theory of learning and to create new models.

1.3 Basic models of rational learning in games

1.3.1 Notation and definitions

Two players [6] play a repeated game, where every iteration (stage, round, period) - the finite static normal form game $G = \langle \mathcal{I}, S, \{u\}, T \rangle$, where $\mathcal{I} = \{1, 2\}$ – set of player (with number $I = |\mathcal{I}|$), $S = S_1 \times S_2$ – players' strategy profiles are Cartesian products of the finite sets of their pure strategies, $S_i = \{s_{i1}, s_{i2}, \ldots, s_{iJ}\}$, $i = \{1, 2\}$. Further, $\{u\} \equiv \{u_1(S), u_2(S)\}$ are payoff functions for each player is defined on S, and $T < \infty$ is the number of periods (rounds) of the finite repeated game, with typical period t. Mixed strategy σ_i of each player i is a probability distribution on the set S_i; it prescribes to each player i a probability $\sigma_i(s_{ij})$ to play her pure strategy $s_{ij} \in S_i$. Set of all mixed strategies i forms $J - 1$ dimensional simplex with a typical element σ_i.The set of mixed strategies of each player includes the set of pure strategies. In a repeated context, it determines the vector of frequencies with which each pure strategy is chosen, $[p^t_{i1}(s_{i1}), \ldots, p^t_{iJ}(s_{iJ})]$, although $\sum_{j=1}^{J} p^t_{ij} = 1$, $p_{ij} \in \sigma_i$. The set of mixed strategies of each player naturally contains the set of pure strategies at $\sigma_i(s_{ik}) = 1, s_{ik} \in S_i$. For a one-period game, the expected payoff of player i is defined as

$$U_i(\sigma) = \mathbb{E}_\sigma(u_i(\sigma)) = \sum_{s_i \in S_i} u_i(s_1, \ldots s_{\mathcal{I}}) \prod_{i=1}^{\mathcal{I}} \sigma_1(s_2)\sigma_1(s_2)$$

where the strategies of the players are assumed to be independent. Payoffs for player 2 are determined similarly.

The sequence of strategies chosen by the players in a dynamic game is called the *history of the game* at time t, and is denoted by h^t, where $h^t = \{s^1, s^2, \ldots, s^t\}$ and $s^t = \{s^t_1, \ldots, s^t_I\}$. Mapping $\xi_i : h^t \to \Delta S_i$, determining which of the pure strategies player i should choose in response to the observed history of the game is called a behavioral strategy of player i. All subsequent definitions naturally generalize to behavioral strategy profiles ξ.

Strategy $\bar{\sigma}_i$ in a one-period game is called *best response* to the profile σ_{-i} if $U_i(\bar{\sigma}_i, \sigma_{-i}) \in \max_{\sigma'_i \in \Delta S_i} U_i(\sigma'_i, \sigma_{-i})$. The set of all best replies of player i to strategy profile σ is denoted by $B_i(\sigma)$, and the set of all the best replies in the profile σ is $B(\sigma) = \prod_{i \in \mathcal{I}} B_i(\sigma)$. If the best reply is a mixed strategy, then each pure strategy $\bar{\sigma}_i$, to which it attaches a positive probability, should give the same expected utility against σ_{-i}. Otherwise, reducing the weight to the less advantageous of the pure strategy would yield a higher expected utility, and $\bar{\sigma}_i$ would not belong to B.

[6] For simplicity, we are limited to models with two players and the same number of J strategies, although in most cases they are summarized on any finite number of players and an unequal number of strategies.

A Nash equilibrium is such a profile of mixed strategies σ^* that is a mutual best reply, i.e. $\sigma^* \in B(\sigma)$. In equilibrium, there are no incentives for any of the participants to unilaterally change their strategies. Nash's theorem states that in any finite non-cooperative game with compact strategy spaces and upper semi-continuous best replies such equilibrium (in pure or mixed strategies) necessarily exists.

1.3.2 Optimal play against Nature: the model of a multi-armed bandit

To illustrate the statistical learning task and its properties in games, Let's start with the degenerate case of game learning — learning in game against "Nature." In the theory of games, the key difference between " nature " and other players is that it does not respond to the actions of others and has stationary strategies. "One-armed bandit" is a slot machine with a handle and reels, that draws at random a sequence of values of which the player receives a prize. Because the expected winnings of such a game are negative, the machine "loots" the player, hence the name.

Now let's imagine that a player sees several such automates in front of her (that's why the bandit "multi-armed", sometimes the name is shortened to simply "bandit problems" (Bergemann and Valimaki, 2008)), that are known only that the expected value from each hand may not be equivalent to the others. How to build the sequence of hand choices? Each of the machines should provide independent from the others and time permanent expected payoff (i.e. stationary), which is unknown to the player for any of the slot machines. The information that is available to the player appears as a result of experimentation - when a player pulls a handle, she observes the result. However, this information is not given for "free", otherwise the player would be unlimited in experimentation. There is something called exploitation-exploration trade-off which is a critical trade-off between what frequency player need to choose for the hand with the highest expected payoff from those already tried, and the frequency of choosing yet of unused arms, i.e., getting new information. The decision to use a new arm is associated with the risk that new will be worse than the best of those about which the information has already been accumulated. For example, if there are ten arms and information has already been collected about nine arms, the expected value on the remaining arm is a priori equal to the average (we have no reason a priori to assume it is better or worse than others), while outcome from the best of of the nine investigated arms is distributed as a maximum of nine such independent averages. If there are enough arms, then by trying only some of them, we risk never try the best one. But brute-force search of all the arms is suboptimal since each new one (as expected) adds less expected value than from the already used best one.

If the arms are independent and the distribution of outcomes for each of them is stationary, then such model is sufficient to determine the "best solution" found back in the (Gittins, 1979). At each move, there is the so-called Gittins index [7]. Strategy "Select the arm with the highest current index" minimizes future regret (formally defined as the difference between the payoff on the best and used hand) from not using other hands.

Formally, this index (the notation follows (Bergemann and Valimaki, 2008)) can be

[7] We can draw an analogy with the Sprague–Grundy function in combinatorial games

defined as follows. Consider the decision problem on an infinite horizon with discrete-time $t = 0, 1, \ldots$. At each moment it is necessary to choose between K arms, denote this choice a_t. Action $a_t = k$ results in payoff with the same indices, x_t^k. This payoff is random and is defined by the realization of a random variable X_t^k. The sequence of choices can change the state of the system, denote this state by s_t. Then the distribution of X_t^k is representable in the form of $F(\cdot; s_t)$, where independence of the arms means that $F^k(\cdot; s_t) = F^k(\cdot; s_t^k)$. The transition function between states is $s_{t+1} = \varphi(x_t^k; s_t)$. We assume that the state variable can be decomposed into K components independent of the other hands, i.e. for all k, $s_{t+1}^k = s_t^k$ if $a_t \neq k$ and $s_{t+1}^k = \varphi(x_t^k; s_t)$ if $a_t = k$.

Then the Gittins index for the time moment τ is defined as

$$g(s_t^k) = \sup_{\tau \geq 2} \frac{\mathbb{E}_x[\sum_{t=1}^{\tau-1} \beta^t X^k(s_t^k)]}{\mathbb{E}_x[\sum_{t=1}^{\tau-1} \beta^t]}$$

where τ is the current round, β is the discount factor, and $X^k(s_t^k)$ is the payoff of state s_t^k. This is the expected utility normalized by the discount factor from the choice of a given arm, where utility is calculated by taking into account the change in utility from experimentation. As we note below, the issue of evaluating the utility of experimentation remains relevant to contemporary literature as well.

However, this index can be difficult to calculate, for which the new approximate methods and variations of the problem statement are being developed, such as richer description (context) of states S_t (survey is represented in (Bubeck and Cesa-Bianchi, 2012)).

The discussion about learning cannot end with this model because it is poorly applicable for game situations with an active opponent reacting to the player's actions. You can't always count on your opponent to play one or the other action (giving different arm gains) independently from the actions of the player herself. That is, the distribution of the values is not only nonstationary but also dependent on the previous actions of the player, which excludes the applicability of successful "multi-armed bandits" solutions in the game with an opponent who is not, in game-theoretic terms, "Nature".

1.3.3 Fictitious play, the first classical algorithm

Fictitious play is a learning rule first described by George Brown (Brown, 1951), who introduced this somewhat obscure but established term. In this family of models, each player assumes that the opponent plays a stationary (possibly mixed) strategy and best response to the empirical frequency of the opponent's strategies. Each participant looks at a history that has already been played and acts solely based on these observations, without taking into account the possibility of an opponent's reaction. Depending on the precise definition of "empirical frequency" and "best response" are possible very different algorithms with different properties, so we discuss only the most basic results.

We begin our consideration of the fictitious play with its simplest variant and, at the same time, with historically the first model of learning dynamics in which players respond to each other's actions: with the dynamics of the best response in the Cournot model. In this classic problem there are two firms, knowing each other's costs and market demand, simultaneously determine volumes the output on which their profits depend. In each period of the repeated game, two firms observe the output decisions made by

both players and set their output on the level corresponding to the best response to the opponent's decision in the preceding period.

Cournot dynamics (best response)

As the simplest example of this behavior, we can consider the degenerate case of a fictitious play, and at the same time historically the first example of learning dynamics in which players respond to actions each other. In the original work of (Cournot, 1960), Cournot was not primarily interested in finding the equilibrium known now as a Cournot-Nash equilibrium, but an adaptive process of mutual "adjustment" of the strategies of the two firms, competing in terms of output in a particular market (in Cournot's book — in the mineral water market). If the output of the two firms is denoted by s_1, s_2, and the utility function by $u_i(s_1, s_2), i \in \{1, 2\}$, then the best response function is — $B_i(s_{-i}) = \arg\max_{\tilde{s}_i} u_i(\tilde{s}_i, s_{-i})$. The standard case is, that the objective function $u(\cdot, s_{-i})$ is strictly concave on the opponent's strategy. Starting with any profile of strategies $s^0 = \{s_1^0, s_2^0\}$, Cournot dynamics assumes that both participants choose the best response to the opponent's strategy chosen on the previous move, i.e. $s_i^t = B_i(s_{-i}^{t-1})$. Due to the concavity of the objective functions, the best response curves decrease by the opponent's strategy and intersect in one point, which guarantees the uniqueness of the mutual best response (equilibrium). Such dynamics are, of course, very "myopic": players in it react only to the opponent's behavior in the previous period, without trying to anticipate her actions. An extension of the Cournot dynamics, taking into account not only the opponent's last move but also the whole history of the game, would be the stationary fictitious play model.

Stationary fictitious play model

Two players play a repeated game $G = \langle \mathcal{I}, S, \{u\}, T \rangle$, and each has J strategies $S_i = \{s_i^1, \ldots, s_i^J\}$. In addition, the initial weights are set or "counters" κ_{ik} for each strategy k of the player i. In the process of creating the history of the game h^t, where $t \in 1, 2, \ldots T$ statistics of what actions your opponent has chosen in previous periods are collected. Players simultaneously choose their strategies, observe each other's decisions, after which they update their beliefs (how the opponent will play) by adding 1 to the "counter" κ_{ik}^t of that strategy s_{-ik}^t, that the opponent chose during this period $t = 1, 2 \ldots$:

$$\kappa_{ik}^t(s_{-ik}^t) = \kappa_{ik}^{t-1} + b, \qquad b = \begin{cases} 1, \, if \, s_{-ik} \in s_{-i}^t \\ 0 \, if \, s_{-ik} \notin s_{-i}^t \end{cases} \tag{1.1}$$

In the Cournot dynamics, player i believed her opponent would choose the same strategy as in the last period. In this, more general case, the belief i that player $-i$ will play the strategy s_{-ik} at time t is defined as the relative weight of this strategy in the empirical frequencies of past the actions of the player $-i$:

$$\gamma_{ik}^t(s_{-i}^t) = \frac{\kappa_{ik}^t(s_{-i}^t)}{\sum_{j=1}^J \kappa_{ij}^t}$$

In the end, player i chooses her own reaction [8], i.e. best response on her current beliefs of how her opponent is playing:

$$BR_i^t(\gamma_i^t) \in \arg \max_{\{s_{ik} \in S_i\}} \mathsf{E}\left(u_i^t(s_{ij}^t, s_{-i}^t) \mid \gamma_i^t\right).$$

Talking about the properties it is worth starting from that the fictitious play converges to the Nash equilibrium in any zero-sum game (Robinson, 1951), in any non-degenerated game 2×2 (Miyasawa, 1961), in any game solved by the iterative method of excluding strictly dominant strategies (Nachbar, 1990). If fictitious play converges to the profile of pure strategies for all players — this profile will be a Nash equilibrium, also if for all players empirical frequency distributions γ_i^t converge, then the profile of the strategies, to which they converge, is the Nash equilibrium. When there is a strict Nash equilibrium in play, then this equilibrium is an absorbing state of fictitious play (Nachbar, 1990), (Fudenberg and Levine, 2009).

Several examples of play for FP

Let's look at a few examples of how a fictitious play works. As a first example, take the simplest game of "Matching pennies", presented in the Table 1.2. In this game, two players at the same time choose Head or Tail. If they chose the same thing, the winner is the first, and if different sides of the coin, the second.

Table 1.2: Matching pennies game

	Head (H)	Tail (T)
Head (H)	1; -1	-1;1
Tail (T)	-1;1	1;-1

Let's start with any profile of initial (zero period) counters, for example, $\kappa_1 = (1, 2)$ and $\kappa_2 = (3, 1)$, that is, the first player may consider the move "Tail" of the second one as more probable, and the second one considers the probable move "Head" of the first. Then in the first period, both participants play BR and weights become (1,3) and (3,2). Following this, player 1 should play H, and player 2 — T until experience convinces her that the former is more likely to be played by T, i.e. to the weights (3,4), when she will have to change strategy on H under the assumption that the opponent changes strategy only then, when the new strategy is strictly better than the old one. In response, player 1 will accumulate evidence in favor of her opponent playing H (starting with (1,5) and until (6,5), when she herself will have to change the strategy to H, and so on. In a long enough perspective, the empirical frequencies of such game converge to a single Nash equilibrium $\{[1/2,\ 1/2], [1/2,\ 1/2]\}$. In this example, the concept of fictitious play is unobjectionable.

However, there can also be complications. The convergence of empirical frequencies does not always capture the essence of the game well. Consider the game "Rock-Paper-Scissors" as an example. In this classic game participants simultaneously choose one of the items: rock beats scissors, paper beats rock, scissors beat paper (in other words, the relation between the strategies is non-transitive) — and the winner takes the payoff

[8] Note that it is allowed that there is not the single best response - in this case the solution is chosen arbitrarily.

Table 1.3: Rock-Paper-Scissors game

	Rock (R)	Scissors (P)	Paper (S)
Rock (R)	0,0	1,-1	-1,1
Scissors (S)	-1,1	0,0	1,-1
Paper (P)	1,-1	-1,1	0,0

of this game period. The only Nash equilibrium in mixed strategies is $\{1/3, 1/3, 1/3\}$, (Table 1.3).

It is not difficult to see that in this case, the participants will respond to each strategy by changing the best response as soon as the player's action frequencies incentivize her opponent to switch to a winning strategy (rock beats scissors, paper wins the rock...). In this case, the empirical frequencies will indeed converge to equilibrium, but players' average payoffs will differ by players with each cycle of strategy change and the difference will increase sharply with increasing period, which we don't expect in a "equilibrium" game. Note that in this game the empirical frequencies will converge to equilibrium not only if the one-period game equilibrium (mixed Nash equilibrium) is played. When players change strategy each time to the best response to their own strategy of the previous round, during this one of them all the time will win, the other will lose, but the empirical frequencies will correspond to the equilibrium.

A stronger criterion would be the requirement of convergence of joint empirical frequencies to the profile of Nash strategies, in this game requirement is met if each position in the winnings matrix occurs with the same frequency (1/9) in the game history h^t. However, even for such a criterion, it is possible to create a simple deterministic rule that does not correspond to the expected understanding of "convergence", e.g., play $(R, R) \to (R, S) \to (R, P) \to (S, R) \ldots, (P, P) \to (R, R)$. Based on the law of large numbers, on the round numbers $t = 1, 10, 20 \ldots$ position (R,R) should occur no more than 1/9 times, which obviously will not be fulfilled.

1.3.4 Definitions of convergence: which criterion is the right one?

In the example of fictitious play, we see the problem of the balance of convergence definition according to which there are several ways to formulate significantly different definitions. Weaker definition corresponds to almost any learning rule, and stronger one corresponds to none, and at first glance, there may not seem to be much difference between the two (Nachbar, 2020). The example of a coordination game (matrix on the table 1.4) can serve as a vivid illustration of the difficulty of finding such a balance. If players begin to repeat a stage game Nash equilibrium play in which they play (A; A) in odd periods and (B; B) in even periods, then the system converges (trivially) to a Nash equilibrium in repeated play. Nachbar (2020) notes that: "strictly speaking, we get a different Nash equilibrium depending on whether the starting date of the continuation game is odd or even.

The convergence of the fictitious play in the general case is not guaranteed. A classic example is given in (Shapley, 1964), to which corresponds, for example, the payment matrix presented in the table 1.5. This game is a variant of the Rock-Paper-Scissors

Table 1.4: Coordination Game ("Battle of the sexes")

	A	B
A	1,1	0,0
B	0,0	1,1

game (1.3), which, however, becomes a nonzero-sum game. This slight difference leads, however, to a significant shift in the rate of accumulation of weights corresponding to the empirical frequencies of the best response: If the initial weights are attributed to participants playing any of the strategy profiles, lying outside the main diagonal, then the dynamics of the fictitious play will attribute to them following the cycle$(T, M) \to (T, R) \to (M, R) \to (M, L) \to (D, L) \to (D, M) \to (T, M) \dots$ Each next part of the cycle will require more and more time, but these dynamics never converges.

Table 1.5: Shapley game

$1 \diagdown 2$	L	M	R
T	0,0	1,0	0,1
M	0,1	0,0	1,0
D	1,0	0,1	0,0

Let's imagine now that one of the players consistently is making moves on the cycle $R \to S \to P \to R \dots$ in RPS. His/her astute opponent, having figured out this simple rule of updating strategies, will be able to always win by shifting his/her own strategy in the right direction also acting on cycle $P \to R \to S \to P$. Such a strategy in terms of empirical frequencies is corresponding to the Nash equilibrium mixed strategy, but the average payoff is -1 for the first player and 1 for the second player. (i.e. formally there is a convergence to equilibrium, but in fact, the first player behaves myopically and constantly loses)

In this case, it is convenient to define convergence in terms of stability of average payoffs rather than in terms of frequencies. If the average winnings of a pair of players do not change by more than ε, we can consider that it has converged. However, the fact of convergence itself does not impose restrictions either on the minimum required value of the average payoff or on the type of opponent. This brings us to the notion of universal convergence or Hannan consistency (Hannan et al., 1957). According to this criterion, a player will almost certainly get at least as much utility as she could have gotten if she had known in advance the frequency of her opponent's strategies σ_i (but not their order, not the strategies themselves on each of their rounds). Technically, it can be defined as:

$$\limsup_{T \to \infty} \left(\max_{\sigma_i} (u_i(\sigma_i, \gamma_i^t) - \frac{1}{T} \sum_t u_i(y^t(h^{t-1})) \right) \leq \varepsilon$$

Where $y^t : H \to \mathbb{R}_+$ is a function of outcomes. Unlike other criteria, universal convergence allows us to judge the quality of a strategy not by theoretical profiles, but by observable values — payoffs. In addition, it implies the ability of the algorithm to play with any type of opponent.

However, the very fact that the rule converges to an equilibrium outcome may still have too much space for interpretation. For example, in the Battle of the Sexes game, in Fig. 1.3.4 universal convergence would provide a player with an average payoff of 2, but with the right alternation of strategies, players could achieve an average of 3. So (Mathevet and Romero, 2012) took advantage of the observations (McKelvey and Palfrey, 2001) that learning algorithms poorly predict experimental outcomes, and compared theoretical predictions, simulation outcomes, and experimental results in terms of mean payoff. An example of this analysis is shown in Fig.1.3.4.

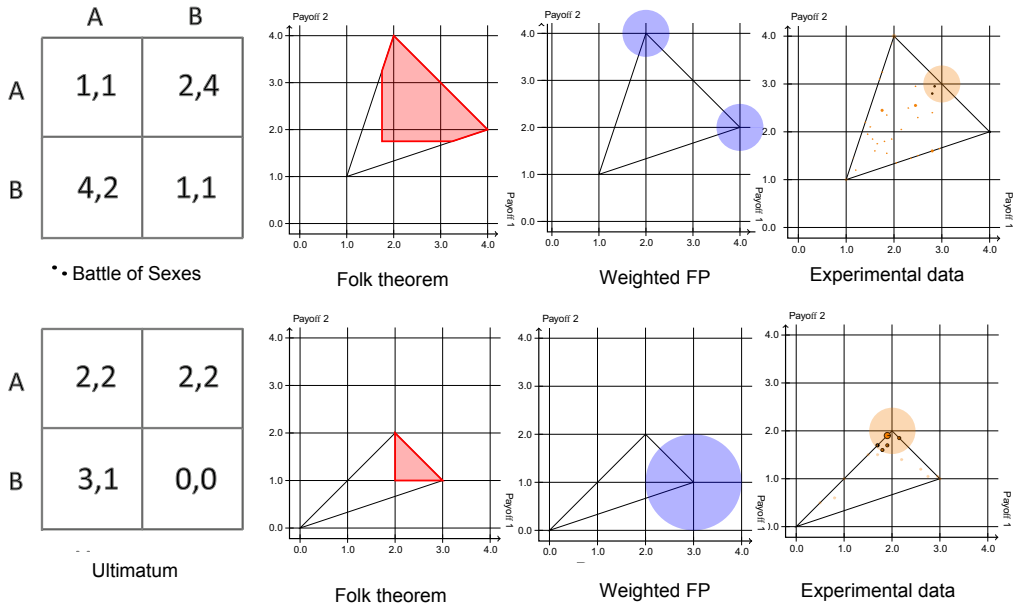


Figure 1.1: Comparative dynamics of fictitious play by (Mathevet and Romero, 2012)

In Fig.1.3.4, the left column shows the studied games in matrix form. The sets of all achievable potential payoff profiles for each of the games are described by closed contours, presented in the graphs to the right of the corresponding matrix. The second column contains illustrations including the set of payoffs that dominate the equilibrium payoffs in mixed strategies according to the folk theorem. The figure in the third column illustrates the convergence results of the simulated weighted fictitious play, and in the latter — payoffs distributions from experimental data.

In the last two cases, the circles indicate payoffs with coordinates in the center of each circle, and the diameters of each circle — the frequencies with which the population plays the matched strategy profile. For the convenience of comparison, the unit radius

corresponds to the whole population [9].

A comparison of the empirical frequencies of the average payoffs shows that the weighted FP converges to results different from the experimental data in both games, which can be clearly seen in the graph. So for example in the coordination game "Battle of the Sexes" people tend to coordinate quite quickly and switch between profiles (A,B), thereby achieving an identical average payoff (with a value equal to 3) for each participant. In contrast, a pair of "fictitious play" algorithms converge to a less "social" outcome (A,A), (B,B) in exactly half of the cases (which also demonstrates the sensitivity of dynamics to the choice in the first round or to the original beliefs). Thus, it can be argued that the relevant models do not describe very well the behavior of actual subjects especially in forward-looking prediction situations, where nevertheless with the right incentives the participants perform well. If each player believes that her opponent's behavior is described by a sequence of independent and equally distributed multinomial random variables, and its a priori beliefs with respect to this distribution are described by the Dirichlet distribution[10], then FP corresponds to a more general model - Bayesian learning (Fudenberg and Levine, 1993) (Nachbar, 2020), which is where we'll get to.

1.3.5 Bayesian learning

A class of Bayesian learning algorithms is often referred to as "rational learning" (Marimon, 1996), since they satisfy the standard axioms of rationality. Assumptions that specify rational preferences (e.g., Savage's subjective expected utility (Savage, 1954)), require from a player consistency of beliefs, including the inclusion of new information in the old system of beliefs according to Bayes' rule. However, compliance with some set of axioms alone cannot guarantee the adequacy of a learning rule.

Bayes' rule is widely and successfully applied in various fields of statistics, so it may seem that following it by a rational player is sufficient for successful convergence to equilibrium (stage game rational behavior). However, the structure of the representations of the opponent and some details of the algorithm formulation are important for learning, which leads to paradoxes.

To demonstrate the features of this class of models, let us characterize the essential

[9]The data in the third column shows the result of 1000 simulations with pairs of algorithms programmed to play a weighted fictitious play with ($\phi \in (0,1)$). Every simulation continued until the average pair payoff did not change by less than 0.01 in 20 consecutive blocks (In particular, the simulation is divided into blocks of 100 periods.). The maximum length in each of the runs was set to 100,000 periods, although the median the convergence length did not exceed 16,800 periods for each of the games. The results, further presented at graphs, are averages, taken for 1000 simulated pairs, for the last 1000 periods. For the experimental data, the sample size was 60 and 70 participants for each of the 2 games respectively. The data used for illustration shows the relative frequency of each possible payoff profile in the last 20 periods of the Supergame. Experimental data were collected in two subgroups with slightly different rules: in the first one, the initial 30 rounds were fixed, and then, starting from the 31st round, the probability of continuation was 0.9, so the expected length of each the supergame was 40. In the second one, starting from the first round, the probability of continuation was 0.99, so the expected length was 100.

[10]Applies to the discrete case of games with more than two strategies available, in this case the Multinomial distribution and the Dirichlet distribution form family of conjugate distributions: for the a priori Dirichlet distribution and multinomial of the likelihood function, the posterior distribution will also be the Dirichlet distribution.

determinants of Bayesian learning in the repeated game:

- Each player has an a priori probability distribution with respect to her opponent's behavioral strategies.
- Based on the game history available to all players after each period, these beliefs are updated by Bayes' rule.
- At each point in time, each player chooses the behavioral strategy that maximizes her expected discounted payoff in all future periods.

Even without imposing additional assumptions, there are difficulties with the dynamics of two players in a repeated game (primarily in formalizing sets of a priori distributions). However, a number of significant assumptions are usually added to the basic assumptions, with which we will start considering the dynamics of Bayesian learning.

A Bayesian learning model against a myopic opponent

The basic assumption is that the Bayesian learning algorithm (hereafter, Bayesian learner) interacts with an environment that is in no way dependent on its actions. [11] This approach can also be motivated as a player's interaction with the result of the averaged actions of many other players perceived as averaged social action. That is even if the agent's actions i affect the social outcome, as long as her actions do not take this dependence into account, she can treat the social outcome as an "exogenously given external world".

The learning rule under consideration prescribes rational players to update their beliefs according to Bayes' rule as history progresses, and choose the best reply $\sigma_i^{t+1}(h^t) \in B_i(x^t)$ in any period t, when they have decisions to make. Each t-th payment of player i depends only on her action $s_i^t \in S_i$ and on the state of the process x^t, for which in games it is natural to take the profile of opponents' strategies s_{-i}^t, so that $X = S_{-i}$ and $u \equiv u(s_i^t, x^t)$, while the set of pairs itself (s_i^t, x^t) – it's the history of the game. h^t.

As an example, consider a "game with nature" — a coin flip that can be biased, and where the goal of the participant playing with nature — is to bet on the correct (more likely) side of the coin. Each such coin toss is a realization of a random variable $X = \{0, 1\}$ with binomial distribution and unknown parameter θ (the true bias of the coin). The player's goal — is to determine the most likely value $\hat{\theta}^t$ after the history h^t, which will automatically allow her to choose $\sigma^t* = 0$ or 1, depending on whether $\hat{\theta}^t \gtrless 0.5$.

Let the a priori estimate of the probability of the outcome $x \in X$ is $\Pr(\theta)$, and there is a sequence of outcomes (tosses) $\mathbf{x}$ with a likelihood function $\Pr(\mathbf{x}|\theta)$. Then the a posteriori value of θ will be determined by Bayes' rule as

$$\Pr(\theta|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\theta)\Pr(\theta)}{\int \Pr(\mathbf{x}|\theta')\Pr(\theta')d\theta'}$$

In the coin flip case, the a priori probability is given by a two-parameter beta distribution, and the likelihood function is given by a binomial distribution, this pair forming

[11] If such an opponent can take into account only last actions but she is not forward-looking it is often referred to in the economic literature as "myopic".

a family of conjugate distributions: if the a priori distribution is a beta distribution and the likelihood function is binomial, then the posterior distribution will also be a beta distribution and the problem is solved analytically (see full description in the appendix). In our example, the dynamics of the game will look quite simple: the Bayesian learner should just flip a coin for a long time, updating the beliefs about the parameter of the beta-binomial distribution, then her prediction the head probability in the limit will converge to the true distribution $\theta^t \to \theta^*$, where θ^* — true value θ (Marimon, 1996). In the general case, $\Pr(\theta|\mathbf{x}) \propto \Pr(\mathbf{x}|\theta)\Pr(\theta)$, and the calculation of the posterior probability depends on the a priori distribution and its parameters.

Problems of Bayesian learning

Even with opponents playing unconditionally, learning procedures are optimal only if the agent's beliefs match the complexity of the environment. You need to have truthful assumptions in advance about how the "actual" environment is arranged. If this representation is simplified, then the agent's predictions may also be far from correct. For example, suppose the bias of a coin θ is not constant, but can vary from period to period (we alternately use two different coins). Bayesian learning rule does not guarantee us success: let, for example, the sequence of outcomes $X = \{0, 1\}$ — a simple alternation of outcomes $(0, 1, 0, 1, ..., 0, 1...)$. Knowing or guessing the rule, it is possible to guess each successive outcome. But the Bayesian learner would to converge to an estimate of $\theta = 1/2$ for this sequence, which would result in an error half the time. In this example, in forming the a priori probability distribution, the possibility of time-dependent bias (including "even-numbered — one, odd-numbered — another") should be taken into account. In general, however, it is difficult to draw up an exhaustive list of what possibilities should be considered.

Another problem is caused by the determinism and impossibility of "experiments". Consider Bayesian learning in the two-armed bandit problem. On one hand, there is a coin with probability $p$of falling out 1 (the other side brings 0), on the other hand, there is a coin with probability q and we need to maximize the discounted sum of occurring outcomes. It turns out that even Bayesian learners with any level of sophistication of a priori beliefs with positive probability converge to the choice of a non-optimal hand in the given problem (Rothschild, 1974). This happens if the player's current information indicates in favor of one hand and she permanently stops choosing the second hand, as she receives no new information about it. Since one hand is chosen only a finite number of times, the score q may not converge to its true value and the student continues to choose the hand with probability p even at $p < q$.

If even under sufficiently strict assumptions limiting the opponent's behavior there are situations in which the Bayesian learner behaves in a non-optimal manner, two questions remain. The first – if two Bayesian learners play with each other, will they converge to equilibrium (specifically, Nash equilibrium)? The second question is whether their beliefs about each other can converge to true beliefs in the limit. We will address these issues in the next section.

Bayesian learning in general form

The main difficulties in using Bayesian learning as a universal rule lie in the issue of how the system of a priori beliefs can be specified.

In general, the player moves from passive learning (where she interacted with an opponent playing unconditionally) to active learning (allowing that the opponent's actions depend in some way on the history of the game). For instance, a sufficiently patient rational player can maintain infinitely complex strategies (Young, 2004) (p.91), such as optimal experimentation, pattern recognition, deliberately misleading the opponent, myopia simulation, etc.

To illustrate these complexities, let us consider a general model of Bayesian learning in a repeated game, such that participants learn from all stories $h^t = (s^1, s^2 \dots s^t)$, including all actions chosen by participants up to and including period t. The set of all possible finite histories is denoted by H. Recall that the sequence of strategies chosen by the players in a dynamic two-player game is called the history of the game at the moment t, and is denoted by h^t, where $h^t = \{s^1, s^2, \dots, s^t\}$ and $s^t = \{s_1^t, \dots, s_I^t\}$. Mapping $\xi_i : h^t \to \Delta S_i$ that determines which of the pure strategies player i should choose in response to the observed game history is called the player's behavioral strategy i. It is assumed that each participant at each moment of time has some model of behavior of her opponents, defined as a mapping from the set of admissible histories into the set of strategies of her opponents, and denoted by $m_i : H_i \to \Delta(S_{-i})$. We denote the set of possible models by $\mathcal{M}_i$, and the set of beliefs that give a positive probability to the set of possible models $\mathcal{M}_i$ — as $\mu_i(\mathcal{M}_i)$. It is clear that under some such models the course of the game will converge to equilibrium (for example, if players are initially convinced that their opponents are initially playing the same Nash equilibrium). The question is how wide is the class of such models in which players rationally trained by Bayes' rule will always converge to equilibrium. This cannot generally be guaranteed, but the first significant result was Kalai and Lehrer (Kalai and Lehrer, 1993), who obtained the characterization of the conditions under which the players' beliefs converge to the true distributions of the opponents' behavioral strategies, and, as a consequence, the mixed strategies of the players converge to Nash equilibrium.

Each behavioral strategy profile $\xi(h^t) \equiv \xi_1(h^t) \times \dots \xi_{\mathcal{I}}(h^t)$, implemented after history h^t, specifies a well-defined probability distribution on a set of possible histories from the player's view i which is denoted by $D_i(\xi_i)$. Beliefs μ_i are called *absolutely continuous*, if for each probability distribution on the set of possible histories $D_i(\xi_i)$ model $m_i \in \mathcal{M}_i$ could be found such that the strategy prescribed by this model (Denote it as $D_i(\mu_i, \xi_i)$) specifies the same probability distribution, i.e. $D_i(\mu_i, \xi_i) = D_i(\xi_i)$, and each of the actually possible histories at the time t these beliefs assign the positive probability.

The main result – if all the strategies $\xi_i(\mu_i)$, generated by beliefs μ, absolutely continuous $D_i(\xi_i)$, then these strategies converge to a Nash equilibrium. By convergence here is assumed such that for each history, the probability distributions given by $\xi_i(\mu_i)$ almost surely match with the equilibrium mixed strategies in the given stage game. Proof of this technically difficult result follows from Blackwell's approachability theorem (see (Fudenberg and Levine, 1998)).

The requirement of absolute continuity is nontrivial. Suppose in an infinitely repeated prisoners' dilemma both players use a " grim-trigger" strategy (cooperation until the

opponent has deceived, and the rejection of cooperation after that) and believe that both will continue to play this strategy as soon as one of them will deceive. If the player's strategies are such that one of them soon is really deceiving, then the strategy of the grim-trigger is being implemented and beliefs turn out to be absolutely continuous. If, however, both players never cheat, then they will never have the opportunity to check this strategy and absolute continuity is broken. Another example for a wider class of games is provided by (Nachbar, 2005): If in "matching pennies" two players have constructed some correct models of behavior of their opponents, then once they recognize their strategies, it will no longer be advantageous for them to maintain equilibrium strategy given by these models, which again violates absolute continuity.

In the general case, it turns out to be impossible to construct a class of beliefs allowing to "learn" parameters of the distribution of the opponent's true beliefs, and, as a consequence, converging to an equilibrium regardless of her actions (Nachbar, 2020). (Based on the same idea for games with uncertainty in payoffs a similar result is obtained in (Foster and Young, 2001)). To clarify this result, let us first consider another example (from (Marimon, 1996))

Table 1.6: Play against "miscoordination environments" (Marimon, 1996)

	A	B
s_1	1	0
s_2	0	1

Let $X = \{A, B\}$, the matrix of payoffs is given by the table 1.6, and μ_i satisfies the condition of absolute continuity with respect to the family of distributions $\nu \in \mathscr{N}$. Let also the process underlying Bayesian learning is the result of the best response of the player B_i according to her behavioral strategy ξ_i. Now let's look at the process of data generation x^t, which prescribes on the best response of the player her opponent to play a disadvantageous strategy for the player, i.e. $Prob(x^t = A|x^{t-1}) > 1/2$, if $\sigma_t(x_{t-1}) = s_2$ and $Prob(x^t = B|x^{t-1}) > 1/2$, if $\sigma_t(x_{t-1}) = s_1$. If such a process lies in $\mathscr{N}$, then the Bayesian learner will be able to "learn it", however B_i^t would no longer be the best reply. With the new $\hat{B}_i^t$ let's bind a new "unprofitable" rule, still from $\mathscr{N}$ etc. As a rule, one cannot "close" this process, i.e. there is no optimal strategy for all modes of play lying in $\mathscr{N}$ and at the same time remain in this class $\mathscr{N}$, given the "feedback" on the optimal strategy. In other words, a player may assume that her opponent is not myopic and change the strategy from s_1 to s_2, but if the other player is also not myopic (both players are rational) and is sufficiently patient, then trying to learn the complex behavior of that opponent leads to complicating that opponent's behavior. What Young (2004)(p.92) says about it: "Indeed, when both actors are rational, the attempt to learn the complex behavior of the opponent frequently leads to still more complex behavior on the part of the learner." and further he notes that for some types of games: "this interactive effect yields behaviors that become arbitrarily complex, and are effectively impossible to learn through the updating of priors".

The general theorem (Nachbar, 2005) states that it is impossible for a rational learning strategy to follow three seemingly natural rules simultaneously:

- **learnability** i.e. of being able to arrive at a state in which beliefs about the opponent's play predict her next move as if the prediction were made under a known true distribution underlying the opponent's play

- **richness**, which means that the beliefs are closed with respect to the type of strategy, that is, if a certain type of strategy is included then together with it all variation of strategy must be included as well (e.g. all strategies with k-period memory)

- **consistency**, what means matching the player's best reply at each moment of the game to her beliefs.

Following the example given by (Nachbar, 2005) to illustrate these properties, let us turn to the already mentioned Bayesian interpretation of the fictitious play. If FP player has a best reply, then all variations of this best reply need to lie in FP strategy space, then belief set satisfies learnability and richness. However, If a fictitious player is faced with a situation where it has an equal Prior, such strategies become i.i.d. over actions, which violate the consistency requirement. This result is robust for any Bayesian learning and for ε-equilibrium strategies(Nachbar, 2005). It does not in itself imply the impossibility of learning and says only that the convergent learning algorithm cannot meet all three above-mentioned properties simultaneously. If we do not assume that the player knows anything about her opponent's utility function (this formulation of the learning problem is called "uncoupled dynamics"), (Hart and Mas-Colell, 2001) shows that convergence to Nash equilibrium against any arbitrary given learning algorithm is impossible for any learning rule at all, not only for deterministic algorithms based on Bayes rules.

1.4 Basic models of bounded-learning in games

1.4.1 Calibration

Let us start our discussion of boundedly rational rules with a statistically noteworthy rule: calibration of predictions. Game learning is closely related to the tasks of predictive statistics: both weather forecasting and predicting the opponent's behavior require analysis of previous experience and can include both point forecasts ("tomorrow it will rain", "the opponent will choose a strategy *tail*"), as well as probabilistic ("the probability of rain tomorrow is 85%", "the best reply is for a given game history will have an 85% probability of choosing *tail*""). Is it difficult to get a "correct" probabilistic forecast to predict an opponent's behavior?

Prediction calibration is defined as the correspondence between the predicted probabilities of events and the frequencies with which these events occur (Foster and Vohra, 1998). For instance, if a quarter of the days of the year were rainy, the better the calibration of the forecast, the closer was the predicted probability of rain to 0.25. Note, however, that the binary sequence prediction 01010101 can also be well-calibrated by both with a 0.5 probability prediction of each of the outcomes, and by an accurate deterministic sequence. On the other hand, a pure-calibrated forecast 10101010, although wrong every time, captures the essence of the process (alternation of zeros and ones) better than a forecast which would predict the probability of one being 0.33. These problems

are similar to the problems of determining the convergence of empirical game frequencies to equilibrium frequencies.

Nevertheless, it is not unhelpful to ask whether it is possible to construct a prediction rule that will be well-calibrated to any future paths of the progress of the predicted sequence? The answer for deterministic rules is simple - no, for a deterministic rule, no matter how it is complex, there will always be a sequence on which it is poorly calibrated. The answer for rules with randomness already depends on the source of the predicted sequence: It could be Nature or an adversarial opponent with the possibility of selecting the next element of the sequence, knowing the prediction for that element and the intermediate case when the opponent can change the sequence as the game progresses, but only knows the distribution of predictions, not the exact predictions for the next turn. As it might be expected, in a game against an omniscient opponent the ability to randomize doesn't help, and the multi-armed bandit task shows that in a game against Nature the answer is positive. For the adaptive opponent a nontrivial result is obtained (Foster and Vohra, 1998) that it is possible to construct such a prediction rule for an arbitrary sequence generated by a non-omniscient opponent that this rule will be well-calibrated. Alternative proof for the existence of calibrated randomized rules can be obtained using the minimax theorem (for details see. (Foster and Vohra, 1997)). This result also gave rise to a question, and how to check whether it was not some Forecast charlatan not possessing any information about the data generation process, but using this rule to create the visibility of good calibration (for details, see the section about calibration in the review (Nachbar, 2020)).

1.4.2 Directional learning

Directional Learning - one of the most common ways to assign bounded rational learning rules with a probabilistic action function (Selten and Stoecker, 1986), (Selten and Buchta, 1999) and (Selten et al., 2005). A classic example of directional learning - target shooting. When there are several attempts the shooter can estimate in which "direction" the result of the shot would be better and for the next shot to shift point of aiming in the same direction (therefore " directional ").

Formally, you can give a qualitative description of three conditions. (Selten, 2004):

1. Discrete time for learning is required $t = 1, 2, \ldots, T$.
2. Should exist a valid parameter υ_t which player chooses in each period.
3. A feedback should be configured to correct the value of the parameter regarding the previous choice.

For example, consider the prisoners' dilemma game. By backward induction (D,D) is the only equilibrium in a repeated game with a finite number of rounds (we will call the whole set of rounds a supergame). However empirically, participants in experiments with this game tend to cooperate in a significant number of rounds.

	C	D
C	5, 5	0,7
D	7, 0	2,2

In terms of the directional learning model (Selten and Stoecker, 1986), it's called silent cooperation. Participants are ready to cooperate up to the last round in supergame, while the opponent does not deviate. But each of the participants naturally predicts that at the end of the opponent will reject. So she is forced to decide which moment she should deceive to do not regret about missed benefit. When choosing a round in the supergame in which the player starts to deviate from cooperative behavior on her own, she will be guided by the experience gained from previous rounds. The intended period of deviation start will be the parameter of directional learning in this game.

Since directed learning is a qualitative rather than quantitative theory, for each individual game it corresponds to some stochastic functions, prescribing the player, instead of randomly choosing to choose more often something that shifting the player in the right direction. Examples of more specific specifications of directed learning could be found in – (Selten and Buchta, 1999), (Cason and Friedman, 1999), (Sadrieh, 1998).

In directional learning, the rule of action probability choice depends on the parameter, which determines the direction in which the player moves. What if this direction cannot be specified explicitly? Reinforcement learning is an important class of bounded-rational models that is appropriate for such a case.

1.4.3 Reinforcement learning

A natural source of non-economic inspiration for game theory is biology, which provides two developed paradigms of adaptation. One of them is borrowed together with the name in evolutionary game theory, but we will consider the second one, which considers adaptation not between generations of organisms, but for the same organism – reinforcement learning. It has its origins in the work of Ivan Pavlov on the formation of conditional reflexes, but in a century of active research has transcended the limits of actual biology and has become an integral part of computer science and psychology. It is based on the simple psychological principle of feedback, i.e. the choice and fixation in the observed behavior of such actions, to which the response from the external environment gives positive stimulation. Positive stimulation may not be sufficient on its own, but it can play a significant role in conjunction with beliefs. E.g. Selten et al. (2007) provide evidence from lab experiment with allocation passengers among two routes that both information and experience are relevant in the long run in the context of learning. At the same time, while modern psychology is skeptical of behaviorism, which takes this model as the sole basis of all psychological processes, on the contrary, it is an actively developing area of research in machine learning (for a more detailed introduction to this area we recommend (Sutton and Barto, 2018)).

Historically, reinforcement learning has its origins in the famous of the psychology of the "law of effect" (Thorndike, 1911), (Thorndike, 1927): "Responses followed by a satisfying effect are strengthened and likely to occur again in a particular situation but responses followed by a dissatisfying effect are weakened and less likely to occur again in a particular situation". The formulation of the reinforcement learning model itself is attributed to psychologists Bush and Mosteller (Bush and Mosteller, 1955). In economics, the application of reinforcement learning models has its origins in the work of (Erev and Roth, 1998).

Model

The reinforcement learning model on repeated games $G = \langle \mathcal{I}, S, \{u\}, T \rangle$ [12] is defined as follows. Each player has J strategies $S_i = \{s_i^1, \ldots, s_i^J\}$. Similar to the fictitious play model, for each player's are defined: the propensity q_{ij}^t for period t, initial propensity q_{ij}^0, and updating rule:

$$q_{ij}^{t+1} = q_{ij}^t + (u(s_j) - u_{min}(s))$$

where $(u(s_j) - u_{min}(s))$ is normalized utility from the strategy used $u(s)$ over minimal possible utility $u_{min}(s)$. If the strategy has not been chosen in this round, then the propensity to play remains unchanged. The pobability to choose a strategy in the following round is defined, like in the fictitious play, as a relative propensity:

$$p_{ik}^t = \frac{q_{ik}^t}{\sum_{j=1}^J q_{ij}^t}$$

Note that, unlike the fictitious play, the propensity to play one or another particular strategy is influenced not only by the success of the strategy in the past but also by the magnitude of the gain. Thus, the learning curve will initially have a steep incline but will become more flatter with time (with increasing t). The value of initial propensities is the only parameter in the classical formulation of the model, but in numerous extensions additional ones often appear. Here are some natural examples (first appearing in the Behaviorist literature of the thirties, e.g. (Watson and Kimble, 2017), or more modern example of a formal model – (Erev et al., 1995)):

- Cutoff parameter ϑ (Erev et al., 1995): probabilities $p_{ij} < \vartheta$ are assumed to be zero. If low probabilities are indistinguishable from zero for the player, significantly improves convergence (in the baseline version of the model, the number of rounds needed to achieve an equilibrium outcome may exceed 10,000).

- The Local Experimentation or Generalization parameter. Parameter, which affects the magnitude of the increase in propensity to play a particular Strategy: of the entire value x, only a fraction $1 - \epsilon$ is adding to the chosen strategy. The remainder ϵ is adding to the closest strategy. Thus ϵ is interpreted as local experiments or errors. Strategies must be interpreted in one dimension: price, quantity, amount given to the other player, etc.

- The recency or memory attrition parameter: players tend to attach more importance to recent events. The effect has long been discussed in the decision-making literature, e.g., in (Estes, 1964). Its application was so popular that there were even several models based on it (see sample-based model (Chmura et al., 2012) for belief-based one, and individual evolutionary learning (Arifovic and Ledyard, 2004) for reinforcement-based). The core idea of such models that agents draw a sample that is based on the past and react according to those policy function (e.g. best reply). But they are generally covered by the later idea of FP or RL with memory parameter. Formally it is given as a $1 - \phi$-adjusted propensity (where ϕ small). This parameter ensures that new observations contribute to the overall learning

[12] Note that it can also be specified for a wider class of games and situations

process, even if there is already extensive experience. In the classical model, inertia increases with increasing propensity: it takes more new observations to notice that the environment has changed.

Properties of the reinforcement learning model

In 2x2 constant sum games with two players and a single equilibrium, reinforcement learning converges to the value of the game (Beggs, 2005). Reinforcement learning and stochastic FP converge, but not necessarily to the Nash equilibrium, in the 2x2 games, zero-sum games, and coordination games (Hofbauer and Hopkins, 2005), with a higher convergence rate for the stochastic versions of FP than for reinforcement learning (Benaim and Hirsch, 1999).

To illustrate the complexity of interpreting the dynamics of reinforcement learning, we again use data from (Mathevet and Romero, 2012), which provides a comparison of theoretical prediction, two-player simulation results with "reinforcement learning" as a base, and the performance of participants playing the experiments in terms of average payoff. A detailed description of the experimental design and computer simulations can be found in subsection 1.3.4

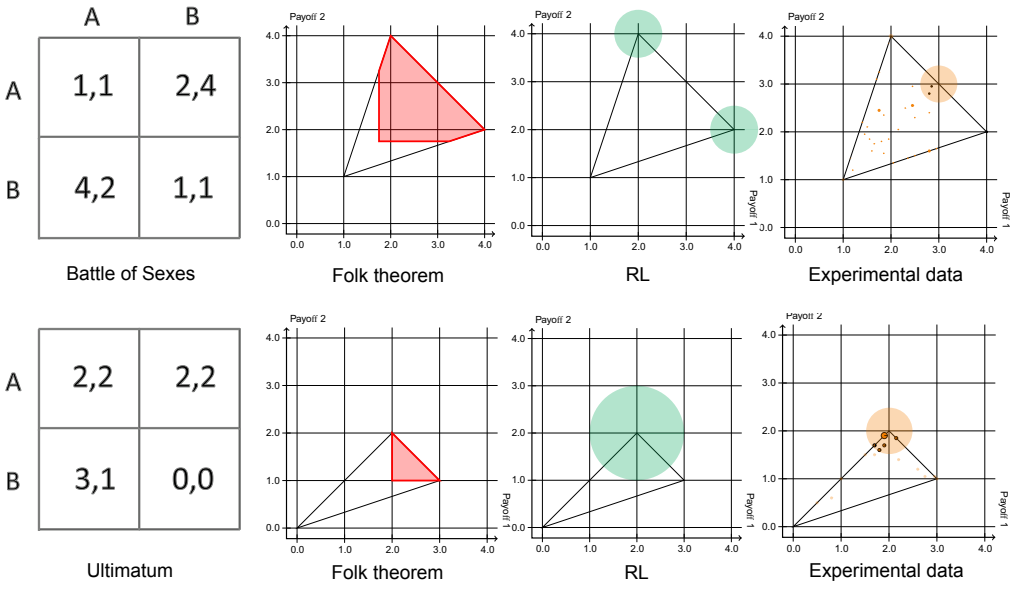


Figure 1.2: Comparative dynamics of reinforcement learning by (Mathevet and Romero, 2012)

In the figure 1.2, as in the previous example, the left column represents the investi-

gated games in matrix form. The sets of all achievable average payoffs for each of the games are described by closed contours shown in the graphs to the right of the corresponding matrix. The second column on the left represents the set of payoffs that dominate the equilibrium average payoffs in mixed strategies according to the folk theorem. The graph in the third column shows the convergence of the simulation of a pair of reinforcement learning algorithms against each other, and in the last column, the – payment distributions for the experimental data. In the latter two cases, the circles denote the payments with coordinates in the center of each circle, and the diameters of each circle are – the frequencies with which the population plays the corresponding profile of strategies. For ease of comparison, the unit radius corresponds to the entire population.

It can be observed that in the first game, the experimental results do not combine with the prediction of a simple reinforcement learning rule. In the second game ("Ultimatum"), reinforcement learning predicts half of the experimental results. In this case, the rationality constraints in the algorithm capture the tendency of participants to play a fairer outcome (2,2) and reinforce it. This can be interpreted as a step towards a more accurate description of the behavior of individuals, albeit one that requires more fine-tuning.

1.4.4 Weighted Fictitious Play

Before proceeding to describe hybrid models, it is useful to consider the way in which a deterministic algorithm turns into a stochastic algorithm. In the extension of classical FP – the Weighted FP model (Cheung and Friedman, 1997), each player i has three counters κ_{is}^t, one for each opponent's action $s \in \{0, 1, \ldots, J\}$. Starting with $\kappa_{is}^0 = 0$, at each time period $t \in \{1, 2, \ldots T\}$, these three counters are updated (enlarged or not). We enlarge the k-th counter κ_{ik}^t by 1 when the opponent's observed action s_{-i}^{t-1} is equal to this k:

$$\kappa_{ik}^t = \kappa_{ik}^{t-1} + \begin{cases} 1, \text{ if } s_{-i}^{t-1} = k \\ 0, \text{ if } s_{-i}^{t-1} \neq k \end{cases} \quad \forall k \in \{0, 1, \ldots, J\}.$$

Belief γ_{ij}^t of player i that his/her opponent $(-i)$ will choose action k at period t is defined as the relative weight, i.e, the empirical frequency of this action:

$$\gamma_{ik}^t := \frac{\kappa_{ik}^t}{\sum_{j=0}^{J} \kappa_{ij}^t} \quad \forall k \in \{0, 1, \ldots, J\}.$$

Unlike deterministic FP that reacts only to the most probable action of the opponent, the Weighted FP algorithm reacts to random actions, appearing with probabilities γ_{ik}^t. Further, a weighted fictitious play initiates a test to determine which of the γ is implemented in other words run hypothetical scenario the result of which is already a specific action of the opponent $\bar{s}_{-i}^t(\gamma)$. Finally, the algorithm plays

$$BR_i^t(\gamma) \in \arg\max_{\{s_{ik}\}} \mathsf{E}\left(u_i^t(s_{ij}^t, s_{-i}^t) \mid \bar{s}_{-i}^t\right).$$

1.4.5 Experience Weighted Attraction Learning (EWA)

Both approaches: both fictitious play and reinforcement learning appear to be reflecting some properties of real-world learning, but not describing the whole thing. The fictitious play pays close attention to the opponent's behavior, but not how (presumably) the best response to that behavior is really good. In contrast, reinforcement learning does not monitor an opponent's behavior and focuses on the success of one's own actions. Therefore, it seems natural to search for an approach that combines them in a single algorithm. One such algorithm, called "experience weighted attraction" (usually abbreviated EWA) (Camerer and Ho, 1999), allows for both fictional play and reinforcement learning and linear combinations thereof as special cases. This comes at the cost of a large number of parameters, causing criticism about the overfitting of such models.

The response to such criticism was the model STEWA (Ho et al., 2007), which is fixing part of the parameters at a "reasonable" level. Experience weighted attraction learning is one of the first models to incorporate elements of reinforcement learning and fictional play with psychological interpretation. Having simple models inside EWA is also convenient for practical reasons because they can automatically be tested inside the model and if they are more accurate, EWA should show it. Moreover, similarities and differences of simple models also become possible to observe in the data.

Formal model description

The underlying model assumptions resemble reinforcement learning. Consider the standard environment of a repeated game $G = \langle \mathcal{I}, S, \{u\}, T \rangle$ with $t \in 1, 2, \ldots T$ periods and J strategies. In addition, introduce a parameter normalizing the experience of previous periods $N(t)$ (at an initial value of $N(0)$) and attractions $A^t_{ij}(s_{ij})$ instead of propensities. p_{ij}. Previous experience is discounted by a factor of ρ and the parameter $N(t)$ follows the rule:

$$N(t) = \rho N(t-1) + 1, t > 1$$

$A^t_{ij}(s_{ij})$ is "attraction" of the j-th strategy of player i at period t (with initial attractions $A^0_{ij}(s_{ij})$). Counting and updating A_{ij} includes three components: discounting the old attraction $\phi N(t-1) A^{t-1}_{ij}(s_{ij})$ (ϕ – discount parameter), taking into account the result of the current round $u_{ij}(s^t_{ij}, s^t_{-ij})$ and the normalization of experience $N(t)$.

$$A^t_{ij}(s_{ij}) = \frac{\phi N(t-1) \times A^{t-1}_{ij} + [\delta + (1-\delta) \times \mathbb{I}(s_{ij}, s^t_i)] \times u_{ij}(s^t_{ij}, s^t_{-ij})}{\rho N_i(t-1) + 1}.$$

where $\mathbb{I}$ is an indicator of the strategy used, δ determines the comparative weight of payoffs from selected and unselected strategies in the attraction function. For example, if $\delta = 0$ then they will be counted as in the reinforcement learning model (consider the chosen strategy), and if $\delta = 1$ then as in the fictitious play model (consider all strategies). $A^0_{ij}(s_{ij})$ and $N(0)$ allow us to adjust the speed of learning in the first rounds of the game or the asymmetry in the attractiveness of the initial strategies, reflecting the initial knowledge of the player.

Similar to reinforcement learning, the probability of choosing strategy j in round t is given by an attraction-dependent objective function.

The original article (Camerer and Ho, 1999) suggests several ways to introduce probability, but the main one is logistic:

$$p_{ij}^{t+1} = \frac{e^{\lambda A_{ij}^t}}{\sum_{w=1}^{k} e^{\lambda A_{iw}^t}}$$

In total, the EWA model has 6 parameters (see table below), so that it can easily "fit" many possible trajectories, as confirmed by simulations (Salmon, 2001). Despite a large number of parameters, EWA performs better than its simpler counterparts, even with penalties imposed on their number.

ρ - discounting	ϕ - discounting
$N(0)$ - strength of experience	$A_{ij}^0(s_{ij})$ - "form" of previous attraction
δ - weight of hypothetical payoff	λ - attraction sensitivity

Recall again, but in tabular form, the relations between EWA and its nested models (note that at $\delta = 1$ the parameter $N(0)$ does not matter and can be any).

ϕ	δ	ρ	$N(0)$	Model
1	1	1	-	Fictitious play
0	1	0	-	Best response by Cournot
$\phi \in (0,1)$	1	ϕ	-	Weighted fictitious play
$\phi \in [0,1]$	0	0	1	Cumulative reinforcement
$\phi \in [0,1]$	0	ϕ	$\frac{1}{1-\phi}$	Average reinforcement

It is worth stating that the mere presence of different parameters makes interpretation difficult. Confusion arises also because the parameters are in non-trivial relations with each other, and their numerical values can be misinterpreted. For instance, we can think that a model changing parameters imitates the player's game, whereas the player does not do any complicated calculations, and if we ask him to describe what he does, we will find parameters like sensitivity to reinforcement (lambda) in a parametric model like EWA that have no analogs in the human description. However, apart from the attractions updating, the model does not change anything within the game, parameter values are fixed. The whole space of parameters implies that the player-human is simply a realization of one of the values of this set. Thus the average values of the parameters in the population are the subject of interest of the researcher having such a model in her toolkit.

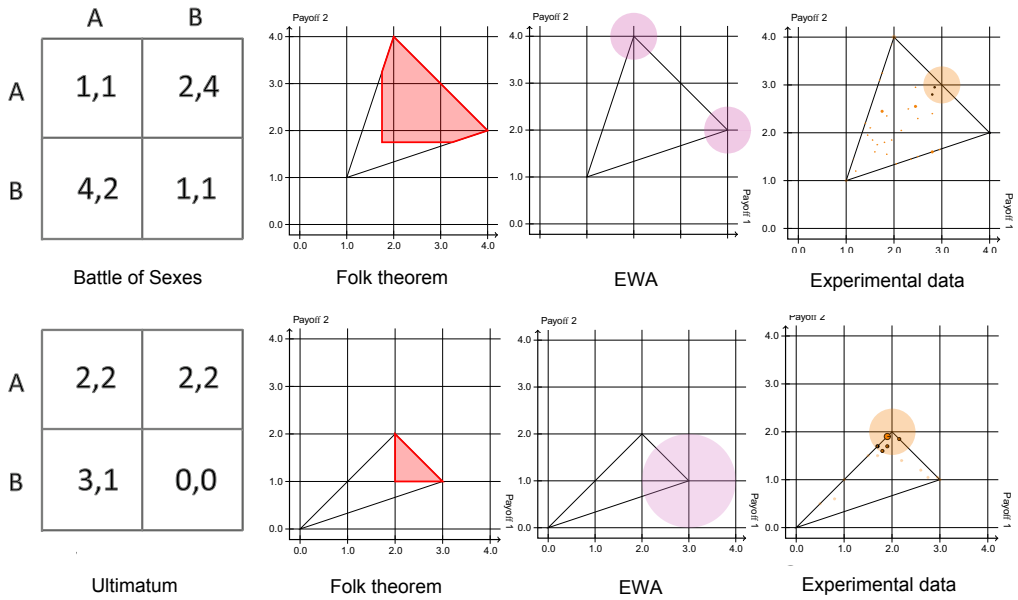


Figure 1.3: Comparative dynamics of EWA by (Mathevet and Romero, 2012)

To illustrate the complexity of interpreting experience weighted attraction learning dynamics again we use the comparison of the average payoffs from the simulation results of the two EWA players and the results of the participants of the experiments from (Mathevet and Romero, 2012). A refined description of the experimental design and computer simulations can be found in the subsection 1.3.4.

In the figure 1.3, as in the previous example, the left column represents the investigated games in matrix form. Sets of all possible payoffs for each of the games are described by the closed contours shown in the graphs to the right of the corresponding matrix. The second column contains illustrations that include the set of achievable average payoffs that dominate the equilibrium average payoffs in mixed strategies according to the folk theorem. The graph in the third column shows the convergence of pairs of experience weighted attraction (EWA) algorithms, and in the last column, the – average payoff distributions for the experimental data. In the last two cases, the circles indicate the payoff of the strategy profile corresponding to the center of each circle. The diameters of each circle reflect the frequency with which the population plays the appropriate strategy profile. For the convenience, the unit radius corresponds to the entire population.

The simulation results, as in the case of previous algorithms, show the mismatch with experiments. From a theoretical point of view experience weighted attraction learning could produce a prediction close to reinforcement learning, with parameters $\delta = 0$, however, at medium parameters it shows dynamics, more like an FP. This result can also

be interpreted as excessive complexity since the algorithm is more tend to use too many parameters for a fairly narrow space of strategies.

Self-tuning EWA (STEWA)

Modified model self-tuning EWA (Ho et al., 2007) contains only one parameter and the rest are "self-tuning". Initial parameters ($N(0)$, A_{ij}^0) are equated to 1 because they reflect a characteristic of the strategic situation rather than the learning process. Of the remaining two discounting parameters (ϕ, ρ) have been consolidated into $\phi_i(t)$. The following remain unchanged δ and λ. The accounting of previous experience in the modified model is given by the change detector – a function $\phi_i(t)$ based on the "Surprise index". It reflects the difference between the players' actions throughout the game and the most recent ones.

Let's define a vector of average game history of opponents players, containing the relative frequency of playing their J strategies, its element is:

$$\eta_{ij}(t) = \frac{\sum_{t=1}^{T} \mathbb{I}(s_{ij}, s_{-i}(t))}{t}$$

.

The latest factual history of the game forms another vector $r_{ij}(t) = \mathbb{I}(s_{ij}, s_{-i}(t))$. "Surprise index" $\Omega_i(t)$ is the sum of squares of deviations between these two vectors:

$$\Omega_i(t) = \sum_{j=1}^{J} (\eta_{ij}(t) - r_{ij}(t))^2$$

It takes values between zero (stable environment) and two (maximum unexpected outcome). In other words, the surprise index reflects the degree of change in the most recent observation relative to a stable story.

The change detector function is now set as:

$$\phi_i(t) = 1 - \frac{1}{2} \cdot \Omega_i(t)$$

Here is an example from the original article (Ho et al., 2007). Suppose your opponent consistently plays the same strategy for several rounds, and then suddenly changes it. In this case. $\phi_i(t)$ equals to $\frac{2t-1}{t^2}$. This reflects the fact that the more repeated choices – the greater the surprise will be the appearance of a different strategy (for $t = 2, 3, 5$ and 10 the value would be .75, .56, .36, and .19).

The attention function, as another self-adjusting parameter, has a similar, psychologically-based nature. Her idea is, that the player's attention shifts to a strategy that has not been chosen, but the hypothetical payoff on it exceeded the current payoff. Attention function $\delta(t)$ is equal to one if $u_i(s_{ij}, s_{-i}(t)) \geq u_i(t)$ and zero in the opposite case.

The STEWA model can be considered as the culmination of classical game learning model theory, but it does not answer many natural questions and leaves a wide field for further research. One such issue is the formal definition and study of experimentation in games, a behavior that is not locally optimal, but is optimal in the long run because it allows one to learn more about the opponent's behavior and response to the player's

actions. As we have seen throughout the review, positive results in learning theory point in one way or another to the need for such behavior to achieve equilibrium. However, it is difficult to investigate because in order to notice experimentation as a deviation from optimal behavior, one must determine which behavior is optimal within the model. For different models, the optimal (and thus experimental) behavior may be different even for the same dataset.

1.5 Sophisticated learning

In this section, we will discuss potential learning models that would better account for cognitive aspects of behavior. In subsection 1.5.1 we will discuss the role of pattern-recognizing rules among Bayesian rules and among universally consistent rules. In subsection 1.5.2 we will discuss the scope of sophistical learning models – where they can be helpful, what limitations they have, and some examples of their application. In subsection **??** we will discuss the extension of this type of model from 2-by-2 normal form games to games with continuous actions and will give some examples.

1.5.1 Why we need another class of models?

One of the main questions before the theory of learning in games was to determine whether two agents that do not know anything about their opponent's beliefs but can adapt over time converge to some kind of equilibrium and what kind if so. The rationality of the players in the equilibrium theory is an assumption that allows incorporating the reaction of the opponent to possible outcomes (we know that the opponent will act according to hisrational preference and thus can calculate its reaction).

From the (Foster and Young, 2003) (and also as we could ascertain from the section on Bayesian learning) we know about the limitation of the rationality concept. Young (2004) summarised it as "even when players are perfectly rational and arbitrarily forward-looking there may be no priors that permit them to learn Nash equilibrium behavior".

The same can be said about socially preferable outcomes, for example, Goyal and Janssen (1996) found that even in a 2-by-2 pure coordination game rationality alone cannot result in coordination. Thus, rational learning has interesting alternatives such as statistical learning that allows some form of deviation from the deterministic best response (usually by experimenting that was discussed in subsections devoted to the bandit model and convergence)

They are universal but simultaneously are very myopic, in situations where human subjects could rapidly recognize the Pareto-efficient outcome (like 010101 pattern in the Battle-of-Sexes game) they will respond only with an effective empirical frequency (randomizing equally 0 and 1), not with an efficient pattern.

The literature (Fudenberg and Levine, 1998)[p.337] describes it as "myopic learning procedures, not in the sense that players do not care about the future but in the strategic sense of lacking concern about the consequences of current play for opponent's future action". As we could see previously (see fig. 1.3.4, fig. 1.2) human subjects are able to coordinate.

"Another conclusion one might draw from our analysis is that the perfect rationality paradigm is not very powerful when it comes to the study of interactive learning

processes. Instead, one could resort to models of boundedly rational behavior or evolutionary models." Here Goyal and Janssen (1996) express a similar sentiment to us and the (Fudenberg and Levine, 1998)[p.4]: "our own views about learning models tend to favor those in which the agents, while not necessarily fully rational, are nevertheless somewhat sophisticated; we will frequently criticize learning models for assuming that agents are more naïve than we feel is plausible".

We can also notice that despite behavioral genesis some rules such as simple RL on actions also do not produce a socially preferable outcome. Fudenberg and Levine (1998)[p.302] claims the need to have "sophisticated learning in the sense that they explicitly attempt to detect patterns" and lists sources of such rules as (i) direct inclusion of behavioral strategies, (ii) conditioning on certain events, and (iii) using internal "experts" as they are usually defined in CS literature.

Let's note that this list is not mutually exclusive and we can have an algorithm that is mixing recognition CDC only on even rounds and CCD only on odd rounds (that is a mixe of (i) and (ii)). However crude exploiting of patterns is very limited in its expressiveness. E.g. if a model can only decide how to alternate between two options according to odd or even periods, it can be successfully implemented in the battle of the sexes (BotS) game but not in prisoners dilemma (PD). [13]

To be useful at least partially in addition to the pattern processing ability model should be able to avoid losing cycles, one simple way is described in Fudenberg Levine p.138 "Random behavior can prevent the player from being "manipulated" by a clever opponent".

Thus a behavioral model can be useful if it manages the trade-off between pattern recognition and manipulation avoidance (including the ability to adapt). Let's take a closer look at what kind of patterns we can consider.

1.5.2 Examples and limitations

As a basic example of sophisticated learning, let's look at an (Fudenberg and Levine, 1998) example of strategic teaching. Consider the game matrix from the table 1.7, and suppose we have an FP column player and a sophisticated row player. A sufficiently patient row player can shift the equilibrium result to a Pareto efficient one. If she for a long time chooses U despite the fact that it is dominated by D, then the outcome will switch to $\{U, R\}$.

Table 1.7

	L	R
U	1,0	3,2
D	2,1	4,0

Here in 1.7 one of the players is assumed to be myopic but for example, in the Battle of the Sexes game, both players may "try to teach" and in that context, teaching may

[13] If the model only imitates an opponent's action then vice versa it is useful only in the PD game. If it combines both and switches between them then there are trajectories where it is useless for both games.

also be perceived as "being stubborn". If both players may be teaching and compete in doing it, we have a situation where we have two instances of the teaching process and one occasion of competition. Naturally, such a problem is considerably harder to model, estimate, and work with in general than a problem where who is teaching and who is learning is clear. To some extent, it is an open question e.g. is strategic teaching possible in the game 1.8.

Table 1.8: Extended Dilemma

	C	D
C	4,4	-1,12
D	12,-1	1,1

Here comparing between $s_1 = \{C^{t=1}, D^{t=2}, \dots, C^{t=n}\}$ and $s_2 = \{C^{t=1}, C^{t=2}, \dots, C^{t=n}\}$ first is preferable both in social and private outcome $U(s_1, s_1) > U(s_2, s_2)$. But how let you opponent know that your action directed to alternating and not to defecting?

We have already mentioned that myopia can be interpreted as a lack of forward-looking behavior. However in practice when we faced the ability to process the simple patterns it is hard to disentangle two of these concepts. Let us consider the situation in the strategy of conditional cooperation also called "Tit-For-Tat".

If a player starts to play tit-for-tat and the second player does likewise, there could be several interpretations: (1) the second player expects the first one to cooperate only if the second will cooperate itself and thus second cooperates in response because it is profitable (2) the second player has a bunch of patterns including "cooperate if opponent cooperates" and such pattern is reinforced by successful application.

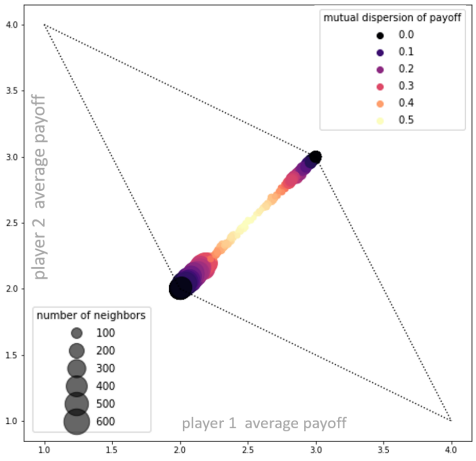
To some extent those narratives are the same, the second being more typical for the evolutionary approach. But we can also frame it as the agent evolves "experts" within itself and thus learns patterns.

This approach does not suffer from the drawbacks of pure teaching because here to learn both players need only an option to "defect if opponent defects" as a negative incentive.

To illustrate it we can consider two situations. In the first player with a tit-for-tat strategy faces a similar opponent and they converge on cooperation. In the second tit-for-tat player faces a simple RL. RL is known to "experiment" and will frequently deviate, but will get defection (thus, punishment) in return.

In a significant share of cases (25%) according to our simulations (see fig. 1.4) simple RL opponents are successfully taught. The ability to counter different types of forward-looking opponents alleviates the induction problem if the number of types is limited.

Figure 1.4: Convergence in average payoff for pair of different type of players: Tit-For-Tat automaton vs RL



Here the axes show the average payoff of a pair of players. 1000 pairs of players are taken such that the first pair is TFT and another is action RL. The length of the game is broken into blocks, with each block lasting 200 rounds. They interact until either the block limit (50) runs out or the deviation of the average payoff taken among each next 5 blocks is less than 0.01. The color shows whether the players have converged or not (100% black means the deviation is smaller than the delta; In other cases the darker the less the oscillation of the payoffs). The size of the circle represents what percentage of the population has converged in a small radius around it. So in this picture 24% of the players have winnings greater than 2.5. and 76% less.

For practical purposes, one source of this limitation is the natural limitation of human opponents' computational resources a la level-k or cognitive hierarchy theories (Stahl, 1993). In general, repeated strategies of fixed length were proposed to narrow the induction problem in evolutionary games via Deterministic Finite Automata (DFA) a long time ago Aumann (1987).

While in evolutionary games each automaton was identical to the player, later in learning literature automata were proposed as a building block for the learning model by (Hanaki, 2004) and later was implemented in algorithm by (Ioannou and Romero, 2014). Algorithms such as strategic EWA show the ability to react to some patterns of coordination (see 1.5).

Even though this is the only currently successful example of applying DFA to learning models, it is not without downsides. In particular, it requires pre-learning that makes it similar to neural nets and prevents fast transfer or immediate adaptability. Models use presimulation to set initial weights, finding initial propensities demands a lot of rounds simulated before play (the median convergence length was close to 16,800 periods for each of the games).

Moreover, there are strict requirements for action space. As a consequence of which when we discuss only games with 2 actions for both players it produces 26 versions of strategies ($2^4 * 2 - 6 = 32 - 6$ where 6 are trivial versions and 2 possible starting states). For a game with 3 actions, it becomes close to $3^9 * 2$.

Thus this model can be applied only to small-dimensional discrete action games.

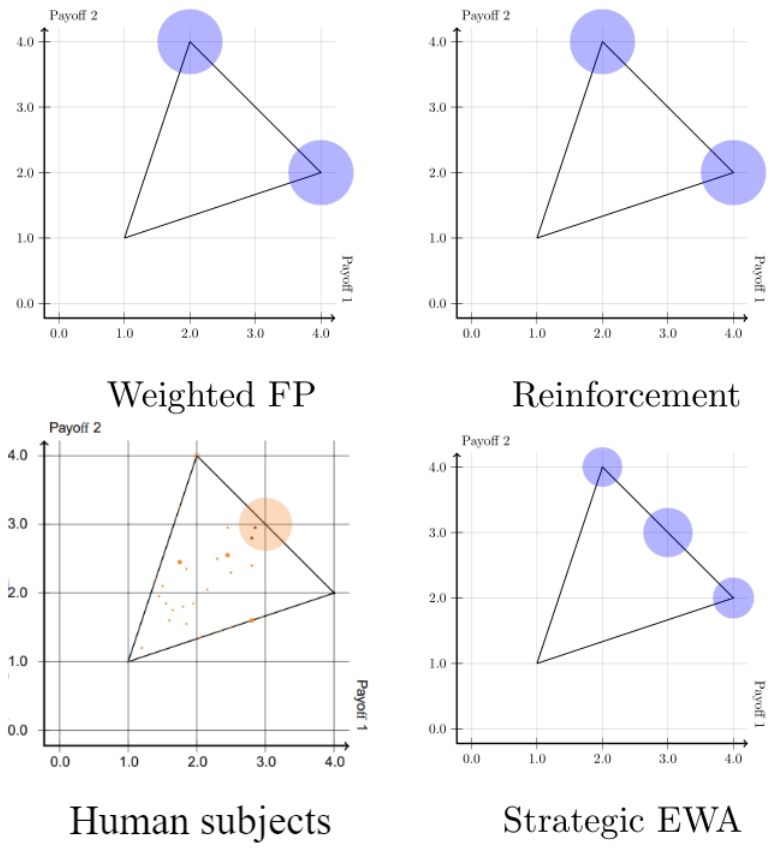


Figure 1.5: Convergence in average payoff for pair of different type of players by Mathevet and Romero (2012)

However, evidence of strategic teaching was investigated by (Duersch et al., 2010) and on the continuous action games in particular in the oligopoly game experiment, however, the behavior exhibited by subjects there has not been represented by any known theoretical learning model.

The authors in the experiment use human' play against several different computer algorithms that are trained according to one of the classical rules (fictitious play, trial & error, best response, imitation, reinforcement learning, EWA). The experimental setting included the linear inverse demand function $max\{109-Quantity, 0\}$ and constant marginal cost, MC = 1. The actions (quantities) are at the interval – $quantity = a \in [0, 100]$. Utility of each player is:

$$u(a_i, a_{-i}) := (max\{109 - a_i - a_{-i}, 0\} - 1) * a_i$$

Table 1.9: Prominent outcomes by theory

	a_i	a_{-i}	u_i	u_{-i}
Cournot-Nash equilibrium	36	36	1296	1296
Symmetric competitive outcome	54	54	0	0
Symmetric collusive outcome	27	27	1458	1458
Stackelberg leader outcome	54	27	1458	729
Stackelberg follower outcome	27	54	729	1458
Monopoly solution	54	0	2916	0

People manage to take the oligopoly from Cournot to Stackelberg by "unraveling" the process underlying your opponent's decision and using his rules. Why is behavior in an oligopoly not represented by a sophisticated learning model? One answer lies in the difficulty of representing strategies since in this setup the set of actions is a continuum. Typically, when trying to construct them, the complexity begins at the action representation stage because it is necessary to specify all variants of actions as a discrete set with steps (for example for the interval from 1 to 100 it can be a set of the sequence 1,2...100). However, in the latter, the algorithms learn too slowly and rather randomly (obviously this is far from human dynamics, as (Duersch et al., 2010) have shown.

1.6 Chapter summary

We usually assume rational preferences and rational behavior in game theory (i.e. multi-person decision theory), because making a decision rationally implies, in some sense, a maximization problem over several possible alternatives. Even when we know that there is more to be said about the underlying decision process, like small probabilities being undervalued in decisions under risk, the generic utility function can be a useful approximation. For example, von Neumann-Morgenstern utility function can fit the data better than the prospect theory (Harless and Camerer, 1994). This intrinsic attractiveness of general solutions explains why theorists spend so much time and effort to discover the general model of rational learning that would fit desirable theoretical properties such as universal consistency or convergence to Nash equilibrium. But results are a mixed bag - there is a following sequence of results from D.Foster's seminar[14] on whether convergence to Nash equilibrium can be achieved: Yes: You can learn NE from a grain of truth (Kalai and Lehrer, 1993). No: Not exactly. (Foster and Young, 2001) Yes: Via exhaustive search-i.e. very slowly (Foster and Young, 2006) No: (Hart and Mas-Colell, 2003) Yes: via public, deterministic calibration which is very slow (Kakade and Foster, 2008). For all but the smallest games, it is basically no. As we can see from this example, the theory is not very coherent in what happens in particular cases.

With too many agents individual input of each of them becomes relatively smaller. At some point they don't account for their own effect on the public outcome, so we can treat public outcome as an external state of the world. As (Marimon, 1996) points out: "In competitive environment an individual agent does not affect the social outcome and, therefore does not create strange correlations out of his optimal actions and mistakes".

[14] http://deanfoster.net/calibration_handout.pdf

However with a few players the situation is qualitatively similar to two-player case and all problems of induction remain.

In this chapter, we have discussed the consequences of these results that purely deterministic rationality cannot guarantee good theoretic properties. Naturally, learning theory turned away from simple deterministic algorithms, such as fictitious play, and towards simple randomized algorithms such as stochastic and weighted fictitious play algorithms. Such rules are appealing because they can play equally well against anyone and due to of randomisation prevent rough manipulation by opponent. However, they can be too crude and fail to detect even the simplest regularities in the opponent's behavior. Sometimes (as in zero-sum games) it is a minor issue but for coordination games this approach is significantly limited. While in zero-sum games patterns can be expected to be short-lived (being predictable will be punished), coordination should reinforce patterns, because being predictable helps to coordinate actions.

A possible trade-off between preventing manipulation on the one hand and pattern recognition on the other could be models involving complex cognitive strategies and randomization simultaneously such as (Ioannou and Romero, 2014) or (Spiliopoulos, 2012). The application of these models in broader than 2-by-2 games contexts, however, has some coincidental and empirical issues. Conceptually they are limited by the size of the subjects' computational capabilities. Empirically it is difficult to distinguish them from their simpler analogs.

However, behavioral models in the context of the question at hand have an advantage because they are limited from above in the complexity of the options they consider (after all, we are talking about players-humans whose beliefs about their opponent's behavior are limited by computational capabilities). In our view, research design could be aimed at investigation these intrinsic properties of players, rather than just predicting their behavior in a limited number of cases. For example, we don't know what human subjects will play in a game from tab. 1.8, nor whether there is a repeated strategy with a limited memory length capable of inducing coordination in a given game (and thus leading to a Pareto optimum). At the end, examining these properties of players in the context of learning refers to the big question of how agents process information in general. We believe that the small answers to these questions in learning that we raise here and in subsequent chapters are also small answers to big questions around information design in economic mechanisms per se.

Supplementary

1.6.1 Bayesian algorithm play against nature

Consider a numerically classical example of Bayesian learning — estimation a coin bias from observations. Let's formulate the game: "Nature creates a coin with head probability r and then flips that coin N times, generating a series of "actions". The second player guess one side before each action, take a 1 if successful, otherwise 0.

The series of values generated by Nature is described by the Bernoulli distribution, and the likelihood function is described by the binomial distribution with parameters θ: r which reflects the probability of occurring heads (coin bias), h - the number of heads in N periods and the number of tails in t (respectively, $N = t + h$).

That is, the probability that r takes some known value, given the known data $x(h = H)$ takes the form $Pr\left(h = H \mid r, h + t\right) = \binom{h+t}{h} r^h (1-r)^t$, where H is the eagle dropout value in the general population.

We denote the a priori distribution of the probability density function r by $g(r) \in [0, 1]$. The posterior distribution is obtained by the product of the likelihood functions $r \to p(t, h|r)$ and the a priori distribution $g(r)$ normalized by the probability of head occurring $p(t, h)$ in n trials:

$$p(r|t = T, h = H) = \frac{p(t, h|r)g(r)}{\int p(t, h|r')g(r')dr'} = \quad \frac{Pr(h = H|r, h + t)g(r)}{\int_0^1 Pr(h = H|r',\ h + t)g(r')dr'}$$

Putting the binomial distribution formula into the posterior distribution formula, we obtain:

$$f(r|t = T, h = H)) = \frac{\binom{h+t}{h} r^h (1-r)^t}{\int_0^1 \binom{h+t}{h} r'h(1-r')dr'} = \frac{r^h (1-r)^t}{\int_0^1 r'h(1-r')^t dr'}$$

This distribution, conjugate a priori for the binomial of the distribution, is called the beta distribution, in the general case its denominator is expressed through the beta function:

$$f\left(r \mid t = T, h = H\right) = \frac{1}{\mathrm{B}\left(h + 1, t + 1\right)} r^h (1-r)^t.$$

When the player maximizes her payoff, she should to guess the side with the higher (according to the a priori distribution) probability. Assume that the a priori distribution r is– uniform on [0, 1], i.e. $g\left(r\right) = 1$. Let's take the following game state as an example, let $n = 10,\ h = 7$, that is, the coin is tossed 10 times and there are 7 heads. What to choose on the 11th move? Since h and t are integers, and the a priori distribution is — uniform, the formula for the posterior beta distribution can also be written in factorials:

$$f\left(r \mid t = T, h = H\right) = \frac{(t + h + 1)!}{\mathrm{h!t!}} r^h (1-r)^t = \frac{(10 + 1)!}{7!3!} r^7 (1-r)^3 = 1320 r^7 (1-r)^3$$

$f\ (r \mid H = 7, T = 3)$ reaches its peak at $r = h\ /\ (h + t) = 0, 7$ The expected value of r for a given distribution is:

$$\mathbb{E}(r) = \int_0^1 r * f(r|H = 7, T = 3)dr = \frac{h + 1}{h + t + 2} = \frac{2}{3}$$

This means that the Bayesian player, choosing the most likely event, must bet on the head in period 11.

Bibliography

(1) J. Arifovic and J. Ledyard. Scaling up learning models in public good games. *Journal of Public Economic Theory*, 6(2):203–238, 2004.

(2) R. J. Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society, pp*, pages 1–18, 1987.

(3) A. Beggs. On the convergence of reinforcement learning. *Journal of economic theory*, 122:1–36, 2005.

(4) M. Benaim and M. Hirsch. Mixed equilibria and dynamical systems arising from fictitious play in perturbed games. *Games and Economic Behavavior*, 72:29–36, 1999.

(5) D. Bergemann and J. Valimaki. Bandit problems. *The New Palgrave Dictionary of Economics:*, 1(8):336–340, 2008.

(6) G. W. Brown. Iterative solutions of games by fictitious play, m in activity analysis of production and allocation, ed, by t. koop# mans. page 376. New York: Wiley, 1951.

(7) S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122, 2012.

(8) R. R. Bush and F. Mosteller. *Stochastic models for learning.* John Wiley & Sons, Inc., 1955.

(9) C. Camerer and H. Ho. Experience-weighted attraction learning in normal form games. *Econometrica*, 67(4):827–874, 1999.

(10) T. N. Cason and D. Friedman. Learning in a laboratory market with random supply and demand. *Experimental Economics*, 2(1):77–98, 1999.

(11) Y. Cheung and D. Friedman. Individual learning in normal form games: Some laboratory results. *Games and Economic Behavior*, 19:46–76, 1997.

(12) T. Chmura, S. J. Goerg, and R. Selten. Learning in experimental 2× 2 games. *Games and Economic Behavior*, 76(1):44–73, 2012.

(13) A. Cournot. *Researches into the Mathematical Principles of the Theory of Wealth, trans.* Kelly, N. Bacon. New York, 1960.

(14) P. Duersch, A. Kolb, and J. Oechssler. Rage against the machines: how subjects play against learning algorithms. *Economic Theory*, 43(3):407–430, 2010.

(15) H. Ebbinghaus. Memory: a contribution to experimental psychology. *Annals of neurosciences*, 20(4):155, 1885/1974.

(16) I. Erev and A. Roth. Predicting how people play games: reinforcement learning in games with unique strategy mixed-strategy equilibrium. *American Economic Review*, 88:848–881, 1998.

(17) I. Erev, D. Gopher, R. Itkin, and Y. Greenshpa. Toward a generalization of signal detection theory to n-person games: The example of two-person safety problem. *Journal of Mathematical Psychologyn*, 39(4):360–375, 1995.

(18) W. K. Estes. Probability learning. In *Categories of human learning*, pages 89–128. Elsevier, 1964.

(19) D. Foster and H. P. Young. On the impossibility of predicting the behavior of rational agents. *Proceedings of the National Academy of Sciences of the USA*, 98(22): 12848–12853, 2001.

(20) D. Foster and P. Young. Regret testing: Learning to play nash equilibrium without knowing you have an opponent. *Theoretical Economics*, 1(3):341–367, 2006.

(21) D. P. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

(22) D. P. Foster and R. V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.

(23) D. P. Foster and H. P. Young. Learning, hypothesis testing, and nash equilibrium. *Games and Economic Behavior*, 45(1):73–96, 2003.

(24) D. Fudenberg and D. Levine. Steady state learning and Nash equilibrium. *Econometrica: Journal of the Econometric Society, pp*, pages 547–573, 1993.

(25) D. Fudenberg and D. Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

(26) D. Fudenberg and D. Levine. Learning and equilibrium. *Annual Review of Economics*, 1(1):385–420, 2009.

(27) D. Fudenberg and D. Levine. Whither game theory? towards a theory of learning in games. *Journal of Economic Perspectives*, 30(4):151–170, 2016.

(28) D. Fudenberg and J. Tirole. *Game theory*. 1991. Cambridge, Massachusetts, 393(12), 80, 1991.

(29) J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 148–177, 1979.

(30) S. Goyal and M. Janssen. Can we rationally learn to coordinate? *Theory and Decision*, 40(1):29–49, 1996.

(31) W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3:367–388, 1982.

(32) N. Hanaki. Action learning versus strategy learning. *Complexity*, 9(5):41–50, 2004.

(33) J. Hannan, M. Dresher, A. Tucker, and P. Wolfe. Approximation to bayes risk in repeated play. In *Contributions to the Theory of Games*, pages 97–139. Princeton Univ. Press, Princeton, 1957.

(34) D. W. Harless and C. Camerer. The predictive utility of generalized expected utility theories. *Econometrica*, 62(6):1251–89, 1994.

(35) S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98:26–54, 2001.

(36) S. Hart and A. Mas-Colell. Uncoupled dynamics do not lead to Nash equilibrium. *American Economic Review*, 93:1830–1836, 2003.

(37) T. Ho, C. F. Camerer, and J. Chong. Self-tuning experience weighted attraction learning in games. *Journal of Economic Theory*, 133:177–198, 2007.

(38) J. Hofbauer and E. Hopkins. Learning in perturbed asymmetric games. *Games and Economic Behavior*, 52(1):133–152, 2005.

(39) C. A. Ioannou and J. Romero. A generalized approach to belief learning in repeated games. *Games and Economic Behavior*, 87:178–203, 2014.

(40) S. M. Kakade and D. P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

(41) E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61:1019–1045, 1993.

(42) R. Marimon. Learning from learning in economics. In *Advances in economic theory:the 7th Word Congress. CUP*. European University Institute, 1996.

(43) L. Mathevet and J. Romero. Predictive repeated game theory: Measures and experiments. 2012.

(44) R. D. McKelvey and T. R. Palfrey. *Playing in the dark: Information, learning, and coordination in repeated games*. California Institute of Technology, California, 2001.

(45) K. Miyasawa. On the convergence of the learning process in a 2 x 2 non-zero-sum game. *Economic Research Program, Princeton University, Research Memorandum No*, 33, 1961.

(46) J. M. J. Murre and J. Dros. Replication and analysis of ebbinghaus forgetting curve. *PLOS ONE*, 10:7, 2015. doi: 10.1371/journal.pone.0120644.

(47) J. Nachbar. Learning in games. In Springer, editor, *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*, pages 485–498. Springer, New York, 2020.

(48) J. H. Nachbar. Evolutionary selection dynamics in games: Convergence and limit properties. *International journal of game theory*, 19(1):59–89, 1990.

(49) J. H. Nachbar. Beliefs in repeated games. *Econometrica*, 73(2):459–480, 2005.

(50) R. Nagel. Unraveling in guessing games: An experimental study. *AmericanEconomic Review*, 85(5):1995, 1995.

(51) J. Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54: 296–301, 1951.

(52) M. Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9:185–202, 1974.

(53) A. Sadrieh. *The alternating double auction market: A game theoretic and experimental investigation*, volume 466. Springer, Science & Business Media, 1998.

(54) T. C. Salmon. An evaluation of econometric models of adaptive learning. *Econometrica*, 69(6):1597–1628, 2001.

(55) L. J. Savage. *The Foundations of Statistics*. New York, Wiley, 1954.

(56) R. Selten. Learning direction theory and impulse balance equilibrium. In *Economics Lab*, pages 147–154. Routledge, 2004.

(57) R. Selten and J. Buchta. Experimental sealed bid first price auctions with directly observed bid functions. In Z. Budescu, Erev, editor, *Games and Human Behavior*, pages 79–104. Essays in the Honor of Amnon Rapoport. Lawrenz Associates,, Lawrenz A. and Mahwah, N.J., 1999.

(58) R. Selten and R. Stoecker. End behavior in sequences of finite prisoner's dilemma supergames a learning theory approach. *Journal of Economic Behavior & Organization*, 7(1):47–70, 1986.

(59) R. Selten, T. Abbink, and R. Cox. Learning direction theory and the winners curse. *Experimental Economics*, 8(1):5–20, 2005.

(60) R. Selten, T. Chmura, T. Pitz, S. Kube, and M. Schreckenberg. Commuters route choice behaviour. *Games and Economic Behavior*, 58(2):394–406, 2007.

(61) L. S. Shapley. Some topics in two-person games. In M. Dresher, L. S. Shapley, and A. W. Tucker, editors, *Advances in Game Theory*, pages 1–28. Princeton University Press, 1964.

(62) L. Spiliopoulos. Pattern recognition and subjectie belief learning in a repeated constant-sum game. *Games and Economic Behavior*, 75:921–35, 2012.

(63) D. O. Stahl. Evolution of smartn players. *Games and Economic Behavior*, 5(4): 604–617, 1993.

(64) R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction. 2nd edition.* MIT press, 2018.

(65) E. L. Thorndike. *Animal Intelligence*. Macmillan, New York, 1911.

(66) E. L. Thorndike. The law of effect. *American Journal of Psychology*, 39:212–222, 1927.

(67) E. Van Damme. *Stability and perfection of Nash equilibria*, volume 339. Springer-Verlag, Berlin, 1991.

(68) J. B. Watson and G. A. Kimble. *Behaviorism*. Routledge, 2017.

(69) H. P. Young. *Strategic learning and its limits*. Oxford University Press, Oxford, 2004.