Learning in games
using the imprecise Dirichlet model *

Erik Quaeghebeur [1,*] Gert de Cooman

SYSTeMS research group, EESA department, Ghent University.
Technologiepark-Zwijnaarde 914, 9052 Zwijnaarde, Belgium.

Abstract

We propose a new learning model for finite strategic-form two-player games based on fictitious play and Walley's imprecise Dirichlet model (1996, J. Roy. Statistical Society B 58, 3–57). This model allows the initial beliefs of the players about their opponent's strategy choice to be near-vacuous or imprecise instead of being precisely defined. A similar generalization can be made as the one proposed by Fudenberg and Kreps (1993, Games Econ. Behav. 5, 320–367) for fictitious play, where assumptions about immediate behavior are replaced with assumptions about asymptotic behavior. We also obtain similar convergence results for this generalization: if there is convergence, it will be to an equilibrium.

Key words: learning, fictitious play, imprecise Dirichlet model, imprecise probability models, two-player games, decision making
MSC: 91A26, 62F15, 91A35

1 Introduction

This paper describes a new approach to learning for finite strategic-form two-player games in the setting of fictitious play. So, consider successive rounds of a game.

Suppose we are in the t-th round, and that at this point, each player has certain beliefs about which strategy his opponent will play in the next, $(t + 1)$-th, round.

After each round of the game the players change their beliefs. Indeed, we assume that each player can observe the strategy his opponent plays, which naturally leads to the setting of so-called fictitious play [2,11]. After each round, this new information affects the player's beliefs about the strategy his opponent will play next. This updating of beliefs is what we call *learning about the opponent's strategy choice*.

In a Bayesian context, such beliefs are represented by a probability distribution on the opponent's strategies (a precise Dirichlet model or PDM), and learning is implemented by updating with Bayes's rule using a likelihood function relating observations to future strategies. The model we propose for representing and updating these beliefs generalizes this Bayesian approach. It is based on Walley's imprecise Dirichlet model or IDM [13]. Using this generalization is justified by the fact that sometimes the assumptions underlying the Bayesian approach are too strong. The most visible difference is that the beliefs are summarized by a convex set of expected mixed strategies, instead of only one.

Now, given such beliefs about his opponent's next strategy choice, and the game's payoff function, which strategy should a player use in the next round to satisfy some optimality criterion? In a Bayesian context, optimal strategies maximize a player's expected payoff. As optimality criteria, we propose two generalizations of the concept of expectation maximization. The main consequence is that – in contrast to the Bayesian context – two different optimal strategies may be incomparable.

If both players use a method of learning and of choosing optimal strategies during the successive rounds of a game, will their behavior converge: Will the optimal strategies they select converge to an equilibrium of the game?[2] Or, when convergence occurs, will it be to an equilibrium? We investigate whether some interesting, existing results [5] for players using a PDM-based learning model can be generalized to the case where the players use our IDM-based learning model.

How is this paper organized? In Sec. 2 we describe its game-theoretic setting and introduce basic game-theoretic concepts and notation. In Sec. 3, we look at how a player can represent and update his beliefs about his opponent's strategy. Both the PDM-based Bayesian model and our IDM-based generalization are improvements of the classical model of fictitious play [2,11]. In Sec. 4 we discuss the player's options for deciding on an optimal strategy. Most importantly, we investigate options that can be used in conjunction with the PDM-based and IDM-based models. Our contribution up until this point is mainly to show how well imprecise probability theory can be allied with game theory. Then, in Sec. 5, before concluding, this alliance is used to generalize results by Fudenberg and Kreps [5] about the convergence of play to equilibria. All proofs are collected in the Appendix at the end.

----
[2] In an equilibrium, no player can increase his payoff by unilaterally changing his strategy.

2 Basic concepts and notation

We consider two-player games and use i as an index for a player and $-i$ as an index for his opponent. This allows for player-neutral notation and formulas.

Both players have a finite set $S^i = \{1, \ldots, N^i\}$ of N^i *pure strategies* s^i, each of which labels a possibly complex description. When the choice of pure strategy is determined through some form of randomization, the player uses a *mixed strategy* $\sigma^i \colon S^i \to [0,1]$. We use $\sigma^i(s^i)$ to denote the chance that s^i is played when σ^i is used. (We use the word *chance* when talking about uncertainty due to randomization. When discussing the uncertainty of a player, we use the word *probability*.) The set of all mixed strategies Σ^i forms a unit simplex in $\mathbb{R}^{N^i}$, where the vertices correspond to the pure strategies and mixed strategies to the convex combinations with weights $\sigma^i(s^i)$, i.e., $\sum_l \sigma^i(l) = 1$. All quantities that can be interpreted as a mixed strategy are denoted with a lower case Greek letter. From now on, 'strategy' implicitly means 'mixed strategy', unless explicitly stated otherwise.

A *strategy profile* is a couple of two strategies, one for each player. The notation for pure and mixed strategy profiles is defined and illustrated in Fig. 1.
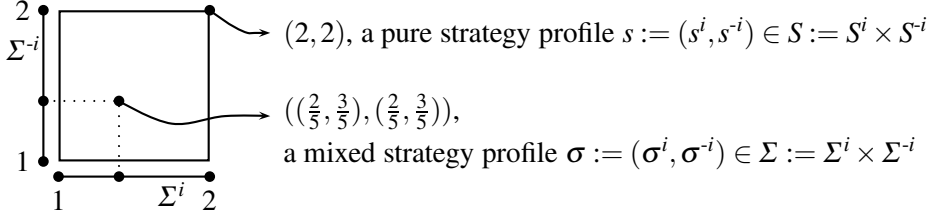

$(2,2)$, a pure strategy profile $s := (s^i, s^{-i}) \in S := S^i \times S^{-i}$

$((\frac{2}{5}, \frac{3}{5}), (\frac{2}{5}, \frac{3}{5}))$,
a mixed strategy profile $\sigma := (\sigma^i, \sigma^{-i}) \in \Sigma := \Sigma^i \times \Sigma^{-i}$

Fig. 1. The set Σ of strategy profiles for players with two pure strategies each and two of its elements: one mixed, σ, and one pure, s.

Since we assume that the randomization mechanisms both players use are independent, the chance that s is played when σ is used, is given by $\sigma(s) := \sigma^i(s^i)\sigma^{-i}(s^{-i})$.

Each round of a game consists of a pure strategy being selected by the players. After every round, each player receives a bounded *payoff* $u^i(s)$ expressed in some linear utility [12]. When the players use mixed strategies, the expected payoff becomes

$$u^i(\sigma) = u^i(\sigma^i, \sigma^{-i}) := \sum_{s \in S} \sigma(s)u^i(s) = \sum_{(s^i, s^{-i}) \in S} \sigma^i(s^i)\sigma^{-i}(s^{-i})u^i(s^i, s^{-i}). \quad (1)$$

During each round, the player chooses a strategy based on assumptions about his expected payoff. With each of his own strategies σ^i there corresponds an *unknown payoff* u_{σ^i} which is a function relating his opponent's (still unknown) strategy choice σ^{-i} to the corresponding (expected) payoff:

$$u_{\sigma^i} \colon \Sigma^{-i} \to \mathbb{R} \colon \sigma^{-i} \mapsto u_{\sigma^i}(\sigma^{-i}) = u^i(\sigma^i, \sigma^{-i}). \quad (2)$$

Such an unknown payoff is a real random variable on Σ^{-i} which we call a *gam-*

ble [12]: choosing a strategy is like participating in a lottery. When choosing his strategy, the player is unsure about the amount of utility he is going to win (or lose).

The game is played repeatedly, and during each round t the players observe the pure strategy profile $s_t := (s_t^i, s_t^{-i})$ that is actually played in that round. The mixed strategy profile that is played cannot be observed, i.e., the opponent's mixed strategy choice remains hidden. All the pure strategy profiles that were played up to and during round t form the *history* after t rounds $h_t = (s_1, \dots, s_k, \dots, s_t)$. Given a history h_t, the pure strategy profile played during round $t' \leq t$ is $h_t(t')$. All the possible histories after t rounds form the Cartesian product set $\mathcal{H}_t := S^t$. When considering an unending number of rounds, we talk about an infinite history $h_\infty = (s_1, s_2, \dots)$, which is an element of $\mathcal{H}_\infty$. The set of all possible histories is $\mathcal{H} := \bigcup_t \mathcal{H}_t$.

Single player histories are also used. These are written down using superscripts as h_t^i, or h_t^{-i} for the opponent. This is done similarly for the other history concepts.

The opponent's history h_t^{-i} can be summarized by how many times the opponent has played each of his pure strategies. This *observed strategy count* is formalized with the function $n^{-i} \colon \mathcal{H}^{-i} \times S^{-i} \to \mathbb{N} \colon (h_t^{-i}, s^{-i}) \mapsto n^{-i}(h_t^{-i}, s^{-i})$, where $n^{-i}(h_t^{-i}, s^{-i})$ is the total number of rounds the opponent has played s^{-i} in the history h_t^{-i}. It immediately follows that $\sum_{s^{-i} \in S^{-i}} n^{-i}(h_t^{-i}, s^{-i}) = t$. When the history h_t^{-i} under consideration is implicit, we use the notation n_t^{-i} for $n^{-i}(h_t^{-i}, \cdot)$. Similarly, we can also consider v^{-i}, the *(relative) frequency of observed strategies*, where $v^{-i}(h_t^{-i}, s^{-i}) = n^{-i}(h_t^{-i}, s^{-i})/t$ is the total fraction of the rounds the opponent played s^{-i} in the history h_t^{-i}. Again, when the history under consideration is implicit, we can write v_t^{-i} for n_t^{-i}/t. We also use count and frequency profiles $n := (n^i, n^{-i})$ and $v := (v^i, v^{-i})$.

3 Assessing the opponent's strategy

In this section, we look at what a player (thinks he) knows about his opponent's strategy choice and how to model this belief. We start out in Sec. 3.1 with the basic assumptions made by the player. Then, we consider possible belief models and how to update them using observations of the pure strategies played during previous rounds in Secs. 3.2 and 3.3, where we introduce the imprecise Dirichlet model. Finally, in Sec. 3.4, we link so-called assessment rules to these belief models.

3.1 Modelling the opponent

In the classical model of *fictitious play* [2,11], the players base their strategy choice on their opponent's so-called accumulated mixed strategy. This corresponds to using v_t^{-i}, the frequency of observed pure strategies played by the opponent in the first t rounds. (Note that this can be interpreted as a strategy of the opponent.)

To make the basic assumption of this model explicit [7, Ch. 2], we can say that *each player is convinced that his opponent is playing a fixed mixed strategy* that is (initially completely) unknown to him. This implies that the players do not try to influence their opponent's strategy choices. They do try to learn as much as they can about this unknown strategy from the observation of their opponent's play. Because of this assumption, the order of the strategies played by the opponent is irrelevant, and the only observational data a player uses is the sufficient statistic n^{-i} or, equivalently, t and v^{-i}.

So, under the assumptions of fictitious play, the opponent is modelled as an unknown *iid-process* (identical independent draws). A player could imagine his opponent drawing marbles from a bag (sampling with replacement from a finite set). Marble types then correspond to pure strategies of the opponent and the relative frequency of each type to the unknown mixed strategy played by the opponent.

Considering the model given for the opponent, just using the frequency of observed strategies v^{-i} as the assessment for his strategy seems natural, but is nevertheless problematic: What is the rationale behind this choice of an assessment, and what is the justification for modifying it once a new round has been played? Classically, an interpretation is given by assuming that the players use a form of Bayesian inference [7, Ch. 2]. This will be described and investigated in the following subsection.

3.2   The precise Dirichlet model

The assessments made in the model of fictitious play are formulated in the framework of Bayesian inference as follows. The player considers all of his opponent's mixed strategies to be possible a priori. So he uses a *prior* probability density function over his opponent's simplex to model his uncertainty about his opponent's mixed strategy choice – assumed to be fixed in time. After observing the pure strategy played by his opponent in that round, he can construct a *likelihood function*. Using Bayes's rule, the prior and the likelihood function are combined to form a *posterior* probability density function over his opponent's simplex. This posterior then models his uncertainty about the opponent's strategy choice, after having observed the pure strategy played by the opponent.

Because the opponent is modelled as an iid-process and has a finite number of pure strategies, the likelihood function, defined on Σ^{-i}, takes the form of a multinomial probability mass function $L(\sigma^{-i}|n_t^{-i}) \propto \prod_{s^{-i} \in S^{-i}} \sigma^{-i}(s^{-i})^{n_t^{-i}(s^{-i})}$. For any history h_t^{-i}, it gives the chance that n_t^{-i} was produced by an opponent using a mixed strategy σ^{-i}.

The basis for our models is the *Dirichlet density* $D(\cdot|r,\rho^{-i})$. It is defined on the interior $\text{int}(\Sigma^{-i})$ of the unit simplex Σ^{-i} by $D(\sigma^{-i}|r,\rho^{-i}) \propto \prod_{s^{-i} \in S^{-i}} \sigma^{-i}(s^{-i})^{r\rho^{-i}(s^{-i})-1}$, where $r \in \mathbb{R}^+$ and $\rho^{-i} \in \text{int}(\Sigma^{-i})$. The family of Dirichlet densities can take on a

variety of shapes, depending on the choice of parameters and thus represent a variety of (prior) beliefs. This is the reason for using this family, together with them being conjugate for multinomial sampling, resulting in posteriors from the same family that are easily obtained. The linear prevision (expectation functional) $P(\cdot|r,\rho^{-i})$ associated with a Dirichlet density $D(\cdot|r,\rho^{-i})$ is called a *precise Dirichlet model* or PDM. It is defined on measurable gambles (real valued random variables on Σ^{-i}). The prevision of $\sigma^{-i}(s^{-i})$, i.e., the prevision of the chance that the opponent plays s^{-i},

$$P(\sigma^{-i}(s^{-i})|r,\rho^{-i}) = \int_{\text{int}(\Sigma^{-i})} \sigma^{-i}(s^{-i})D(\sigma^{-i}|r,\rho^{-i})\mathrm{d}\sigma^{-i}, \qquad (3)$$

turns out to be equal to the parameter $\rho^{-i}(s^{-i})$. This implies that $P(\sigma^{-i}|r,\rho^{-i})$, the player's expected value for the strategy played by the opponent, is ρ^{-i}.

A prior PDM, $P(\cdot|r_0,\rho_0^{-i})$, is the model used by the player to represent his initial uncertainty about his opponent's strategy. The parameters' subscript (0 here) indicates the number of observations on which the model is based. This prior model can be *updated* to a posterior model after one or more rounds of the game, i.e., after observing n_t^{-i}. This amounts to normalizing the product of the prior $D(\cdot|r_0,\rho_0^{-i})$ and the likelihood function $L(\cdot|n_t^{-i})$, resulting in a new Dirichlet density $D(\cdot|r_t,\rho_t^{-i})$, with updated parameters [3]

$$r_t = r_0 + t \quad \text{and} \quad \rho_t^{-i} = \frac{r_0\rho_0^{-i}+n_t^{-i}}{r_0+t} = \frac{r_0}{r_0+t}\rho_0^{-i} + \frac{t}{r_0+t}v_t^{-i}. \qquad (4)$$

The posterior PDM, $P(\cdot|r_t,\rho_t^{-i})$, which represents the updated uncertainty about his opponent's mixed strategy after observing n_t^{-i}, is thus easily obtained by updating the parameters as shown above in Eq. (4).

In Eq. (4), we see that the expected strategy is a convex mixture of the expected strategy ρ_0^{-i} chosen prior to any observation and the frequency of observed strategies v_t^{-i}. As t becomes large relative to r_0, the expected strategy is mainly determined by the observations. Considering that all observations have equal weight, this means that r_0 can be interpreted as the number of imaginary rounds that determine ρ_0^{-i} and is thus related to the trust the player has in his initial choice.

The problem the players are now faced with is how to choose the prior parameters r_0 and ρ_0^{-i}. If the player initially knows nothing about his opponent, every choice seems arbitrary. For any choice of prior PDM, an initial expected strategy ρ_0^{-i} is fixed. This has very strong behavioral implications, as we shall see in Sec. 4.2: it yields an essentially unique optimal strategy for the player to follow in response to the opponent's strategy, and we feel this to be unwarranted given the player's initial ignorance about what his opponent will do. Therefore, we are of the opinion that by using this classical approach, the player initially assumes more than he actually knows, and uses an assessment model that is too precise and is therefore not a good model for prior ignorance. In the next subsection, we propose a model that alleviates this arbitrariness by allowing for some imprecision.

---

[3] Correspondence with Walley's [13] notation: $t \leftrightarrow N$, $r_0 \leftrightarrow s$, $\rho_0^{-i} \leftrightarrow t$, and $n_t^{-i} \leftrightarrow n$.

3.3   The imprecise Dirichlet model

We propose that initially, when we have no information concerning our opponent's strategy, all $\rho_0^{-i} \in \Sigma^{-i}$ are considered as possible initial expected strategies. To express this in mathematical terms, we use the set of Dirichlet densities $\mathscr{D}(r_0, \mathscr{R}_0^{-i}) = \{D(\cdot | r_0, \rho_0^{-i}) : \rho_0^{-i} \in \mathscr{R}_0^{-i}\}$, where $\mathscr{R}_0^{-i} = \mathrm{int}(\Sigma^{-i})$, removing the arbitrary choice of an initial ρ_0^{-i}. A choice for r_0 still has to be made, however.

Instead of working with a linear prevision, we now work with a *lower prevision* $\underline{P}$ and an *upper prevision* $\overline{P}$, defined as the lower and upper envelopes (i.e., infima and suprema) of the set of linear previsions $\mathscr{P}' = \{P(\cdot | D) : D \in \mathscr{D}\}$. Because $\overline{P}(\cdot | \mathscr{D}) = -\underline{P}(-\cdot | \mathscr{D})$, the upper prevision is implicitly known when we only specify a lower prevision. Instead of working with $\mathscr{P}'$, we work with the closed convex hull $\mathscr{P} = \overline{\mathrm{co}}(\mathscr{P}')$ for mathematical practicality. The lower and upper envelopes do not change, and there is a one-to-one relationship between lower previsions and closed convex sets of previsions. Walley [12] gives an extensive treatment of imprecise probability models such as lower and upper previsions.

For any subset $\mathscr{R}^{-i}$ of the interior $\mathrm{int}(\Sigma^{-i})$ of the opponent's simplex, the lower prevision $\underline{P}(\cdot | r, \mathscr{R}^{-i})$ is called an *imprecise Dirichlet model* or IDM [13]. We use such models as generalizations of PDMs in order to represent the player's knowledge about his opponent's strategy. The lower prevision of the opponent's mixed strategy σ^{-i} is calculated component-wise as $\underline{P}(\sigma^{-i}(s^{-i}) | r, \mathscr{R}^{-i}) = \inf_{\rho^{-i} \in \mathscr{R}^{-i}} P(\sigma^{-i}(s^{-i}) | r, \rho^{-i}) = \inf_{\rho^{-i} \in \mathscr{R}^{-i}} \rho^{-i}(s^{-i})$. So, loosely speaking, the expected chance for the opponent to play s^{-i} is at least $\inf_{\rho^{-i} \in \mathscr{R}^{-i}} \rho^{-i}(s^{-i})$ and, similarly, not higher than $\sup_{\rho^{-i} \in \mathscr{R}^{-i}} \rho^{-i}(s^{-i})$.

The prior IDM, $\underline{P}(\cdot | r_0, \mathscr{R}_0^{-i})$, with $\mathscr{R}_0^{-i} = \mathrm{int}(\Sigma^{-i})$, is the model used by the player to represent his initial uncertainty about his opponent's mixed strategy. The choice $\mathscr{R}_0^{-i} = \mathrm{int}(\Sigma^{-i})$ has been argued [13] to result in a good model for prior ignorance. Additional prior information could correspond to a more specific choice for $\mathscr{R}_0^{-i}$. Using regular extension [12, App. J] – as is done implicitly by Walley [13, Sec. 2.3] –, we can update this model after one or more rounds of the game, i.e., after observing n_t^{-i}. This amounts to updating every linear prevision in $\mathscr{P}(r_0, \mathscr{R}_0^{-i})'$ as shown in Sec. 3.2, which results in an updated set of linear previsions $\mathscr{P}(r_t, \mathscr{R}_t^{-i})'$, where $r_t = r_0 + t$ and

$$\mathscr{R}_t^{-i} = \frac{r_0 \mathscr{R}_0^{-i} + n_t^{-i}}{r_0 + t} := \left\{ \rho_t^{-i} = \frac{r_0 \rho_0^{-i} + n_t^{-i}}{r_0 + t} : \rho_0^{-i} \in \mathscr{R}_0^{-i} \right\} = \frac{r_0}{r_0 + t} \mathscr{R}_0^{-i} + \frac{t}{r_0 + t} v_t^{-i}. \qquad (5)$$

The corresponding updated IDM is $\underline{P}(\cdot | r_t, \mathscr{R}_t^{-i})$. When we consider that $\mathscr{P}(r_t, \mathscr{R}_t^{-i}) = \overline{\mathrm{co}}(\mathscr{P}(r_t, \mathscr{R}_t^{-i})')$, the set of possible expected strategies is the closure $\mathrm{cl}(\mathscr{R}_t^{-i})$ of $\mathscr{R}_t^{-i}$. This set is convex and compact. Initially it is $\mathrm{cl}(\mathscr{R}_0^{-i}) = \mathrm{cl}(\mathrm{int}(\Sigma^{-i})) = \Sigma^{-i}$.

As is shown at the end of Eq. (5), the expression for the set of possible expected strategies is a convex mixture of (i) the set $\mathscr{R}_0^{-i}$, representing our initial ignorance,

whose weight decreases as t increases (as more observations become available), (ii) the frequency of observed strategies v_t^{-i}, whose weight increases as t increases. It is similar to the one found for a PDM in Eq. (4). Again, r_0 can be seen as the weight accorded to the initial beliefs.

3.4 Assessment rules

A player's assessments about his opponent's strategy after any round t are modelled by either a PDM $P(\cdot|r_t, \rho_t^{-i})$ or an IDM $\underline{P}(\cdot|r_t, \mathscr{R}_t^{-i})$, depending on which type of prior uncertainty model he uses. These previsions contain all the information the player has about the unknown mixed strategy he thinks his opponent is playing.

Fudenberg and Kreps [5] introduce the concept of an *assessment rule* μ^{-i}. It determines the opponent's strategies the player believes are most likely to be played in the next round, based on the observed history h_t^{-i} and some initial beliefs. Because we are focusing on the models used to represent the assessment, we emphasize that the assessment rule is a function of the current parameters' value – denoted generically by q_t –, belonging to a model-dependent set $\mathscr{Q}_t$ of possible values. So an assessment rule is defined as the map $\mu^{-i} \colon \mathscr{Q}_t \to \wp(\Sigma^{-i}) \colon q_t \mapsto \mu^{-i}(q_t)$, where $\wp(\Sigma^{-i})$ denotes the power set of Σ^{-i}. Whenever the parameters are implicit, we use $\mu_t^{-i} = \mu^{-i}(q_t)$.

In contrast to Fudenberg and Kreps [5], we allow for assessment rules that correspond to more than one mixed strategy, i.e., that are *set-valued*. Such a set contains all the opponent's strategies the player believes are most likely to be played in the next round. No distinction is made between strategies in the set, but they are all considered more likely than the strategies that are not in the set. We also use profiles $\mu = (\mu^i, \mu^{-i})$ of assessment rules, which then correspond to a subset of Σ.

When the player uses a PDM, $q_t = (r_t, \rho_t^{-i})$ and – identifying singletons with elements – we have $\mu_t^{-i} = \rho_t^{-i}$. When he uses an IDM, $q_t = (r_t, \mathscr{R}_t^{-i})$ and $\mu_t^{-i} = \mathrm{cl}(\mathscr{R}_t^{-i})$. Classical fictitious play as described by Brown [2] and Robinson [11] can be seen as a limit case for $r_0 \to 0$ of both the PDM and the IDM. In this case $\mu_t^{-i} = v_t^{-i}$.

When considering extensions to fictitious play, Fudenberg and Kreps [5] defined some classes of assessment rules. To make them fit in the more general context of this paper, we reformulate them to allow for set-valued assessment rules. A profile of assessment rules belongs to a certain class if both its components belong to it.

An assessment rule μ^{-i} is *adaptive* if it attaches diminishing importance to earlier parts of the history, as the number of rounds increases. Formally: for all t and all $\varepsilon > 0$

$$\exists t^* > t \colon \forall t' > t^* \colon \forall h_{t'}^{-i} \in \mathscr{H}_{t'}^{-i} \colon \forall \sigma^{-i} \in \mu_{t'}^{-i} \colon \ \sigma^{-i}(s^{-i}) < \varepsilon, \qquad (6)$$

for every pure strategy s^{-i} that was not played in the last $t' - t$ rounds.

8

Among all adaptive assessment rules, those that converge in some sense to the frequency of observed strategies v_t^{-i} are of special interest. An assessment rule is called *asymptotically empirical* if for every infinite history h_∞^{-i} with partial histories h_t^{-i}, it holds that

$$\lim_{t \to \infty} \sup_{\sigma^{-i} \in \mu_t^{-i}, s^{-i} \in S^{-i}} |v^{-i}(h_t^{-i}, s^{-i}) - \sigma^{-i}(s^{-i})| = 0. \tag{7}$$

From Eqs. (4) and (5) it can be deduced that the assessment rules associated with the PDM and IDM are asymptotically empirical and thus adaptive.

4 Deciding on an optimal strategy

In this section, we investigate how the player can choose a strategy for the next round, that is in some sense optimal for him. First, in Sec. 4.1, we recall some typical ways of (partially) deciding on which strategy to use. Then, we look at how the belief models for the opponent we introduced previously, the PDM and IDM, can be used together with these ways of deciding (Secs. 4.2 and 4.3). Finally, in Sec. 4.4, we make the link with so-called behavior rules.

4.1 Strategy types

A player's strategy τ^i is said to strictly dominate a strategy σ^i when its corresponding unknown payoff is strictly higher, i.e., when $u_{\tau^i} \geq u_{\sigma^i}$ pointwise and $u_{\tau^i} \neq u_{\sigma^i}$. A strategy σ^i is called *inadmissible* if there is another strategy τ^i that dominates it, otherwise it is *admissible*. The set of admissible strategies is nonempty and can be written as a connected union of convex subsets of Σ^i spanned by pure strategies. This property, and the other properties we mention in this section can be seen to hold by using ideas from Walley's book [12, Sec. 3.9] and looking at the set $\{u_{\sigma^i} : \sigma^i \in \Sigma^i\}$, which forms a convex polytope in $\mathbb{R}^{N^{-i}}$.

A rational player can use admissibility as a criterion to limit the number of strategies among which he has to decide. As can be seen from the definition, admissibilit in no way depends on the opponent's strategy choice. This implies that admissibility cannot be directly incorporated into decisions based on the PDM or the IDM, which are assessment models for this unknown strategy. However, this need not exclude our using it separately.

Now suppose that the player knows his opponent is going to play a strategy σ^{-i} in Σ^{-i}. Then, a *best reply* to σ^{-i} is a strategy for which the corresponding unknown payoff in σ^{-i} is maximal. All admissible strategies are best replies to some σ^{-i} and there are

admissible best replies for all σ^{-i}, but not all best replies must be admissible. The set of all best replies to σ^{-i} is denoted by $BR^i(\sigma^{-i})$, so

$$BR^i(\sigma^{-i}) = \mathrm{argmax}_{\sigma^i \in \Sigma^i} u_{\sigma^i}(\sigma^{-i}). \tag{8}$$

It is interesting to note that $BR^i(\sigma^{-i})$ is a convex subset of Σ^i, spanned by a subset of the player's pure strategies, so there exist only a finite number of different sets of best replies. The collection of best replies to the strategies in $\mathscr{S}^{-i} \subseteq \Sigma^{-i}$ is given by $BR^i(\mathscr{S}^{-i}) := \bigcup_{\sigma^{-i} \in \mathscr{S}^{-i}} BR^i(\sigma^{-i})$, which in general is not a convex subset of Σ^i anymore. An illustration of the best reply map is given in Fig. 2.
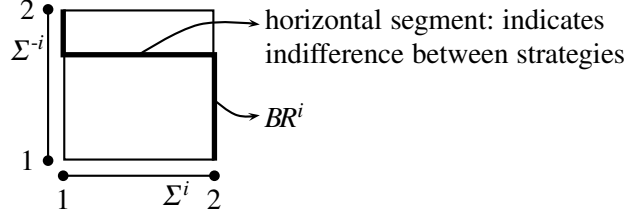

Fig. 2. An illustration of the best reply map for players with two pure strategies. The bold line gives the graph of BR^i.

In sharp contrast to admissibility, it is clear that a player using best replies needs a model of his opponent. Even though a player using a PDM or IDM does not know what his opponent's strategy is going to be, the same idea, maximizing expected payoff, is applicable, as will be shown in the next subsections.

Up to now, we have made no assumptions involving the opponent's payoff u^{-i}. But when the player knows (or has suspicions that make him act as if he knows) he is playing a strictly competitive game (e.g., zero-sum games), his opponent supposedly chooses a strategy that minimizes the player's own payoff. Then it could be rational to play the *maximin strategy* that maximizes this minimal payoff. There always exist (admissible) maximin strategies. When the opponent is known to restrict himself to $\mathscr{S}^{-i} \subseteq \Sigma^{-i}$, the set of all maximin strategies, denoted $MM^i(\mathscr{S}^{-i})$, is

$$MM^i(\mathscr{S}^{-i}) = \mathrm{argmax}_{\sigma^i \in \Sigma^i} \inf_{\sigma^{-i} \in \mathscr{S}^{-i}} u_{\sigma^i}(\sigma^{-i}). \tag{9}$$

Note that $MM^i(\mathscr{S}^{-i}) \subseteq BR^i(\mathscr{S}^{-i})$ and $MM^i(\sigma^{-i}) = BR^i(\sigma^{-i})$, so a maximin strategy is a special type of best reply.

As with best replies, the idea behind maximin strategies can be used in conjunction with an assessment model such as the IDM, but not for the PDM (as we shall see in Sec. 4.3).

In classical fictitious play, a player uses a best reply to v_t^{-i}, the frequency of observed strategies; i.e., he uses an element of $BR^i(v_t^{-i})$. When using models such as the PDM and IDM, we have a different form of information: previsions $P(\cdot|r, \rho^{-i})$ and $\underline{P}(\cdot|r, \mathscr{R}^{-i})$. So we do not know an opponent's strategy or his possible set of strategies.

And although we do know his (set of) expected strategies, we are – without further motivation – not justified in considering best replies to such expected strategies. However, we shall see further on that, because linear previsions and payoff functions are linear, we can nevertheless treat expected strategies as if they were actual strategies, and consider best replies to them as optimal.

4.2 *Optimal strategies under a* PDM

We now use the ideas behind the definition of a best reply in a setting where a probability distribution over all the opponent's strategies is given, instead of a single strategy. We are going to look for a strategy for which the prevision of the corresponding gamble is maximal. This (maximizing expected utility) is the usual approach in Bayesian decision making [4,1]. So we are looking for $\text{argmax}_{\sigma^i \in \Sigma^i} P(u_{\sigma^i}|r, \rho^{-i})$.

Using Eqs. (2), (1) and (3), the prevision of the unknown payoff u_{σ^i} becomes

$$P(u_{\sigma^i}|r,\rho^{-i}) = \int_{\text{int}(\Sigma^{-i})} u_{\sigma^i}(\sigma^{-i}) D(\sigma^{-i}|r,\rho^{-i}) \mathrm{d}\sigma^{-i} = \sum_{s^{-i} \in S^{-i}} u_{\sigma^i}(s^{-i}) P(\sigma^{-i}(s^{-i})|r,\rho^{-i})$$
$$= \sum_{s \in S} u^i(s) \sigma^i(s^i) \rho^{-i}(s^{-i}) = u_{\sigma^i}(\rho^{-i}). \tag{10}$$

Equation (8) then shows that $\text{argmax}_{\sigma^i \in \Sigma^i} P(u_{\sigma^i}|r,\rho^{-i}) = \text{argmax}_{\sigma^i \in \Sigma^i} u_{\sigma^i}(\rho^{-i}) = BR^i(\rho^{-i})$, a result that is based essentially on the bilinear character of the payoff u^i and on the linearity of the linear prevision $P(\cdot|r,\rho^{-i})$. It justifies the approach that is usual with fictitious play [7], where the PDM is only used to get the expected strategy ρ^{-i} and $BR^i(\rho^{-i})$ is the (nonempty) set of optimal strategies.

Another line of reasoning, which will prove useful in Sec. 4.3, can be used to get $\text{argmax}_{\sigma^i \in \Sigma^i} P(u_{\sigma^i}|r,\rho^{-i})$ as the set of optimal strategies. When looking for a best reply to a strategy σ^{-i}, the player's own strategies σ^i are ordered according to their expected payoff $u_{\sigma^i}(\sigma^{-i})$. This means that the relative order of two strategies σ^i and τ^i is determined by looking at the difference in payoff $u_{\sigma^i}(\sigma^{-i}) - u_{\tau^i}(\sigma^{-i})$. The generated order is linear: σ^i can either be better, worse, or as good as τ^i.

With a PDM, the order is determined by the previsions of the payoff differences $P(u_{\sigma^i} - u_{\tau^i}|r,\rho^{-i})$. Because $P(\cdot|r,\rho^{-i})$ is linear, this order is also linear. So then an optimal strategy σ^i is a maximal element of the order. This corresponds to the nonnegativity of $P(u_{\sigma^i} - u_{\tau^i}|r,\rho^{-i})$ for all τ^i, which is equivalent to the criterion of optimality given above,

$$\min_{\tau^i \in \Sigma^i} P(u_{\sigma^i} - u_{\tau^i}|r,\rho^{-i}) \geq 0 \tag{11}$$
$$\Leftrightarrow P(u_{\sigma^i}|r,\rho^{-i}) \geq \max_{\tau^i \in \Sigma^i} P(u_{\tau^i}|r,\rho^{-i})$$
$$\Leftrightarrow \sigma^i \in \text{argmax}_{\tau^i \in \Sigma^i} P(u_{\tau^i}|r,\rho^{-i}) = BR^i(\rho^{-i}).$$

So when using a PDM as an assessment model for his opponent, the player makes

11

a choice ρ_0^{-i} for his initial model parameter. This implies that he initially plays a strategy in $BR^i(\rho_0^{-i})$, a set which can be strongly dependent on ρ_0^{-i}. Now suppose, for the sake of the argument, that after a large number of rounds t, it so happens that $\rho_t^{-i} = \rho_0^{-i}$. This means that the player's set of optimal strategies after observing the opponent's response during t rounds, is the the same as when he sets out to play. So this type of assessment model does not allow us to distinguish between decisions based on different numbers of observations. The behavioral implications of the model (which strategy to play next) are always equally strong.

4.3 Optimal strategies under an IDM

When using an IDM, we cannot just maximize the prevision of an unknown payoff because we are working with a set of linear previsions $\mathscr{P}(r, \mathscr{R}^{-i})$, or equivalently, with lower and upper previsions $\underline{P}(\cdot|r, \mathscr{R}^{-i})$ and $\overline{P}(\cdot|r, \mathscr{R}^{-i})$.

What we do is use the IDM to generate an order of the player's strategies. Compared to τ^i, a strategy σ^i is

- equally good when both $\underline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) = 0$ and $\overline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) = 0$,
- strictly better when $\underline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) > 0$, or, equivalently,
- strictly worse when $\overline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) < 0$, and
- incomparable when none of the above hold.

The 'strictly better'-relationship is now only a strict *partial* order, caused by the fact that $\underline{P}(\cdot|r, \mathscr{R}^{-i})$ is not linear. We are mainly interested in the maximal elements of this order, i.e., those that are undominated under the 'strictly better'-relationship. A maximal element is a strategy σ^i that is not strictly worse than any other strategy, so for which $\min_{\tau^i \in \Sigma^i} \overline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) \geq 0$. This criterion is a clear generalization of Eq. (11), the criterion found when using a PDM. (Other criteria are of course possible, our criterion is equivalent to *maximality* [12, Sec. 3.9] and, because of the convexity of the set of gambles, to *E-admissibility* [10].)

By rewriting the criterion, we get a more explicit description of the optimal strategies. Using the definition of an upper prevision, linearity, and Eq. (10), we find

$$\min_{\tau^i \in \Sigma^i} \overline{P}(u_{\sigma^i} - u_{\tau^i}|r, \mathscr{R}^{-i}) \geq 0 \qquad (12)$$
$$\Leftrightarrow \min_{\tau^i \in \Sigma^i} \max_{P \in \mathscr{P}(r, \mathscr{R}^{-i})} P(u_{\sigma^i} - u_{\tau^i}) \geq 0$$
$$\Leftrightarrow \min_{\tau^i \in \Sigma^i} \sup_{\rho^{-i} \in \mathscr{R}^{-i}} \left[u_{\sigma^i}(\rho^{-i}) - u_{\tau^i}(\rho^{-i}) \right] \geq 0.$$

Considering that Σ^i and $\mathrm{cl}(\mathscr{R}^{-i})$ are compact, convex sets and that unknown payoffs are bilinear functions, the maximin theorem [12, Sec. E6] applies, so the criterion becomes $\max_{\rho^{-i} \in \mathrm{cl}(\mathscr{R}^{-i})} \left[u_{\sigma^i}(\rho^{-i}) - \max_{\tau^i \in \Sigma^i} u_{\tau^i}(\rho^{-i}) \right] \geq 0$. It can, and will, for any ρ^{-i} in $\mathrm{cl}(\mathscr{R}^{-i})$, only be satisfied when $\sigma^i = BR^i(\rho^{-i})$, which means $BR^i(\mathrm{cl}(\mathscr{R}^{-i}))$ is the

(nonempty) set of optimal strategies. All strategies in this set are valid optimal choices, but when compared amongst themselves, they are either equivalent or incomparable. So within this set no strategy is strictly better than any other.

Whenever the player knows he is playing a strictly competitive game, it is rational for him to try and limit his losses. When we defined maximin strategies, the player knew what set $\mathscr{S}^{-i}$ his opponent was choosing his strategy from. Now his beliefs are contained in an IDM $\underline{P}(\cdot|r,\mathscr{R}^{-i})$ and to limit his losses, he tries to maximize the lower prevision of his unknown payoff. (This would make no sense when using a PDM, as lower and upper previsions coincide.) This means an optimal strategy σ^i is defined by $\sigma^i \in \mathrm{argmax}_{\tau^i \in \Sigma^i} \underline{P}(u_{\tau^i}|r,\mathscr{R}^{-i})$. Rewriting this criterion gives a more explicit description of the optimal strategies for this case. The definition of a lower prevision gives $\mathrm{argmax}_{\sigma^i \in \Sigma^i} \underline{P}(u_{\sigma^i}|r,\mathscr{R}^{-i}) = \mathrm{argmax}_{\sigma^i \in \Sigma^i} \min_{P \in \mathscr{P}(r,\mathscr{R}^{-i})} P(u_{\sigma^i})$, which is equal to $\mathrm{argmax}_{\sigma^i \in \Sigma^i} \min_{\rho^{-i} \in \mathrm{cl}(\mathscr{R}^{-i})} u_{\sigma^i}(\rho^{-i})$ by Eq. (10). Using Eq. (9), the definition of a maximin strategy, we see that the optimal strategies now correspond to the (nonempty) set $MM^i(\mathrm{cl}(\mathscr{R}^{-i}))$,[4] which is a subset of $BR^i(\mathrm{cl}(\mathscr{R}^{-i}))$, the optimal strategies when no assumptions about the opponent's payoff are made.

So now, when using an IDM as an assessment model for his opponent, the player chooses $\mathscr{R}_0^{-i} = \mathrm{int}(\Sigma^{-i})$ as his initial model parameter. This implies that he initially plays a strategy in $BR^i(\mathrm{cl}(\mathscr{R}_0^{-i})) = BR^i(\Sigma^{-i})$ or a strategy in $MM^i(\Sigma^{-i})$, which is a classical maximin strategy [1, Ch. 5]. This corresponds to the weakest possible rational behavior. After a number of rounds t, the player uses a strategy in $BR^i(\mathrm{cl}(\mathscr{R}_t^{-i}))$ or in $MM^i(\mathrm{cl}(\mathscr{R}_t^{-i}))$. Equation (5) shows that $\mathrm{cl}(\mathscr{R}_t^{-i})$ gets smaller as the number of observations increases, and so the behavioral implications also get stronger. Using an IDM, it is thus possible to distinguish decisions based on different amounts of observational data, in contrast to the situation when using a PDM. As the number of observations gets very large, the behavioral implications of an IDM often tend to be the same as those of a PDM, which will be illustrated when discussing absorption to strict equilibria in Sec. 5.2. That this is not a general rule will be illustrated when discussing convergence to mixed equilibria in Sec. 5.3.

4.4 Behavior rules

The player's behavior during round t is the way he chooses a strategy. For a rational player, this behavior is based on the assessments about his opponent's strategy and on his own and possibly his opponent's payoff.

In Sec. 4.2 we have seen that a rational player using a PDM $P(\cdot|r_t,\rho_t^{-i})$ as his assessment of the opponent's mixed strategy (assumed fixed), must choose any

---

[4] This illustrates that the choice of decision criterion can be separated from the choice of (imprecise) prior model, casting a new light on an old discussion between subjective Bayesians [9] and game theorists [8].

strategy in $BR^i(\rho_t^{-i})$ in order to make an optimal decision, irrespective of the payoff u^{-i} of his opponent. In Sec. 4.3, we have shown that similarly a rational player using an IDM $\underline{P}(\cdot|r_t, \mathscr{R}_t^{-i})$ as his assessment of the opponent's mixed strategy, must choose any strategy in $BR^i(\text{cl}(\mathscr{R}_t^{-i}))$, to make an optimal decision. In the latter case, however, knowledge about the opponent's payoff u^{-i} can lead him to further refine his choice, e.g., to $MM^i(\text{cl}(\mathscr{R}_t^{-i})) \subseteq BR^i(\text{cl}(\mathscr{R}_t^{-i}))$ when playing a strictly competitive game.

Together with assessment rules (see Sec. 3.4), Fudenberg and Kreps [5] introduce the concept of a *behavior rule* ϕ^i, which determines the strategy the player will use in the next round $(t+1)$, based on the observed history h_t and some initial beliefs. When $q_t \in \mathscr{Q}_t$ are the parameters of the player's belief model, a behavior rule is defined as a map $\phi^i: \mathscr{Q}_t \to \Sigma^i: q_t \mapsto \phi^i(q_t)$. Whenever the parameters are implicit, we use $\phi_t^i = \phi^i(q_t)$. We also use profiles of behavior rules $\phi = (\phi^i, \phi^{-i})$. So, if ϕ_{t-1} is equal to the profile s, then s will be played in round t.

It is possible to write the behavior rules for the models we have discussed as a function of the assessment rules for these models. (In general the assessment rules contain less information than the assessments (a PDM or an IDM). This is why we derived the optimal strategies directly from these assessments.) For the PDM, $q_t = (r_t, \rho_t^{-i})$, so
$$\phi^i(r_t, \rho_t^{-i}) \in BR^i(\mu^{-i}(r_t, \rho_t^{-i})) = BR^i(\rho_t^{-i}).$$
For the IDM, $q_t = (r_t, \mathscr{R}_t^{-i})$, so

$\phi^i(r_t, \mathscr{R}_t^{-i}) \in BR^i(\mu^{-i}(r_t, \mathscr{R}_t^{-i})) = BR^i(\text{cl}(\mathscr{R}_t^{-i})),$
$\phi^i(r_t, \mathscr{R}_t^{-i}) \in MM^i(\mu^{-i}(r_t, \mathscr{R}_t^{-i})) = MM^i(\text{cl}(\mathscr{R}_t^{-i}))$ (for strictly competitive games).

In many cases the above equations do not define a unique behavior rule, more often so when using an IDM than when using a PDM, but they say no more than this. To get a unique rule, one ideally uses other justifiable criteria (such as admissibility), but otherwise has to resort to an arbitrary choice.

Behavior rules determine which histories are possible. A history is called *compatible* with the behavior rules ϕ of the players, if it can be generated (with non-zero chance) by these behavior rules. Explicitly, this means that for every pure strategy profile $h_t(t')$, with $1 \leq t' \leq t$, that is a component of a compatible history, the chance $\phi_{t'-1}(h_t(t'))$ is strictly positive. This implies that the randomization devices used by the players can select the pure strategies in $h_t(t')$ with non-zero chance. (The uncertainty we talk about in this paragraph is due to the randomization mechanisms used by the players, which is why we use the word chance.)

To consider extensions of fictitious play, Fudenberg and Kreps [5] defined some classes of behavior rules. We generalize some of these to allow for set-valued assessment rules. A profile of behavior rules belongs to a certain class if both its components belong to it.

A behavior rule ϕ^i is called *myopic* relative to an assessment rule μ^{-i} if it maximizes

the player's immediate expected payoff, i.e., when $\phi_t^i \in BR^i(\mu_t^{-i})$ for all t and h_t. It is clear that the behavior rules we have defined above for the models using the PDM and the IDM are myopic relative to the assessment rules we have defined in Sec. 3.4.

Let us look at a larger class of behavior rules, i.e., those for which immediate expected payoff only has to be maximized asymptotically. We call a behavior rule ϕ^i *strongly asymptotically myopic* relative to the assessment rule μ^{-i} if, for some sequence $\varepsilon_t > 0$ with $\lim_{t\to\infty} \varepsilon_t = 0$ and for all t and histories h_t^{-i}, it holds that

$$\forall \sigma^{-i} \in \mu_t^{-i} \colon \forall \tilde{s}^i \in S^i \text{ such that } \phi_t^i(\tilde{s}^i) > 0 \colon \ u_{\tilde{s}^i}(\sigma^{-i}) + \varepsilon_t \geq \max_{s^i \in S^i} u_{s^i}(\sigma^{-i}). \quad (13)$$

This concept allows us to formulate results for a larger class of models than only the ones using the PDM or the IDM. It is used in Theorems 3 and 4 of the next section.

5 Convergence results

In this section we give some results about the convergence of play to equilibria. These results are formulated using the assessment rules and behavior rules introduced in Secs. 3.4 and 4.4. They are similar to the ones presented by Fudenberg and Kreps [5], but allow for set-valued assessment rules μ. Because randomization devices are used – to be able to play mixed strategies – these results hold with chance 1; this fact is not explicitly mentioned further on. The models based on the PDM or the IDM are used as examples. After defining equilibria, we look at convergence to pure equilibria in Sec. 5.2 and at convergence to mixed equilibria in Sec. 5.3.

5.1 Equilibria

An *equilibrium* is a strategy profile σ_* for which the payoff for both players cannot be increased if one of them changes his strategy, while his opponent's strategy remains unchanged. It is well known that this corresponds to the strategy profile being a fixed point of the combined best reply mapping (defined as $BR(\sigma) = BR^i(\sigma^{-i}) \times BR^{-i}(\sigma^i) \subseteq \Sigma$). So σ_* is an equilibrium if and only if $\sigma_* \in BR(\sigma_*)$.

A *strict equilibrium* is a profile s_* of pure strategies that is its own unique best reply, i.e., for which $s_* = BR(s_*)$. A non-strict equilibrium is called a *mixed equilibrium*. The concepts above are illustrated in Fig. 3, where we specify some games by giving the best reply graph for both players (this is equivalent to giving their payoff functions).
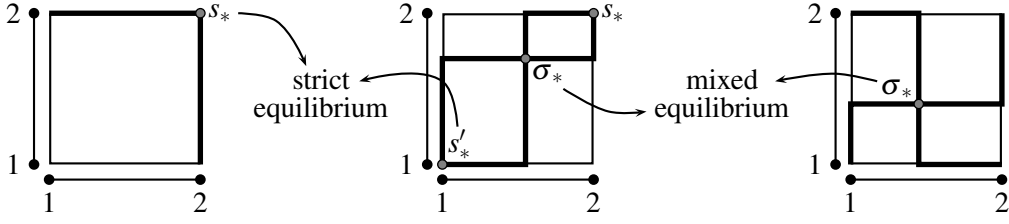
15

Fig. 3. Examples of equilibria for players with two pure strategies. On the left, there is one strict equilibrium; in the middle, two strict and one mixed; on the right, one mixed. Equilibria correspond to intersections of the graphs of BR^i and BR^{-i} (see Fig. 2).

5.2 Convergence to strict equilibria

First, let us look at when convergence to a strict equilibrium is guaranteed to occur. One situation was described by Fudenberg and Kreps [5, Proposition 3.0]:

Theorem 1 (absorption to a strict equilibrium) *If there is a strict equilibrium s_* that is played in round t of a history h_t compatible with myopic behavior rules ϕ relative to the assessment rule $\mu = (\rho^i, \rho^{-i})$ of players using a PDM, then s_* will be played during all subsequent rounds $t' > t$.*

In Fig. 4, we give an illustration of a situation where absorption occurs. In this situation, both players use a PDM with $r_0 = 2$ and the best reply as a behavior rule. The prior assessments are described by μ_0. The table lists all the data necessary to determine the assessment rules μ_t for t up to 5 and the subsequent strategy profiles $\phi_t = s_{t+1} = (s^i_{t+1}, s^{-i}_{t+1})$ played; ties were broken arbitrarily. From round 5 onward, only the strict equilibrium s_* is played. In the picture, we show the evolution of μ_t with increasing round number t.

t	$\rho^i_t(2), \rho^{-i}_t(2)$	ϕ_t
0	$0, \frac{1}{2}$	$(2,1)$
1	$\frac{1}{3}, \frac{1}{3}$	$(1,2)$
2	$\frac{1}{4}, \frac{1}{2}$	$(2,1)^\star$
3	$\frac{3}{5}, \frac{2}{5}$	$(1,2)^\star$
4	$\frac{1}{3}, \frac{1}{2}$	s_*
5	$\frac{3}{7}, \frac{4}{7}$	s_*

$$\mu_t = (\rho^i_t, \rho^{-i}_t)$$
$$= \frac{t}{r_0+t}\nu_t + \frac{r_0}{r_0+t}\rho_0$$
$$\rho^i_t(1) = 1 - \rho^i_t(2)$$
$$\rho^{-i}_t(1) = 1 - \rho^{-i}_t(2)$$
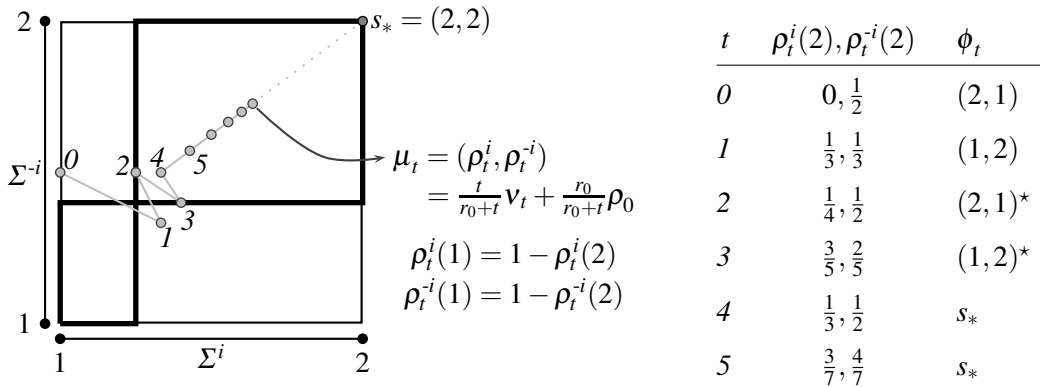
Fig. 4. An illustration of absorption when both players use a PDM with $r_0 = 2$ and respectively $\rho^{-i}_0 = (\frac{1}{2}, \frac{1}{2})$ and $\rho^i_0 = (1, 0)$. Recall that a player's assessments μ^{-i}_t about his opponent's strategy are determined by the assessment model's parameters; here, these are the number of (real and imaginary) rounds r_t and the expected strategy ρ^{-i}_t. The latter evolves with the frequency of observed strategies ν^{-i}_t, which reflects the opponent's behavior ϕ^{-i}_t. Round numbers t are given in italics. Of all the moves ϕ_t given in the table, those marked with '$\star$' could have been different. For the meaning of the bold lines, see Figs. 2 and 3.

We want to stress the fact that absorption does not necessarily occur when the assessment rule μ is not singleton-valued. This is illustrated in Fig. 5. Here, both players use an IDM with $r_0 = 2$ and the best reply as a behavior rule. The prior assessments are described by $\mu_0 = \Sigma$. The table lists the information necessary to determine the assessment rule μ_t for the following rounds and the subsequent strategy profiles ϕ_t played (again, an arbitrary choice was made between the different possible strategy profiles). In the picture, we again show the evolution of μ_t with increasing round number t. Even though $s_3 = \phi_2 = s_*$, absorption to this strict equilibrium does not occur as $\phi_3 \neq s_*$.

$$\mu_t = \mathrm{cl}(\mathscr{R}_t^i) \times \mathrm{cl}(\mathscr{R}_t^{-i})$$
$$= \frac{t}{r_0+t} v_t + \frac{r_0}{r_0+t} \Sigma$$
$$v_t^i(1) = 1 - v_t^i(2)$$
$$v_t^{-i}(1) = 1 - v_t^{-i}(2)$$

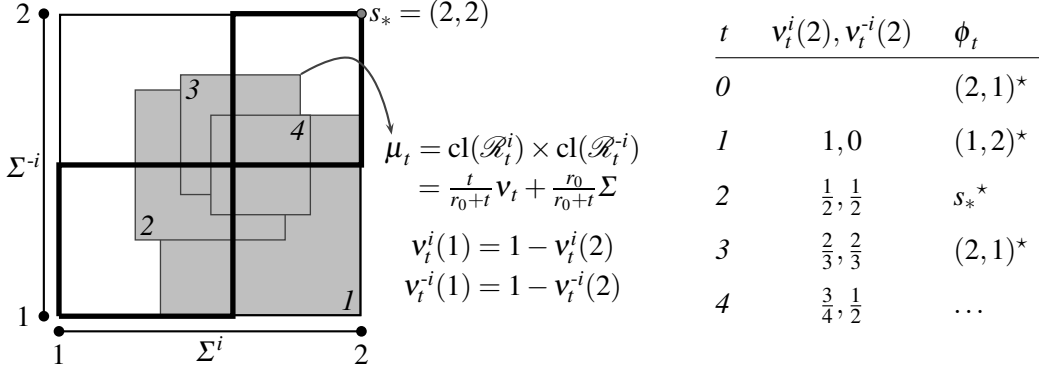| t	$v_t^i(2), v_t^{-i}(2)$	ϕ_t
0		$(2,1)^\star$
1	$1,0$	$(1,2)^\star$
2	$\frac{1}{2}, \frac{1}{2}$	$s_*{}^\star$
3	$\frac{2}{3}, \frac{2}{3}$	$(2,1)^\star$
4	$\frac{3}{4}, \frac{1}{2}$	$\cdots$

Fig. 5. An illustration of non-absorption when both players use an IDM with $r_0 = 2$. See Fig. 4 for an explanation about the notation.

For set-valued μ, the result *can* be obtained by strengthening the requirements:

Theorem 2 (conditional absorption to a strict equilibrium) *If, for some history h_t compatible with myopic behavior rules ϕ relative to the assessment rule $\mu = \mathrm{cl}(\mathscr{R}^i) \times \mathrm{cl}(\mathscr{R}^{-i})$ of players using an IDM, the strategy profile ϕ_t cannot differ from the strict equilibrium s_*, then s_* will be played during all subsequent rounds $t' > t$.*

Look at the condition "[if] ϕ_t cannot differ from the strict equilibrium s_*": due to myopia $\phi_t \in BR(\mu_t)$, so this condition is satisfied if and only if $BR(\mu_t) = s_*$.

Theorem 2 is illustrated in Fig. 6. It shows a possible continuation of the situation in Fig. 5; from round 10 onwards, absorption has occurred. This illustrates that after a sufficient number of rounds, the behavior of a player using an IDM often tends to be the same as the behavior of a player using a PDM. The main reason is that, as the area of μ_t becomes smaller, it behaves more and more like a point, i.e., intersections of μ_t with the best reply mappings BR^i and BR^{-i} become rare.

We can also investigate when convergence to a strict equilibrium is guaranteed to have occurred. This leads to a generalization of a result of Fudenberg and Kreps [5, Proposition 4.1] to set-valued assessment rules.

Theorem 3 (repeated play of a pure strategy profile) *Consider an infinite history h_∞ in $\mathscr{H}_\infty$ such that for some t_0, a pure strategy profile s_* is played in all*

17

$$\mu_t = \frac{t}{r_0+t}\nu_t + \frac{r_0}{r_0+t}\Sigma$$

$$\nu_t^i(1) = 1 - \nu_t^i(2)$$
$$\nu_t^{-i}(1) = 1 - \nu_t^{-i}(2)$$

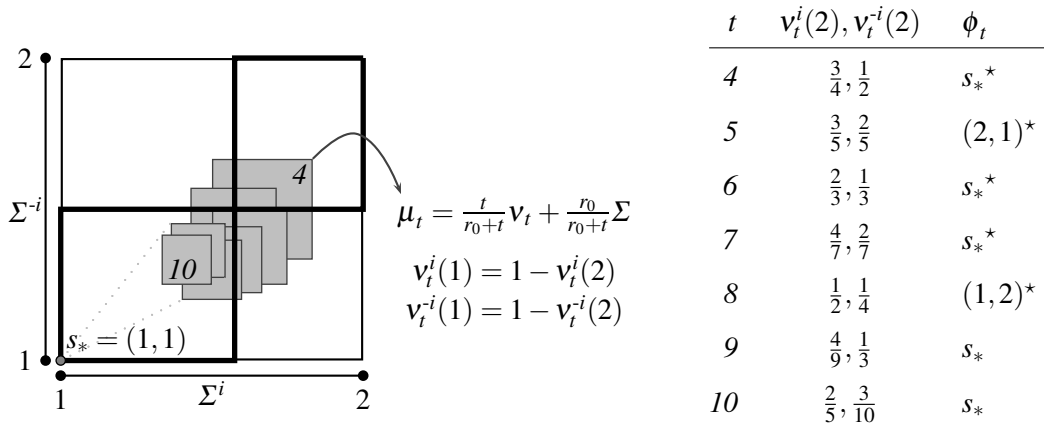| t	$\nu_t^i(2), \nu_t^{-i}(2)$	ϕ_t
4	$\frac{3}{4}, \frac{1}{2}$	$s_*^\star$
5	$\frac{3}{5}, \frac{2}{5}$	$(2,1)^\star$
6	$\frac{2}{3}, \frac{1}{3}$	$s_*^\star$
7	$\frac{4}{7}, \frac{2}{7}$	$s_*^\star$
8	$\frac{1}{2}, \frac{1}{4}$	$(1,2)^\star$
9	$\frac{4}{9}, \frac{1}{3}$	s_*
10	$\frac{2}{5}, \frac{3}{10}$	s_*

Fig. 6. An illustration of conditional absorption when both players use an IDM with $r_0 = 2$. See Fig. 4 for an explanation about the notation.

subsequent rounds. If h_∞ is compatible with behavior rules ϕ that are strongly asymptotically myopic relative to the adaptive assessment rules μ, then s_* is an equilibrium.

The conditions on the assessment rules of this theorem are satisfied when the players use an IDM or a PDM. Note, however, that this result is also valid for other models satisfying the conditions.

5.3 Convergence to mixed equilibria

Because absorption cannot be generalized straightforwardly to mixed equilibria, we immediately look at when convergence to a mixed equilibrium is guaranteed to have occurred. This leads to the following generalization to set-valued assessment rules of another result of Fudenberg and Kreps [5, Proposition 4.2].

Theorem 4 (repeated play of a mixed strategy profile) *Let the infinite history h_∞ in $\mathcal{H}_\infty$ be such that for some mixed strategy profile σ_*, $\lim_{t\to\infty} \nu(h_t, \cdot) = \sigma_*$ holds, where h_t is a partial history of h_∞ for every t. If the infinite history h_∞ is compatible with behavior rules ϕ that are strongly asymptotically myopic relative to the assessment rules μ that are asymptotically empirical, then σ_* is an equilibrium.*

The conditions on the assessment rules of this theorem are satisfied when the players use an IDM or a PDM. Again, this result is not restricted to these models.

Even though the focus in this paper is mostly on the use of the IDM and not on the interpretation of 'learning to play a mixed strategy', we would like to finish this section by showing that learning using an IDM (or other models with set-valued assessment rules) can remedy some pathological behavior of the PDM (or other models with singleton-valued assessment rules). Fudenberg and Kreps [5, Sec. 5] argue that when using a PDM, most of the time no mixed strategy is played, but that the players jump between pure strategies in cycles.

18

This is illustrated with the *battle of the sexes* [3], where for some initial assessments a suboptimal strategy profile is constantly played, even though convergence to the mixed equilibrium occurs.[5] This is shown in Fig. 7, where both players use a PDM with $r_0 = \frac{1}{2}$ and the best reply as a behavior rule. The prior assessments are described by μ_0 (the value of its components is given at the top of the table). We see that only the suboptimal strategy profiles $(1,1)$ and $(2,2)$ are played, even though convergence to $\sigma_* = ((\frac{2}{3}, \frac{1}{3}), (\frac{2}{3}, \frac{1}{3}))$ occurs for both μ_t and ν_t.

t	$\rho_t^i(2), \rho_t^{-i}(2)$	ϕ_t
0	$\frac{3}{5}, \frac{3}{5}$	$(1,1)$
1	$\frac{3}{15}, \frac{3}{15}$	$(2,2)$
2	$\frac{13}{25}, \frac{13}{25}$	$(1,1)$
3	$\frac{13}{35}, \frac{13}{35}$	$(1,1)$
4	$\frac{13}{45}, \frac{13}{45}$	$(2,2)$
5	$\frac{23}{55}, \frac{23}{55}$	$(1,1)$

$\mu_t = (\rho_t^i, \rho_t^{-i})$

$\rho_t^i(1) = 1 - \rho_t^i(2)$
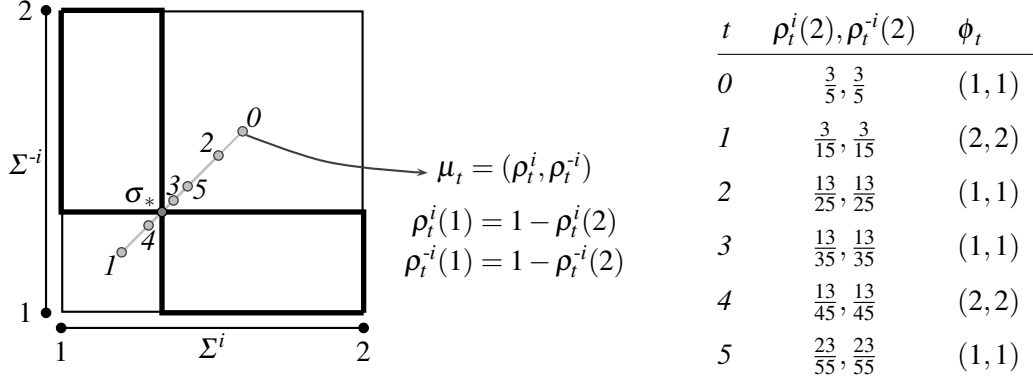$\rho_t^{-i}(1) = 1 - \rho_t^{-i}(2)$

Fig. 7. An illustration of pathological gameplay with the battle of the sexes when both players use a PDM with $r_0 = \frac{1}{2}$ and $\rho_0^{-i} = (\frac{2}{5}, \frac{3}{5}) = \rho_0^i$. No move could have been different. See Fig. 4 for an explanation about the notation.

When the players use an IDM for the same game, the same history can be played, but it is now very likely that the pathological correlated gameplay does not occur. This is illustrated in Fig. 8, where both players use an IDM with $r_0 = \frac{1}{2}$ and the best reply as a behavior rule. The prior assessments are described by $\mu_0 = \Sigma$. Even though
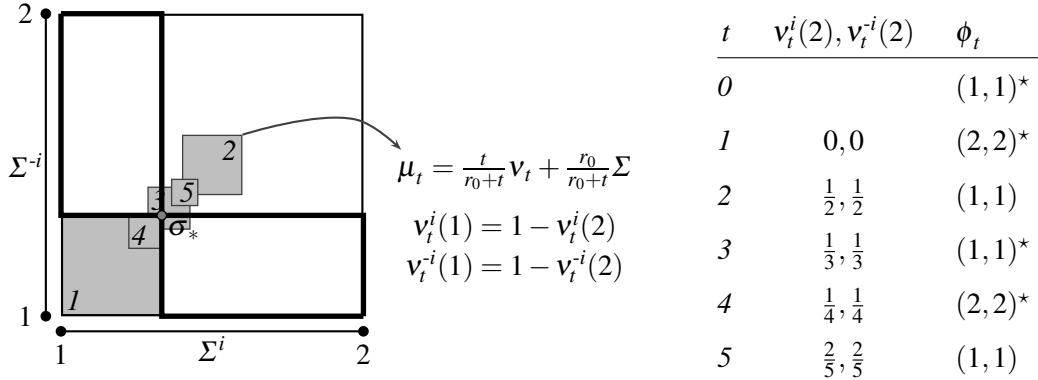

t	$\nu_t^i(2), \nu_t^{-i}(2)$	ϕ_t
0		$(1,1)^\star$
1	$0, 0$	$(2,2)^\star$
2	$\frac{1}{2}, \frac{1}{2}$	$(1,1)$
3	$\frac{1}{3}, \frac{1}{3}$	$(1,1)^\star$
4	$\frac{1}{4}, \frac{1}{4}$	$(2,2)^\star$
5	$\frac{2}{5}, \frac{2}{5}$	$(1,1)$

$\mu_t = \frac{t}{r_0+t}\nu_t + \frac{r_0}{r_0+t}\Sigma$

$\nu_t^i(1) = 1 - \nu_t^i(2)$
$\nu_t^{-i}(1) = 1 - \nu_t^{-i}(2)$

Fig. 8. Gameplay with the battle of the sexes when using set-valued assessment rules when both players use an IDM with $r_0 = \frac{1}{2}$. See Fig. 4 for an explanation about the notation.

we took the partial history shown in this example to be the same as in Fig. 7, it is

---

[5] Fudenberg and Levine [6] propose cautious variants of fictitious play to address this problem and others. They focus on a modification of the players' behavior rules, while we – in this paper – focus mainly on a modification of the players' assessment rules.

clear that when the choice between the different possible ϕ_t is made arbitrarily (or using some targeted method), it is highly likely that the pathological gameplay is interrupted at some point. In the table, we indicate all the strategy profiles that could have been chosen differently with a star. This illustrates what we hinted at at the end of Sec. 4.3, that after a sufficient number of rounds, the behavior of a player using an IDM can sometimes be very different from the behavior of a player using a PDM, thanks to using set-valued assessment rules instead of point-valued ones.

6 Overview of the IDM's advantages and conclusions

So which of the two models for assessment we have discussed, the PDM and the IDM, should be used? Of course this depends on the situation. Both are based on a form of Bayesian inference. Both should reflect the available data and the assumptions made. Neither should depend on any hidden or unjustified assumptions.

Assumptions about the initial strategy choice ρ_0^{-i} of the opponent, necessary for using a PDM, are often unjustified. For example, the assumptions of fictitious play do not say anything about the initial strategy choice. So then we can only say something about the initial strategy choice by making additional assumptions. This is the main reason we have proposed using the IDM, for which this is not necessary (cf. Sec. 3.3). One consequence of initially having less information, is that the behavior resulting from using an IDM (instead of a PDM) is less decisive, i.e., the set of optimal strategies can be larger (cfr. Secs. 4.2 and 4.3). This is something very natural and honest: when taking decisions, one should be less decisive when one has less information.

Another interesting result similarly involves the influence of additional information. Knowing that the game is strictly competitive does not restrict the set of optimal strategies when using a PDM, but may restrict this set when using an IDM (cf. Sec. 4.3).

Considering the similarities and differences between the PDM and the IDM, it is not surprising that this is reflected in the resulting behavior. As we have illustrated, the similarity allows absorption to strict equilibria to remain possible (cf. Sec. 5.2). The difference allows some pathological forms of gameplay to become less likely (cf. Sec. 5.3).

The basis for the differences is the possibility of having assessment rules that are set-valued instead of point-valued (cf. Sec. 3.4). We have investigated the consequences of using such set-valued assessment rules and have shown that existing results about convergence to strict and mixed equilibria occurring can be generalized to this context (cf. Sec. 5). However, the possibility of obtaining behavior different from the classical learning models is probably the most interesting aspect of the model we

propose in this paper. We think that the best way to achieve some goal or generate some specific behavior is by formulating additional optimality criteria that further restrict the set of optimal strategies in a way that encourages achieving the goal or that favors the specific behavior. A speculative example: to avoid absorption to a strict equilibrium, it might be useful to discourage use of the same pure strategy in subsequent rounds.

With regard to extending the results of this paper, it is clear that the approach to learning we have given here for two-player games can be generalized to multiplayer games. There are two immediate options [7]. When the opponents are assumed to play independently, a separate IDM can be used for each of the opponents. When this assumption is not made, one IDM on the set spanned by all the tuples of the opponents' pure strategies can be used. We will not speculate on the generalization of the convergence results to the multiplayer case.

A   Appendix: proofs

Proof of Theorem 2 (includes Theorem 1 as a special case). The fact that ϕ_t cannot be different from s_* means that $BR(\mathscr{R}_t^i \times \mathscr{R}_t^{-i}) = s_*$, or, for both players, that

$$BR^i(\mathscr{R}_t^{-i}) = \bigcup_{\rho_t^{-i} \in \mathscr{R}_t^{-i}} \operatorname{argmax}_{\sigma^i \in \Sigma^i} u_{\sigma^i}(\rho_t^{-i}) = s_*^i. \tag{A.1}$$

So the strategy s_*^i is the unique best reply (payoff maximizer) for all $\rho_t^{-i} \in \mathscr{R}_t^{-i}$. By myopia $\phi_{t+1}^i \in BR^i(\mathscr{R}_{t+1}^{-i})$ and for the round $t+1$ we can write

$$BR^i(\mathscr{R}_{t+1}^{-i}) = \bigcup_{\rho_{t+1}^{-i} \in \mathscr{R}_{t+1}^{-i}} \operatorname{argmax}_{\sigma^i \in \Sigma^i} u_{\sigma^i}(\rho_{t+1}^{-i}).$$

By using the bilinearity of the payoff and

$$\rho_{t+1}^{-i} = \frac{r_0}{r_{t+1}}\rho_0^{-i} + \frac{1}{r_{t+1}}n_{t+1}^{-i} = \frac{r_t}{r_{t+1}}\left[\frac{r_0}{r_t}\rho_0^{-i} + \frac{1}{r_t}n_t^{-i}\right] + \frac{1}{r_{t+1}}s_*^{-i} = \frac{r_t}{r_{t+1}}\rho_t^{-i} + \frac{1}{r_{t+1}}s_*^{-i},$$

this can be rewritten as

$$BR^i(\mathscr{R}_{t+1}^{-i}) = \bigcup_{\rho_t^{-i} \in \mathscr{R}_t^{-i}} \operatorname{argmax}_{\sigma^i \in \Sigma^i}\left[\frac{r_t}{r_{t+1}}u_{\sigma^i}(\rho_t^{-i}) + \frac{1}{r_{t+1}}u_{\sigma^i}(s_*^{-i})\right].$$

Because of Eq. (A.1), s_*^i is the unique maximizer of the first term. Because s_* is a strict equilibrium, this is also the case for the second term. This implies that s_*^i is also the unique maximizer of their sum, or that $BR^i(\mathscr{R}_{t+1}^{-i}) = s_*^i$. Thus, by myopia, $\phi_{t+1} = s_*$. Induction completes the proof. $\square$

Lemma 5 (used to prove Theorems 3 and 4) *Consider an infinite history h_∞ in $\mathscr{H}_\infty$, adaptive assessment rules μ and a strategy profile σ_* such that for each player i there is a sequence of strictly positive reals δ_t (indexed by the round*

number t) converging to 0 *for which it holds for all t that*

$$\forall \sigma^{-i} \in \mu_t^{-i}: \forall s^{-i} \in S^{-i}: \exists \lambda_{s^{-i},t}^{\sigma^{-i}} \in \mathbb{R}:$$

$$|\lambda_{s^{-i},t}^{\sigma^{-i}}| < \delta_t \quad and \quad \sigma^{-i} = [1 - \textstyle\sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}}]\sigma_*^{-i} + \textstyle\sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}} s^{-i}. \quad \text{(A.2)}$$

If h_∞ is compatible with behavior rules ϕ that are strongly asymptotically myopic relative to the assessment rules μ, then σ_* is an equilibrium.

Proof of Lemma 5. We give a proof by contradiction and suppose *ex absurdo* that σ_* is not an equilibrium. We use $\lambda_{\sigma_*^{-i},t}^{\sigma^{-i}}$ as a shorthand for $1 - \sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}}$. Because the behavior rules ϕ are strongly asymptotically myopic (Eq. (13)) we can write for each player i, for all t and every $\sigma^{-i} \in \mu_t^{-i}$ that for all pure strategies $\tilde{s}^i$ for which $\phi_t^i(\tilde{s}^i) > 0$, it holds that $u_{\tilde{s}^i}(\sigma^{-i}) + \varepsilon_t \geq \max_{s^i \in S^i} u_{s^i}(\sigma^{-i})$, where $\varepsilon_t > 0$ is some sequence converging to 0. Or, because of Eq. (A.2) and the bilinearity of the payoff, again we can write for each player i, for all t and every $\sigma^{-i} \in \mu_t^{-i}$ that for all pure strategies $\tilde{s}^i$ for which $\phi_t^i(\tilde{s}^i) > 0$, it holds that

$$\lambda_{\sigma_*^{-i},t}^{\sigma^{-i}} u_{\tilde{s}^i}(\sigma_*^{-i}) + \textstyle\sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}} u_{\tilde{s}^i}(s^{-i}) + \varepsilon_t \geq$$
$$\max_{s^i \in S^i}[\lambda_{\sigma_*^{-i},t}^{\sigma^{-i}} u_{s^i}(\sigma_*^{-i}) + \textstyle\sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}} u_{s^i}(s^{-i})]. \quad \text{(A.3)}$$

Because σ_* is not an equilibrium, it holds for at least one player – called j – that $\sigma_*^j \notin BR^j(\sigma_*^{-j})$. Because – as mentioned in Sec. 4.1 – every $\sigma^j \in BR^j(\sigma_*^{-j})$ is a convex combination of the $s^j \in BR^j(\sigma_*^{-j})$, there is a strategy $\hat{s}^j \notin BR^j(\sigma_*^{-j})$ such that $\sigma_*^j(\hat{s}^j) > 0$. So for this strategy $\hat{s}^j$,

$$\max_{s^j \in S^j} u_{s^j}(\sigma_*^{-j}) - u_{\hat{s}^j}(\sigma_*^{-j}) = \gamma > 0. \quad \text{(A.4)}$$

We now show that this implies that there is some t^* such that $\phi_t^j(\hat{s}^j) = 0$ for all $t > t^*$. *Ex absurdo*, assume that this does not hold, then for all t^* there is some $t' > t^*$ such that $\phi_{t'}^j(\hat{s}^j) > 0$, or in other words, there is some subsequence $\phi_{t'}^j$ of ϕ_t^j such that $\phi_{t'}^j(\hat{s}^j) > 0$ for all t'. Eq. (A.3) holds in particular for $i = j$, $t = t'$ and $\tilde{s}^j = \hat{s}^j$, so by substituting Eq. (A.4) in Eq. (A.3), we find that for all t'

$$\lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} \max_{s^j \in S^j} u_{s^j}(\sigma_*^{-j}) - \lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} \gamma + \textstyle\sum_{s^{-j} \in S^{-j}} \lambda_{s^{-j},t'}^{\sigma^{-j}} u_{\hat{s}^j}(s^{-j}) + \varepsilon_{t'} \geq$$
$$\max_{s^j \in S^j}[\lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} u_{s^j}(\sigma_*^{-j}) + \textstyle\sum_{s^{-j} \in S^{-j}} \lambda_{s^{-j},t'}^{\sigma^{-j}} u_{s^j}(s^{-j})],$$

or

$$\varepsilon_{t'} \geq \lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} \gamma + \textstyle\sum_{s^{-j} \in S^{-j}} \lambda_{s^{-j},t'}^{\sigma^{-j}} \overbrace{[\max_{s^j \in S^j} u_{s^j}(s^{-j}) - u_{\hat{s}^j}(s^{-j})]}^{\geq 0}$$
$$> \lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} \gamma - \delta_{t'} \textstyle\sum_{s^{-j} \in S^{-j}}[\max_{s^j \in S^j} u_{s^j}(s^{-j}) - u_{\hat{s}^j}(s^{-j})].$$

Because $\gamma > 0$, $\lim_{t' \to \infty} \varepsilon_{t'} = 0$, $\lim_{t' \to \infty} \delta_{t'} = 0$, and $\lim_{t' \to \infty} \lambda_{\sigma_*^{-j},t'}^{\sigma^{-j}} = 1$, this cannot

hold, so for sufficiently large t^*, $\phi_t^j(\hat{s}^j) = 0$ if $t > t^*$. Put in another way: there is a strategy $\hat{s}^j$, for which $\sigma_*^j(\hat{s}^j) > 0$, that is not in the history from t^* onwards. Because of the adaptivity of the assessment rules (Eq. (6)), $\sigma^j(\hat{s}^j)$ will be arbitrarily small for all $\sigma^j \in \mu_t^j$ for sufficiently large t. This contradicts (j's opponent's version of) Eq. (A.2), which states that $\sigma^j(\hat{s}^j)$ is at least $\lambda_{\sigma_*^{-j},t}^{\sigma^{-j}} \sigma_*^j(\hat{s}^j) - \delta_t$ (strictly positive for sufficiently large t). We conclude that σ_* must be an equilibrium. $\quad\square$

Proof of Theorem 3. Considering that the assessment rules are adaptive (Eq. (6)), it holds for both players i that for all t and all $\varepsilon > 0$

$$\exists t_\varepsilon > t\colon \forall t' > t_\varepsilon\colon \forall h_{t'}^{-i} \in \mathcal{H}_{t'}^{-i}\colon \forall s^{-i} \in S^{-i} \text{ such that } n_{t'}^{-i}(s^{-i}) = n_t^{-i}(s^{-i})\colon \forall \sigma^{-i} \in \mu_{t'}^{-i}\colon$$
$$\sigma^{-i}(s^{-i}) < \varepsilon.$$

We can always choose $t_\varepsilon > t_0$ (implying that $\phi_{t'} = s_*$ for all $t' > t_\varepsilon$). Thus $n_{t'}^{-i}(s^{-i}) = n_{t_0}^{-i}(s^{-i})$ will hold for every $s^{-i} \neq s_*^{-i}$. This means every $\sigma^{-i} \in \mu_t^{-i}$ can be written as $\sigma^{-i} = \sigma^{-i}(s_*^{-i})s_*^{-i} + \sum_{s^{-i} \in S^{-i}, s^{-i} \neq s_*^{-i}} \sigma^{-i}(s^{-i})s^{-i}$, where $\sigma^{-i}(s^{-i}) < \varepsilon$ if $s^{-i} \neq s_*^{-i}$, due to adaptivity. Because ε can be chosen arbitrarily small, we can create a sequence converging to 0 that fits the conditions of Lemma 5. Applying Lemma 5 completes the proof. This is done by identifying $\sigma_* = s_*$ and $\lambda_{s^{-i},t}^{\sigma^{-i}} = \sigma^{-i}(s^{-i})$ for all s^{-i} in S^{-i}. $\quad\square$

Proof of Theorem 4. Considering that the assessment rules are asymptotically empirical (Eq. (7)), i.e., that

$$\forall \varepsilon > 0\colon \exists t_0 > 0\colon \forall t > t_0\colon \ \sup_{\sigma^{-i} \in \mu_t^{-i}, s^{-i} \in S^{-i}} |v^{-i}(h_t^{-i}, s^{-i}) - \sigma^{-i}(s^{-i})| < \tfrac{\varepsilon}{2}$$

and using the assumptions of Theorem 4, i.e., that

$$\forall \varepsilon > 0\colon \exists t_1 > 0\colon \forall t > t_1\colon \ \sup_{\sigma^{-i} \in \mu_t^{-i}, s^{-i} \in S^{-i}} |\sigma_*^{-i}(s^{-i}) - v^{-i}(h_t^{-i}, s^{-i})| < \tfrac{\varepsilon}{2},$$

it holds for both players i that for all $\varepsilon > 0$ there exists a $t_2 = \max(t_0, t_1)$ such that for all $t > t_2$ and for all $s^{-i} \in S^{-i}$

$$\sup_{\sigma^{-i} \in \mu_t^{-i}} |\sigma_*^{-i}(s^{-i}) - \sigma^{-i}(s^{-i})|$$
$$= \sup_{\sigma^{-i} \in \mu_t^{-i}} |\sigma_*^{-i}(s^{-i}) - v^{-i}(h_t^{-i}, s^{-i}) + v^{-i}(h_t^{-i}, s^{-i}) - \sigma^{-i}(s^{-i})|$$
$$\leq \sup_{\sigma^{-i} \in \mu_t^{-i}} (|\sigma_*^{-i}(s^{-i}) - v^{-i}(h_t^{-i}, s^{-i})| + |v^{-i}(h_t^{-i}, s^{-i}) - \sigma^{-i}(s^{-i})|)$$
$$< \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon.$$

This implies that every $\sigma^{-i} \in \mu_t^{-i}$ can be written as $\sigma^{-i} = \sigma_*^{-i} + \sum_{s^{-i} \in S^{-i}} \lambda_{s^{-i},t}^{\sigma^{-i}} s^{-i}$, where $\lambda_{s^{-i},t}^{\sigma^{-i}} < \varepsilon$ for all $t > t_2$. Because ε can be chosen arbitrarily small, we can create a sequence converging to 0 that fits the conditions of Lemma 5. Applying Lemma 5 again completes the proof. $\quad\square$

References

[1] J. O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd Edition, Springer Series in Statistics, Springer, New York, 1985.

[2] G. W. Brown, Iterative Solutions of Games by Fictitious Play, in: T. C. Koopmans (Ed.), Activity Analysis of Production and Allocation, no. 13 in Cowles commission monographs, Wiley, New York, 1951, pp. 374–376.

[3] M. D. Davis, Game Theory, A Nontechnical Introduction, Republication of the revised 1983 Edition, Dover Publications, Mineola, New York, 1997.

[4] M. H. DeGroot, Optimal statistical decisions, Wiley Classics Library 2004 Edition, Wiley, Hoboken, New Jersey, 1970.

[5] D. Fudenberg, D. M. Kreps, Learning Mixed Equilibria, Games Econ. Behav. 5 (1993) 320–367.

[6] D. Fudenberg, D. K. Levine, Consistency and cautious fictitious play, Journal of Economic Dynamics and Control 19 (1995) 1065–1089.

[7] D. Fudenberg, D. K. Levine, The Theory of Learning in Games, MIT Press, Cambridge, Massachusetts, 1998.

[8] J. C. Harsanyi, Subjective Probability and the Theory of Games: Comments on Kadane and Larkey's Paper, Management Sci. 28 (2) (1982) 120–124.

[9] J. B. Kadane, P. D. Larkey, Subjective Probability and the Theory of Games, Management Sci. 2 (2) (1982) 113–120.

[10] I. Levi, The Enterprise of Knowledge, MIT Press, London, 1980.

[11] J. Robinson, An iterative method of solving a game, Ann. Math. 54 (1951) 296–301.

[12] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.

[13] P. Walley, Inferences from Multinomial Data: Learning about a Bag of Marbles, J. Roy. Statistical Society B 58 (1996) 3–57.