Game-Theoretic Learning Using the Imprecise Dirichlet Model

E. QUAEGHEBEUR
Ghent University, Belgium

G. DE COOMAN
Ghent University, Belgium

Abstract

We discuss two approaches for choosing a strategy in a two-player game. We suppose that the game is played a large number of rounds, which allows the players to use observations of past play to guide them in choosing a strategy.

Central in these approaches is the way the opponent's next strategy is assessed; both a precise and an imprecise Dirichlet model are used. The observations of the opponent's past strategies can then be used to update the model and obtain new assessments. To some extent, the imprecise probability approach allows us to avoid making arbitrary initial assessments.

To be able to choose a strategy, the assessment of the opponent's strategy is combined with rules for selecting an optimal response to it: a so-called best response or a maximin strategy. Together with the updating procedure, this allows us to choose strategies for all the rounds of the game.

The resulting playing sequence can then be analysed to investigate if the strategy choices can converge to equilibria.

Keywords

game theory, fictitious play, equilibria, imprecise Dirichlet model, learning

1 Introduction

In [4] and [5], Fudenberg et al. have proved a number of convergence results concerning methods for learning optimal strategies in a game-theoretic context. They show that these results hold in particular for *fictitious play* in strictly competitive two-player games in strategic form. In this context, a player bases his learning method on the assumption that his opponent uses a fixed, but unknown, mixed strategy. The pure strategies that his opponent actually plays are consequently assumed to be iid observations of the random *multinomial process* that has this mixed strategy as its probability mass function. The player then uses a Bayesian statistical updating scheme, where the prior is chosen from among a class of models that is conjugate with the multinomial likelihood function, namely the Dirichlet priors, mainly because such a choice allows for simple updating rules.

In the present work, we investigate how this learning method is influenced by replacing the Dirichlet priors by so-called imprecise Dirichlet priors, first introduced by Walley [9], and we provide generalisations for Fudenberg's convergence results that can be applied to the new learning method.

1.1 The Game

We consider *strictly competitive two-player games in the strategic form*; [3, Chapter 2], [5, Chapter 1]. One player is denoted by i and his opponent by $-i$, where $i \in \{-1, 1\}$.

Player i has a finite set $S^i = \{1, \dots, N^i\}$ of *pure strategies* s^i. After each round of the game, he receives a (possibly negative) *pay-off* $u^i(s^i, s^{-i})$, with $s^i \in S^i$ and $s^{-i} \in S^{-i}$. This pay-off is assumed to be expressed in units of some predetermined linear utility, e.g. probability currency; [7, Sections 13 and 14], [8, Section 2.2.2].

Instead of choosing a pure strategy, player i can also choose a so-called *mixed strategy* σ^i, which is a probability mass function on the set S^i. This amounts to using a randomisation device that chooses a pure strategy from S^i, with the probabilities for each pure strategy defined by the mixed strategy σ^i. These can be written as a vector of length N^i with $\sum_{s^i} \sigma^i(s^i) = 1$. We denote the set of these mixed strategies by Σ^i. In what follows, unless otherwise indicated, s^i will always be an element of S^i and σ^i will always be an element of Σ^i.

When using mixed strategies, only the *expected pay-off* can be calculated,

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

It should be clear that pure strategies can be considered as border-case, or degenerate, mixed strategies. The set of all mixed strategies Σ^{-i} can be represented as the unit simplex in $\mathbb{R}^{N^{-i}}$. Pure strategies correspond to the vertices of the simplex. The distance between two strategies is measured using the sup-norm,[1]

$$d(\sigma^{-i}, \tau^{-i}) = \sup_{s^{-i} \in S^{-i}} |\sigma^{-i}(s^{-i}) - \tau^{-i}(s^{-i})|.$$

Observe that the convex unit simplex is compact under this norm.

1.2 Our Objective

We wish to formulate a procedure that guides the players in their strategy choices in such a way, that, using the information they have at their disposal, their expected pay-off is in some sense optimal.

[1] This allows for a nice interpretation, but any norm generating the usual topology could be used.

2 Assessing the Opponent's Strategy

It is essential that the information player i has about the strategy σ^{-i} that his opponent will play, is modelled in a manner that is useful, in light of the objective above, for choosing a strategy σ^i in response to σ^{-i}. In this section we describe two uncertainty models for representing such information. The first is a precise probability model, the second is imprecise.

2.1 Gambles

The available information about his opponent's strategy σ^{-i} will lead player i to accept or reject gambles whose outcome depends on σ^{-i}. Both the uncertainty models described later intend to model player i's behavioural dispositions toward such gambles. A *gamble* X on Σ^{-i} is a bounded real-valued map on Σ^{-i}. It represents an uncertain reward: it yields the amount $X(\sigma^{-i})$ if player $-i$ decides to play the mixed strategy σ^{-i}. The set of all gambles on Σ^{-i} is denoted by $\mathcal{L}(\Sigma^{-i})$; [8, Section 1.5.6]. Two types of gambles are of special interest.

If player i decides to play strategy σ^i, then the game will result in an expected pay-off that still depends on the strategy σ^{-i} that his opponent will play. Thus, we can associate the *strategy gamble* X_{σ^i} on Σ^{-i} with this strategy σ^i by defining $X_{\sigma^i}(\sigma^{-i}) = u^i(\sigma^i, \sigma^{-i})$ for all σ^{-i} in Σ^{-i}. It represents the uncertain expected pay-off for player i if he chooses strategy σ^i. Every gamble in the subset $\mathcal{K}^i = \{X_{\sigma^i} : \sigma^i \in \Sigma^i\}$ of $\mathcal{L}(\Sigma^{-i})$ is thus an *uncertain expected pay-off*. The distance between two strategy gambles is measured using the sup-norm,

$$d(X_{\sigma^i}, X_{\tau^i}) = \sup_{\sigma^{-i} \in \Sigma^{-i}} |X_{\sigma^i}(\sigma^{-i}) - X_{\tau^i}(\sigma^{-i})|.$$

Proposition 1 *The set of strategy gambles $\mathcal{K}^i$ is convex and compact under the* sup*-norm topology on Σ^{-i}.*

Another type of gamble on Σ^{-i}, specifically associated with a pure strategy s^{-i}, is the *evaluation gamble* $Y_{s^{-i}} : \Sigma^{-i} \to [0,1]$ defined by $Y_{s^{-i}}(\sigma^{-i}) = \sigma^{-i}(s^{-i})$. This definition implies that $\sum_{s^{-i}} Y_{s^{-i}} = 1$. Each of these gambles yields the unknown probability mass of the pure strategy s^{-i} defined by (the unknown) probability mass function σ^{-i}. Using this notation, the vector $Y^{-i} = (Y_1, \ldots, Y_{N^{-i}})$ of evaluation gambles returns to the unknown mixed strategy $\sigma^{-i} = Y^{-i}(\sigma^{-i})$ itself.

Using Eq. (1), it is possible to write each strategy gamble as a linear combination of evaluation gambles,

$$X_{\sigma^i} = \sum_{s^{-i} \in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i}) \sigma^i(s^i) \right) Y_{s^{-i}}. \tag{2}$$

2.2 The Precise Dirichlet Model

First we consider a model that specifies the information available to player i as a *linear prevision* P on some subset of $\mathcal{L}(\Sigma^{-i})$; [2, Chapter 3], [8, Section 2.8]. $P(X)$ is player i's *fair price*, or *prevision*, for the gamble X, i.e., the unique real number such that he is disposed to buy the gamble X for all prices $p < P(X)$ and to sell X for all prices $p > P(X)$.

If we define $\pi_P = P(Y^{-i}) = (P(Y_1), \dots, P(Y_{N^{-i}}))$, then the properties of linear previsions allow us to conclude that $\sum_{s^{-i}} \pi_P(s^{-i}) = 1$ and $0 \leq \pi_P(s^{-i}) \leq 1$. We see that π_P is a possible mixed strategy for the opponent. It is player i's prevision of the strategy that his opponent will play. Using Eq. (2) and the linearity of the operator P, we can write for the prevision of the strategy gamble X_{σ^i}:

$$P(X_{\sigma^i}) = \sum_{s^{-i} \in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i}) \sigma^i(s^i) \right) \pi_P(s^{-i}) = X_{\sigma^i}(\pi_P), \tag{3}$$

i.e., the expected pay-off if the opponent were actually to play strategy π_P.

The linear prevision P we shall use here is a *precise Dirichlet model* (PDM) $P(\cdot \mid \beta_t, \rho_t)$, where $\beta_t > 0$ and ρ_t is a mixed strategy in the interior $\operatorname{int}(\Sigma^{-i})$ of Σ^{-i}, i.e., $\rho_t(s^{-i}) > 0$ for all $s^{-i} \in S^{-i}$. This PDM is defined for all measurable gambles X on Σ^{-i} by

$$P(X \mid \beta_t, \rho_t) = \frac{1}{B(\beta_t, \rho_t)} \int_{\Sigma^{-i}} X(\sigma^{-i}) f(\sigma^{-i} \mid \beta_t, \rho_t) \mathrm{d}\sigma^{-i}, \tag{4}$$

where f and the normalisation constant B define the parametrised[2] Dirichlet probability density function,

$$f(\sigma^{-i} \mid \beta_t, \rho_t) = \prod_{s^{-i} \in S^{-i}} \sigma^{-i}(s^{-i})^{\beta_t \rho_t(s^{-i}) - 1} \quad \text{and} \quad B(\beta_t, \rho_t) = \frac{\prod_{s^{-i}} \Gamma(\beta_t \rho_t(s^{-i}))}{\Gamma(\beta_t)}.$$

When using such a PDM, the prevision π_P of the strategy his opponent will play coincides with ρ_t:

$$\pi_P = \pi_{P(\cdot \mid \beta_t, \rho_t)} = P(Y^{-i} \mid \beta_t, \rho_t) = \rho_t.$$

This means that for the calculation of $P(X_{\sigma^i} \mid \beta_t, \rho_t)$ we don't need to use Eq. (4), but that we can use Eq. (3), replacing π_P by ρ_t:

$$P(X_{\sigma^i} \mid \beta_t, \rho_t) = \sum_{s^{-i} \in S^{-i}} \left(\sum_{s^i \in S^i} u^i(s^i, s^{-i}) \sigma^i(s^i) \right) \rho_t(s^{-i}) = X_{\sigma^i}(\rho_t).$$

[2] We use a non-standard parametrisation, because it is more convenient in this context; [9].

2.3 The Imprecise Dirichlet Model

Next, we consider an imprecise probability model for the information player i has about his opponent's strategy. This can always be made to take the form of a coherent *lower prevision* $\underline{P}$ on some subset of $\mathcal{L}(\Sigma^{-i})$; [8, Section 2.3]. $\underline{P}(X)$ specifies player i's supremum acceptable price for buying the gamble X, i.e., it is the greatest real number p such that he is disposed to buying the gamble X for all prices strictly smaller than p.

The lower prevision $\underline{P}$ we shall use here is an *imprecise Dirichlet model* (IDM) $\underline{P}(\cdot \mid \beta_t, M_t)$, where $\beta_t > 0$ and $M_t \subseteq \operatorname{int}(\Sigma^{-i})$; [9]. This IDM is defined for all measurable gambles X on Σ^{-i} as the lower envelope of a set of PDM's (with a common β_t, but each with their own ρ_t),

$$\underline{P}(X \mid \beta_t, M_t) = \inf\{P(X \mid \beta_t, \rho_t) \colon \rho_t \in M_t \subset \Sigma^{-i}\}. \tag{5}$$

3 Choosing an Optimal Strategy

When choosing an optimal strategy, it is important to be clear on what defines optimality. In this game-theoretic context, it is desirable to attain a pay-off that is as high as possible, but on the other hand it may also be important to limit possible losses. These are the guiding criteria in our search for optimal strategies [6, Section 3.8].

3.1 Admissible Strategies, Maximin Strategies, Best Replies

If for two strategies τ^i and σ^i, the pay-off for τ^i is always at least as high as that for σ^i, i.e., $X_{\tau^i} \geq X_{\sigma^i}$ or in other words $(\forall \sigma^{-i} \in \Sigma^{-i})(X_{\tau^i}(\sigma^{-i}) \geq X_{\sigma^i}(\sigma^{-i}))$, we say that τ^i *dominates* σ^i—or that X_{τ^i} dominates X_{σ^i}; [5, Section 1.7.2].

A strategy $\sigma^i \in \Sigma^i$, or its corresponding strategy gamble $X_{\sigma^i} \in \mathcal{K}^i$, is called *inadmissible* if there is another strategy τ^i that strictly dominates it: $X_{\tau^i} \geq X_{\sigma^i}$ and $X_{\sigma^i} \neq X_{\tau^i}$. Otherwise, it is called *admissible*. We consider an admissible strategy to be more optimal than an inadmissible strategy. However, the discussion of, and the results deduced for, the learning models below is not essentially affected when this distinction is not made.

Now suppose that player i knows that his opponent will play some strategy in $M \subseteq \Sigma^{-i}$, but nothing more. When playing σ^i, his expected pay-off will at least be $\inf_{\sigma^{-i} \in M} X_{\sigma^i}(\sigma^{-i})$. An *M-maximin strategy* τ^i maximises this minimal pay-off:

$$\tau^i \in \operatorname*{argmax}_{\sigma^i \in \Sigma^i} \inf_{\sigma^{-i} \in M} X_{\sigma^i}(\sigma^{-i}).$$

Proposition 2 *There are admissible M-maximin strategies for any compact subset M of Σ^{-i}.*

When $M = \Sigma^{-i}$, player i doesn't have a clue about his opponent's strategy choice, and the corresponding Σ^{-i}-maximin strategy is simply called a *maximin strategy*.

Corollary 1 *There are always admissible maximin strategies.*

At the other extreme, player i knows his opponent will play a strategy σ^{-i}. Any corresponding $\{\sigma^{-i}\}$-maximin strategy is called a *best reply* to σ^{-i}. The set of all best replies to σ^{-i} is denoted by $BR^i(\sigma^{-i})$.

Corollary 2 *There are always admissible best replies to any strategy σ^{-i} in Σ^{-i}.*

This set of best replies has some interesting properties.

Proposition 3 *For all σ^{-i} in Σ^{-i}, $BR^i(\sigma^{-i})$ is a compact and convex subset of Σ^i. Moreover, if $\sigma^i \in BR^i(\sigma^{-i})$ and $\sigma^i(s^i) > 0$ for some $s^i \in S^i$, then $s^i \in BR^i(\sigma^{-i})$.*

For $M \subseteq \Sigma^{-i}$, the collection of best replies to strategies in M is denoted by $BR^i(M)$ and given by

$$BR^i(M) = \bigcup_{\sigma^{-i} \in M} BR^i(\sigma^{-i}).$$

Proposition 4 *For any subset M of Σ^{-i} that is convex and closed, the M-maximin strategies make up a subset of $BR^i(M)$.*

Corollary 3 *There are always admissible best replies to any convex and closed subset M of Σ^{-i}.*

3.2 Optimal Strategies and the PDM

When using a linear prevision P, any admissible strategy σ^i that maximises $P(X_{\sigma^i})$ is called a *Bayes strategy*. This name refers to the fact that it is an optimal strategy in the usual Bayesian sense of maximising expected utility; [8, Section 3.9].

Eq. (3) tells us that $P(X_{\sigma^i}) = X_{\sigma^i}(\pi_P)$. This means that τ^i is a Bayes strategy whenever $\tau^i \in \operatorname{argmax}_{\sigma^i} X_{\sigma^i}(\pi_P)$. This gives the following result.

Proposition 5 *The set of the Bayes strategies corresponding to a linear prevision P is given by the admissible strategies of $BR^i(\pi_P)$.*

If player i's model for his opponent's strategy is a PDM $P(\cdot \mid \beta_t, \rho_t)$, we find that his optimal (Bayes) strategies are simply the admissible strategies of $BR^i(\rho_t)$.

3.3 Optimal Strategies and the IDM

When using a coherent lower prevision $\underline{P}$, a *maximal strategy* is any admissible strategy σ^i for which $\min_{\tau^i \in \Sigma^i} \overline{P}(X_{\sigma^i} - X_{\tau^i}) \geq 0$; see [8, Section 3.9] for motivation.[3]

[3] To see that this definition generalises that of a Bayes strategy, consider that

$$\sigma^i \in \operatorname*{argmax}_{\tau^i \in \Sigma^i} P(X_{\tau^i}) \Leftrightarrow P(X_{\sigma^i}) \geq \max_{\tau^i \in \Sigma^i} P(X_{\tau^i}) \Leftrightarrow \min_{\tau^i \in \Sigma^i} P(X_{\sigma^i} - X_{\tau^i}) \geq 0.$$

We shall use the notation $\mathcal{M}(\underline{P})$ for the set of linear previsions P that dominate $\underline{P}$ on its domain.

Proposition 6 *A strategy σ^i is maximal under $\underline{P}$*

$\Leftrightarrow$ *σ^i is a Bayes strategy under some P in $\mathcal{M}(\underline{P})$;*

$\Leftrightarrow$ *σ^i is an admissible best reply to π_P for some $P \in \mathcal{M}(\underline{P})$, i.e., the admissible $\sigma^i \in BR^i(M_{\underline{P}})$, where $M_{\underline{P}} = \{\pi_P \colon P \in \mathcal{M}(\underline{P})\} \subseteq \Sigma^i$.*

Corollary 4 *There are maximal strategies under $\underline{P}$.*

There is another optimality criterion associated with a lower prevision $\underline{P}$: an admissible mixed strategy σ^i is called *$\underline{P}$-maximin* if it maximises the lower prevision $\underline{P}(X_{\tau^i})$ of all strategy gambles X_{τ^i}, i.e., if $\sigma^i \in \operatorname{argmax}_{\tau^i \in \Sigma^i} \underline{P}(X_{\tau^i})$; [8, Section 3.9]. Since a coherent lower prevision $\underline{P}$ is the lower envelope of its set of dominating linear previsions (see [8, Theorem 3.3.3]), we see that

$$\underline{P}(X_{\tau^i}) = \min_{P \in \mathcal{M}(\underline{P})} P(X_{\tau^i}) = \min_{\sigma^{-i} \in M_{\underline{P}}} X_{\tau^i}(\sigma^{-i}),$$

and consequently, the admissible mixed strategy σ^i is $\underline{P}$-maximin if and only if $\sigma^i \in \operatorname{argmax}_{\tau^i \in \Sigma^i} \min_{\sigma^{-i} \in M_{\underline{P}}} X_{\tau^i}(\sigma^{-i})$, i.e., if it is $M_{\underline{P}}$-maximin. We know from Section 3.1 that all the $M_{\underline{P}}$-maximin strategies also belong to $BR^i(M_{\underline{P}})$.

Corollary 5 *For any coherent lower prevision $\underline{P}$, there are $\underline{P}$-maximin strategies. They coincide with the admissible $M_{\underline{P}}$-maximin strategies, and are in particular also maximal strategies under $\underline{P}$.*

If player i models his uncertainty about his opponent's strategy by an IDM $\underline{P}(\cdot \mid \beta_t, M_t)$, we have proved the following results, using the continuity of Y^{-i} and the properties of $\mathcal{M}(\underline{P}(\cdot \mid \beta_t, M_t))$.

Theorem 1 *If M_t is a subset of $\operatorname{int}(\Sigma^{-i})$, then the set $M_{\underline{P}(\cdot|\beta_t,M_t)}$ is the closed convex hull $\overline{\operatorname{co}}(M_t)$ of M_t.*

We thus find that the optimal strategies in this imprecise model are the admissible elements of $BR^i(\overline{\operatorname{co}}(M_t))$. Moreover, if player i wants to play it safe (maximise his minimal expected gains), he can use admissible $\overline{\operatorname{co}}(M_t)$-maximin strategies.

4 Playing the Game Over and Over Again

We now turn our attention to how the proposed models, the PDM and the IDM, can be used when a number of rounds of the game are played. We specifically look at the way observations of past play can change the assessments of a player and we formulate an algorithm to guide the players in their strategy choices.

4.1 Learning from Past Play

After playing t rounds of the game, player i has observed a so-called *history* $\zeta_t^{-i} \in \mathcal{Z}_t^{-i} = (S^{-i})^t$ of the pure strategies $\zeta_t^{-i}(k)$, $k = 1, \dots, t$, that his opponent has played.

If player i supposes that his opponent plays a fixed mixed strategy σ^{-i},[4] which is of course not necessarily the case, the order of the strategies in the history does not matter and the observed strategies can be considered as outcomes of a multinomial iid process. As a sufficient statistic for σ^{-i} he can then use the N^{-i}-tuple n^{-i} of *observed occurrences* for which each component $n^{-i}(s^{-i})$ is the number of times his opponent has played $s^{-i} \in S^{-i}$, and which is consequently a random variable with the multinomial distribution. The total number t of rounds played is also equal to $\sum_{s^{-i}} n^{-i}(s^{-i})$. The N^{-i}-tuple of *observed frequencies* $\frac{n^{-i}}{t}$ is denoted by κ_t^{-i} and can be considered to be an element of Σ^{-i}.

The likelihood function for n^{-i} is

$$L_{n^{-i}}(\sigma^{-i}) = \frac{t!}{\prod_{s^{-i}} n^{-i}(s^{-i})!} \prod_{s^{-i} \in S^{-i}} \sigma^{-i}(s^{-i})^{n^{-i}(s^{-i})}.$$

Using Bayes' rule, we can now update (see e.g. [5, Chapter 2]) a prior Dirichlet density function $f(\sigma^{-i} \mid \beta_0, \rho_0)$ with the observations n^{-i},

$$\begin{aligned} f(\sigma^{-i} \mid \beta_0, \rho_0, n^{-i}) &= \frac{1}{P(L_{n^{-i}} \mid \beta_0, \rho_0)} f(\sigma^{-i} \mid \beta_0, \rho_0) L_{n^{-i}}(\sigma^{-i}) \\ &= f(\sigma^{-i} \mid \beta_0 + t, \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}) \\ &= f(\sigma^{-i} \mid \beta_t, \rho_t). \end{aligned}$$

We see that the posterior density function $f(\sigma^{-i} \mid \beta_t, \rho_t)$ is still a Dirichlet density function. This means that that the Dirichlet density functions constitute a *conjugate* family of density functions for the multinomial sampling likelihood function $L_{n^{-i}}$. Observe that $P(L_{n^{-i}} \mid \beta_0, \rho_0)$ has to be non-zero, which is guaranteed by $\beta_0 > 0$ and $\rho_0 \in \operatorname{int}(\Sigma^{-i})$.

4.2 Updating a Dirichlet model

When updating a prior PDM $P(\cdot \mid \beta_0, \rho_0)$ after t rounds, we find that we simply have to update the parameters,

$$\beta_0 \to \beta_t = \beta_0 + t \quad \text{and} \quad \rho_0 \to \rho_t = \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}, \tag{6}$$

[4] This corresponds to the underlying assumption used in so-called *fictitious play*; [5, Chapter 2].

to obtain the posterior PDM $P(\cdot \mid \beta_t, \rho_t)$. It is clear that first updating with n^{-i} and then updating the new model with m^{-i} is equivalent to updating the original model with $n^{-i} + m^{-i}$.

When updating a prior IDM $\underline{P}(\cdot \mid \beta_0, M_0)$ after t rounds, the answer is a bit more complicated. It is possible that there are n^{-i} for which $\underline{P}(L_{n^{-i}} \mid \beta_0, M_0) = 0$ even with $M_0 \subseteq \mathrm{int}(\Sigma^{-i})$, i.e., for $\underline{P}(L_{n^{-i}} \mid \beta_0, M_0) > 0$ we need $P(L_{n^{-i}} \mid \beta_0, \rho_0) > 0$ for all $\rho_0 \in \overline{\mathrm{co}}(M_0)$. However, using the notion of *regular extension*, we can find a unique posterior IDM that is coherent with $\underline{P}(\cdot \mid \beta_0, M_0)$ and that satisfies the additional rationality axiom of *regularity*; [8, Appendix J]. This posterior lower prevision turns out to be the lower envelope of the updated PDM's,

$$\inf_{\rho_0 \in M_0} P(X \mid \beta_0 + t, \frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t}) = \inf_{\rho_t \in M_t} P(X \mid \beta_t, \rho_t) = \underline{P}(X \mid \beta_t, M_t),$$

where β_t and M_t are the parameters of the updated *IDM*,

$$\beta_0 \to \beta_t = \beta_0 + t \quad \text{and} \quad M_0 \to M_t = \{\frac{\beta_0\rho_0 + n^{-i}}{\beta_0 + t} : \rho_0 \in M_0\}. \tag{7}$$

4.3 Iterative Playing Algorithm: Assess, Decide and Update

Our generic guiding algorithm for player i playing multiple rounds of a strictly competitive two-player game consists of three steps; [4, Section 3]. Assume that t rounds have already been played, and that the history ζ_t^{-i} of the pure strategies played by the opponent during these rounds is available to player i. He is about to play a new round and uses some model to describe the information he has.

1. Player i has to make an assessment $\mu^i(\zeta_t^{-i})$ about the data that are relevant for his strategy choice: to this end, he uses an *assessment rule* μ^i.

2. Player i has to use a *decision rule* ϕ^i to choose a strategy $\phi^i(\zeta_t^{-i})$ to play, using his assessments $\mu^i(\zeta_t^{-i})$.

3. After the round is played, player i should use the observation of his opponent's strategy to *update* his information.

Let us now see what this algorithm becomes for the two types of uncertainty models described above.

When using a PDM $P(\cdot \mid \beta_t, \rho_t)$, we can formulate the following implementation of the algorithm.

1. Let $\mu^i(\zeta_t^{-i}) = \rho_t = \pi_{P(\cdot \mid \beta_t, \rho_t)}$, the prevision of the opponent's strategy.

2. Let $\phi^i(\zeta_t^{-i})$ be some (admissible) element of $BR^i(\rho_t)$.

3. Update the PDM to $P(\cdot \mid \beta_{t+1}, \rho_{t+1})$ using Eq. (6).

Initially, player i has to choose a ρ_0 and β_0. The parameter β_0 can be interpreted as the number of *pseudocounts*[5] associated with the initial prevision of his opponent's strategy ρ_0, for which any choice is arbitrary (if it is not based on some information).

When using an IDM $\underline{P}(\cdot \mid \beta_t, M_t)$, we can formulate two different implementations of the algorithm, different only in their choice of behaviour rule.

1. Let $\mu^i(\zeta_t^{-i}) = \overline{\mathrm{co}}(M_t) = M_{\underline{P}(\cdot|\beta_t, M_t)}$.
2. (a) If we consider maximality as the optimality criterion, then let $\phi^i(\zeta_t^{-i})$ be some (admissible) element of $BR^i(\overline{\mathrm{co}}(M_t))$.

   (b) If we consider maximinity as the optimality criterion, then let $\phi^i(\zeta_t^{-i})$ be some (admissible) $\overline{\mathrm{co}}(M_t)$-maximin strategy.
3. Update the IDM to $\underline{P}(\cdot \mid \beta_{t+1}, M_{t+1})$ using Eq. (7).

Initially, player i has to choose an M_0 and a number of pseudocounts β_0. When he has no information available, an obvious choice for M_0 is $\mathrm{int}(\Sigma^{-i})$, which corresponds to so-called near-ignorance [8, Section 4.6.9]. The choice for the best reply behaviour rule or the maximin behaviour rule will not influence the results of Section 5 in any way.

5 Equilibria and Convergence

Now that we have two learning models, the PDM and the IDM, at our disposal, we can investigate the game-play that results from using them. We start by giving some definitions that are essential for the ensuing analysis.

5.1 Strategy Profiles and Equilibria

To be able to analyse the game-play that results from the assessment and behaviour rules discussed in Section 4.3, we introduce some new notation and recall the concept of an equilibrium.

A couple of strategies of the players is called a *strategy profile*, which can be pure $s = (s^i, s^{-i}) \in S = S^i \times S^{-i}$, or mixed $\sigma = (\sigma^i, \sigma^{-i}) \in \Sigma = \Sigma^i \times \Sigma^{-i}$. A corresponding *profile history* after t rounds of play is denoted by $\zeta_t \in \mathcal{Z}_t = S^t$.

The notation $\sigma(s)$ corresponds to $(\sigma^i(s^i), \sigma^{-i}(s^{-i}))$. Likewise, we write

$$\begin{aligned} BR(\sigma) &= BR^i(\sigma^{-i}) \times BR^{-i}(\sigma^i) \subseteq \Sigma, \\ \mu(\zeta_t) &= \mu^i(\zeta_t^{-i}) \times \mu^{-i}(\zeta_t^i) \subseteq \Sigma, \\ \phi(\zeta_t) &= (\phi^i(\zeta_t^{-i}), \phi^{-i}(\zeta_t^i)) \in \Sigma. \end{aligned}$$

[5] In the literature, the values 1 and 2 are found for prior models that are not based on any information; [9].

An *equilibrium* is a strategy profile for which the pay-off for both players cannot be increased if one of them changes his strategy, while his opponent’s strategy remains unchanged; [3]. This means that

$$\sigma_* \text{ is an equilibrium } \Leftrightarrow (\forall i)\Big(u^i(\sigma_*) = \max_{\tau^i \in \Sigma^i} u^i(\tau^i, \sigma_*^{-i})\Big) \Leftrightarrow \sigma_* \in BR(\sigma_*).$$

If $s_* = BR(s_*)$, then s_* is a *strict equilibrium.*[6] A game can have multiple (strict) equilibria.[7]

5.2 Assessment Rules

The definitions in this section and in the next are generalisations of the definitions given by Fudenberg and Kreps in [4] to learning models with assessments $\mu^i(\zeta_t^{-i})$ that are set-valued rather than point-valued.

An important characterisation of possible assessment rules can be made by looking at what the influence is of different parts of a history.

We say that a assessment rule μ^i is *adaptive* if it attaches diminishing importance to earlier parts of the history, as the number of rounds t increases. This means that for all t and all $\varepsilon > 0$,

$$(\exists T > t)\big(\forall t' > T - t\big)\big(\forall \zeta_{t+t'}^{-i} \in \mathcal{Z}_{t+t'}^{-i}\big)\big(\forall \sigma^i \in \mu^i(\zeta_{t+t'}^{-i})\big)\big(\sigma^i(s^i) < \varepsilon\big),$$

for every pure strategy s^i that was not played in the last t' rounds (did not appear in the t' last components of $\zeta_{t+t'}^{-i}$).

A specific subcategory of the adaptive assessment rules can be defined using the observed frequencies κ_t^{-i} of strategies played by the opponent. An assessment rule μ^i is called *asymptotically empirical* if for every *infinite history* $\zeta_\infty^{-i} \in \mathcal{Z}_\infty^{-i}$ it holds that $\lim_{t\to\infty} \sup_{\sigma^{-i} \in \mu^i(\zeta_t^{-i})} d(\sigma^{-i}, \kappa_t^{-i}) = 0$, where the ζ_t^{-i} are partial histories of the selected infinite history ζ_∞^{-i}.

Using the updating formulae (6) and (7), we obtain the following result.

Theorem 2 *The assessment rules of the PDM and the IDM are asymptotically empirical, and thus adaptive.*

5.3 Behaviour Rules

It is clear that the behaviour rules ϕ determine which histories are possible. A history is called *compatible* with the behaviour rules ϕ used by the players if it can be generated (with non-zero probability) by these behaviour rules. Explicitly, this

[6] By Proposition 3, only pure strategy profiles can be strict equilibria.

[7] When only admissible strategies are considered optimal, some equilibria might not be playable. There is always at least one admissible equilibrium.

means that for every pure profile $\zeta_t(k)$, $k = 1, \ldots, t$, that is a component of a compatible profile history, both components of $\phi(\zeta_{k-1})(\zeta_t(k))$ are strictly positive, so the randomisation devices used by the players can select the pure strategies $\zeta_t^i(k)$ and $\zeta_t^{-i}(k)$ with non-zero probability.

It is useful to know to what degree the behaviour rules ϕ used by the players succeed in attaining the objective of optimality (see Section 1.2). The characterisations in this section do just this, and give a clear interpretation of this objective, keeping in mind that the players suppose that their opponent plays an unknown, but fixed, mixed strategy.

We call a behaviour rule ϕ^i *set-myopic* relative to the assessment rule μ^i if, for all t and histories ζ_t^{-i}, it holds that $\phi^i(\zeta_t^{-i}) \in BR^i(\mu^i(\zeta_t^{-i}))$. When the assessments $\mu^i(\zeta_t^{-i})$ are point-valued, the prefix 'set' in set-myopic is dropped.

We now define a weakening of the notion of a set-myopic behaviour rule. We call a behaviour rule ϕ^i *strongly asymptotically set-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 ζ_t^{-i}, it holds that

$$\big(\forall \sigma^{-i} \in \mu^i(\zeta_t^{-i})\big)\big(\forall \tilde{s}^i \in S^i \text{ such that } \phi^i(\zeta_t^{-i})(\tilde{s}^i) > 0\big)$$
$$\Big(u^i(\tilde{s}^i, \sigma^{-i}) + \varepsilon_t \geq \max_{s^i \in S^i} u^i(s^i, \sigma^{-i})\Big).$$

Using the definitions from Section 4.3 the next result is immediate.

Theorem 3 *The behaviour rules for the IDM are set-myopic and the behaviour rule for the PDM is myopic.*

5.4 Convergence to equilibria

An interesting theorem about strict equilibria follows directly from the definitions of a strict equilibrium and of a myopic behaviour rule; [4].

Theorem 4 (absorption to a strict equilibrium) *If there is a strict equilibrium s_* that is played in some round t of a profile history ζ_t compatible with a myopic behaviour rule ϕ, then s_* will be played during all subsequent rounds $t' > t$.*

This theorem holds for the PDM (with myopic behaviour rules), due to Theorem 3, but not for the IDM (with set-myopic behaviour rules), because we have been able to show that the selected mixed strategy $\phi^i(\zeta_t^{-i})$ under both optimisation criteria can still be different from s_*^i due to the fact that $\mu^i(\zeta_t^{-i}) = \overline{\text{co}}(M_t)$ is a set. It is possible to tighten the conditions, to obtain a result that also works for the IDM, i.e., to make sure that best reply only contains s_*.

Theorem 5 (conditional absorption to a strict equilibrium) *If, for some profile history ζ_t compatible with set-myopic behaviour rules ϕ, the strategy profile $\phi(\zeta_t)$ cannot be different from the strict equilibrium s_*, then s_* will be played during all subsequent rounds $t' > t$.*

For equilibria s_* of pure strategies that aren’t necessarily strict, the following result is found.

Theorem 6 (repeated play of a pure strategy profile) *Consider an infinite history ζ_∞ in $\mathcal{Z}_\infty$ such that for some t, a pure strategy profile s_* is played in all subsequent rounds. If ζ_∞ is compatible with behaviour rules ϕ that are strongly asymptotically set-myopic relative to the adaptive assessment rules μ, then s_* is an equilibrium.*

This theorem can be used for both the PDM and the IDM, due to Theorems 2 and 3, even if both players don’t use the same model. For example, one player can use the IDM and his opponent the PDM, or two players can use the IDM, each using a different optimality criterion.

For mixed equilibria σ_*, the following result about the convergence of the observed game-play to a mixed equilibrium, is found.

Theorem 7 (repeated play of a mixed strategy profile) *Let the infinite history ζ_∞ in $\mathcal{Z}_\infty$ be such that for some mixed strategy profile σ_*, it holds that for both players $i \in \{-1,1\}$*

$$\lim_{t\to\infty} \kappa_t^{-i} = \sigma_*^{-i}.$$

If the infinite history ζ_∞ is compatible with behaviour rules ϕ that are strongly asymptotically set-myopic relative to the assessment rules μ that are asymptotically empirical, then σ_* is an equilibrium.

As before, due to Theorems 2 and 3, this theorem can be used for both the PDM and the IDM.

Theorems 6 and 7 can only say that convergence has occurred, but do not indicate when convergence will occur. They could be useful for finding equilibria in large games. As these theorems are generalisations to set-valued assessment rules of theorems found in [4], their proofs are (not always trivial) modifications of the ones found there.

6 Conclusions

6.1 General Remarks

Both the learning models discussed above accomplish our objective of optimality of the expected pay-off quite well. Their convergence properties also favour their use in game theory, notably in the search for equilibria.

The PDM has already been studied in the literature and the learning model based on it is often called *fictitious play* in a game-theoretic setting. The IDM has also been used in different contexts; see [9] for the presentation of the IDM itself and [1], [10], [11] and [12] for examples of possible applications in other areas.

In Section 5 we have in fact generalised the results of Fudenberg and Kreps in [4, Sections 3 and 4], where point-valued assessments $\mu^i(\zeta_t^{-i})$ are used, to set-valued assessments. This is why we formulated Theorems 6 and 7 for a broader class of learning models than our Dirichlet models, which allows Section 5 to be seen as a generalisation of [4, Sections 3 and 4], and not only as a group of results for the PDM and IDM.

We haven't discussed the choice of a specific strategy $\phi^i(\zeta_t^{-i})$ from among the optimal ones. But, if for a specific application other, additional, criteria are available, then using them at this stage will not influence the convergence results in any way.

6.2 PDM vs. IDM

If we compare the PDM to the IDM, the first thing to be said is that the PDM is a special case of the IDM, where $M_0 = \{\rho_0\}$. This immediately indicates the most important advantage of the IDM over the PDM, the possibility of not having to make an arbitrary initial choice, as there is no need to choose one specific prior.

The second advantage of the learning model using the IDM is that it reflects, in its assessment $\mu^i(\zeta_t^{-i})$, the amount of information on which it is based. This corresponds to the fact that the distances between elements of M_t shrink with increasing t. So the model becomes more precise as more observations come in, in the sense that all elements of M_t will lie closer and closer to the ρ_t of any PDM that could have been used.[8]

One disadvantage of the IDM is that it is a more complex model (the player has to work with sets of strategies instead of one strategy). This difference could be reflected by the calculation load for both models.

Acknowledgements

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction initiated by the Belgian state, Prime Minister's Office for Science, Technology and Culture. It was also financed by a Ph.D. grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The scientific responsibility rests with the authors.

References

[1] BERNARD, J.-M. Non-parametric inference about an unknown mean using the imprecise dirichlet model. In *ISIPTA '01 – Proceedings of the Second In-*

[8] It is interesting to note that the successive M_t are similar to one another, because the updating procedure of Eq. (7) corresponds to a contraction and a translation of M_t on the simplex Σ^{-i}.

ternational Symposium on Imprecise Probabilities and Their Applications, G. de Cooman, T. L. Fine, and T. Seidenfeld, Eds. Shaker Publishing, Maastricht, 2000, pp. 40–50.

[2] DE FINETTI, B. *Theory of Probability*, vol. 1. John Wiley & Sons, Chichester, 1974. English Translation of *Teoria delle Probabilità*.

[3] FRIEDMAN, J. W. *Game Theory with Applications to Economics*. Oxford University Press, New York, 1989.

[4] FUDENBERG, D., AND KREPS, D. M. Learning mixed equilibria. *Games and Economic Behaviour 5* (1993), 320–367.

[5] FUDENBERG, D., AND LEVINE, D. K. *The Theory of Learning in Games*, vol. 2 of *The MIT Press Series on Economic Learning and Social Evolution*. The MIT Press, Cambridge, Massachusets and London, England, 1998.

[6] MYERSON, R. B. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, Massachusetts, 1991.

[7] SMITH, C. A. B. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series A 23* (1961), 1–37.

[8] WALLEY, P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[9] WALLEY, P. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B 58* (1996), 3–57. With discussion.

[10] WALLEY, P., AND BERNARD, J.-M. Imprecise probabilistic prediction for categorical data. Tech. Rep. CAF-9901, Laboratoire Cognition et Activités Finalisées, Université de Paris 8, Paris, January 1999.

[11] ZAFFALON, M. Statistical inference of the naive credal classifier. In *ISIPTA '01 – Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, G. de Cooman, T. L. Fine, and T. Seidenfeld, Eds. Shaker Publishing, Maastricht, 2000, pp. 384–393.

[12] ZAFFALON, M. The naive credal classifier. *Journal of Statistical Planning and Inference 105* (2002), 5–21.

Erik Quaeghebeur is a member of the SYSTeMS research group at Ghent University.
address: Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium.
e-mail: erik.quaeghebeur@ugent.be

Gert de Cooman is a member of the SYSTeMS research group at Ghent University.
address: Technologiepark – Zwijnaarde 914, 9052 Zwijnaarde, Belgium.
e-mail: gert.decooman@ugent.be