ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

THE POSSIBLE AND THE IMPOSSIBLE IN MULTI-AGENT LEARNING

H. Peyton Young

Number 304

January 2007

Manor Road Building, Oxford OX1 3UQ

The Possible and the Impossible in Multi-Agent Learning

H. Peyton Young

Department of Economics, University of Oxford

Johns Hopkins University & The Brookings Institution

Email: peyton.young@economics.ox.ac.uk

Abstract

The paper surveys recent work on learning in games and delineates the boundary between forms of learning that lead to Nash equilibrium and forms that lead to weaker notions of equilibrium (or none at all).

JEL Classification: C7, D83. Keywords: *equilibrium, learning, dynamics*

Acknowledgements. I wish to thank Dean Foster and Rakesh Vohra for helpful comments on an earlier version.

The Possible and the Impossible in Multi-Agent Learning

H. Peyton Young

Interactive learning is inherently more complex than single-agent learning, because the act of learning changes the object to be learned. If agent A is trying to learn about agent B, A's behavior will naturally depend on what she has learned so far, and also on what she hopes to learn next. But A's behavior can be observed by B, hence B's behavior may change as a result of A's attempts to learn it. The same holds for B's attempts to learn about A.

This feedback loop is a central and inescapable feature of multi-agent learning situations. It suggests that methods which work for single-agent learning problems may fail in multi-agent settings. It even suggests that learning could fail in general, that is, there may exist situations in which *no* rules allow players to learn one another's behavior in a completely satisfactory sense. This turns out to be the case: in the next section I formulate an *uncertainty principle* for strategic interactions which states that if there is enough ex ante uncertainty about the other players' payoffs (and therefore their potential behaviors), there is no way that rational players can learn to predict one another's behavior, even over an infinite number of repetitions of the game (Foster and Young, 2001; for earlier results in the same spirit see Binmore (1987) and Jordan (1991, 1993)).

Admittedly this and related impossibility theorems rest on very demanding assumptions about agents' rationality, and what it means for them to "learn" their opponents' behavior. Under less restrictive conditions more positive results can be attained, as we shall see in section 3. Thus the purpose of this note is not to claim that multi-agent learning is impossibly difficult, but to try to identify the boundary -- insofar as we now know it -- between the possible and the

impossible in multi-agent learning situations. These issues are discussed in greater depth in Young (2004).

2. Model-based learning

The accompanying perspectives paper by Shoham, Powers and Grenager (2006), hereafter referred to as SPG, forms my jumping-off point. They too draw attention to the fact that multi-agent learning is inherently more complex than single-agent learning. They also make a useful distinction between model-based and model-free forms of learning, which I shall follow here. Using essentially their language, a model-based learning scheme has the following elements:

1. Start with a model of the opponent's strategy.
2. Compute and play a best [or almost best] response.
3. Observe the opponent's play and update your model.
4. Goto step 2.

SPG leave the concept of "model" open, but here I shall suggest a general definition. Namely, a *model-based learning method* is a function that maps any history of play into a prediction about what one's opponents will do next period, that is, to a probability distribution over the opponents' actions conditional on the history so far. This definition encompasses many forms of pattern recognition. The key feature of a model-based learning rule, however, is not what patterns it is able to identify in the data, but how it uses these patterns to forecast the opponents' next moves.

Many game-theoretic learning methods fall into this category. Fictitious play is a simple example: each agent predicts that his opponent will use the distribution next period that he used cumulatively up until now. More generally, Bayesian

updating is a model-based learning procedure: each agent updates his beliefs about the repeated-game strategy of the opponents (conditional on the observed history), which leads to a prediction of their behavior next period.

What exactly do we mean by "learning" in this context? A natural definition is that players "learn" if they eventually succeed in predicting their opponents' behavior with a high degree of accuracy (Foster and Young, 2001). This idea can be given greater precision as follows. Suppose that you are engaged in a two-player game. Given a history h_t to time t, let p_t be your prediction of the opponent's next-period behavior, conditional on h_t. Let q_t be your opponent's actual intended behavior next period, conditional on h_t. Notice that both p_t and q_t are probability distributions over the opponent's action space (which we assume is finite). Thus p_t and q_t lie in an m-dimensional simplex for some nonnegative integer m. The *predictive error* in period t is $\|p_t - q_t\|$. We could say that you *learn to predict* if $\|p_t - q_t\| \to 0$ almost surely as $t \to \infty$. A less demanding definition would be that the mean square error goes to zero: $(1/t) \sum_{s \le t} \|p_s - q_s\|^2 \to 0$ almost surely as $t \to \infty$. We shall say that the former is *learning to predict in the strong sense* and the latter is *learning to predict in the weak sense*.

There is a well-known condition in statistics that guarantees that all players will learn to predict in the strong sense. Namely, it suffices that each player's *forecast* of the others' behavior, conditional on his own behavior, never exclude events that have positive probability under their actual joint behavior. This is the *absolute continuity condition* (Blackwell and Dubins, 1962; Kalai and Lehrer, 1993).

So far we have said nothing about what determines agents' behavior, only what it means for them to learn. In game theory, a standard assumption is that behavior is *rational*: at each point in time, given what has happened to date, the

players' behavioral strategies are optimal given their forecasts of what is going to happen at all future dates. If we combine rationality with the absolute continuity condition (which guarantees good prediction), then we get convergence to Nash equilibrium along the play path (Kalai and Lehrer, 1993).

Suppose, however, that each player is ignorant of his opponent's payoff function. If the opponent is rational, his strategy will depend -- perhaps quite intricately -- on what his payoffs are. Hence the first player will have difficulty forecasting the second player's strategy unless he can gather enough information along the play path to deduce what the latter is optimizing. The same holds for the second player trying to forecast the behavior of the first. This turns out to be impossible in principle when there is enough ex ante uncertainty about the payoffs.

Theorem 1 (Foster and Young, 2001). *Consider an n-person game on a finite joint action space A, where the n|A|possible payoffs defining G are drawn i.i.d. via a continuous density f that is bounded away from zero on an open interval. G is determined once and for all before play begins. Assume the players are forward-looking and rational, with discount factors less than unity, they know their own realized payoffs, and they use forecasting rules that do not depend on the opponents' realized payoffs.*

There is a positive probability that: i) at least one of the players will not learn to predict even in the weak sense; and ii) the players' period-by-period behaviors do not converge to any Nash equilibrium of the repeated game. Furthermore, if the support of f is a sufficiently small interval, then conclusions i) and ii) hold with probability one.

A consequence of this result is that *there exist no general, model-based procedures for multi-agent learning when players are perfectly rational and they have sufficiently incomplete knowledge of their opponents' payoff functions*.

A crucial condition for theorem 1 to hold is that the unknown payoffs are distributed over some interval. If instead they were known to lie in a finite set, or even in a countable set, the result can fail. In this case one can tailor the forecasting rules to take account of the restricted set of payoffs that the opponent could be using, and thereby satisfy absolute continuity. The second crucial condition for theorem 1 is rationality: agents must optimize *exactly*. If instead agents almost optimize, as in smoothed fictitious play (Fudenberg and Kreps, 1993; Fudenberg and Levine, 1993), the result does not necessarily hold.

In my view the first of these conditions (lack of knowledge) is more important than the second (perfect rationality). For one thing the second condition is merely an ideal statement about behavior, there is little or no support for the notion that subjects optimize *exactly*. By contrast the first condition seems quite realistic: a player can hardly be expected know the von Neumann Morgenstern payoffs of his opponent with any precision; surely the most that can be hoped for is that he knows they lie within some range.

I now sketch a model-based, multi-agent learning method that gets around the preceding impossibility result by relaxing rationality a bit, while maintaining the assumption about complete lack of knowledge. The method is structured along the lines of statistical hypothesis testing. Assume, for the moment, that there are two players, 1 and 2, with finite action spaces A_1 and A_2. Let Δ_i be the simplex of probability distributions on A_i. At time t, agent 1's *model* is that agent 2 is going to play a fixed distribution $p_{2t} \in \Delta_2$ in all future periods. Given this model, agent 1 chooses a smoothed best response $q_{1t} \in \Delta_1$. Similarly, agent 2's model at time t is some $p_{1t} \in \Delta_1$ and her smoothed best response is $q_{2t} \in \Delta_2$. Hypothesis testing takes the following form for each player. Let s be a large positive integer (the *sample size*) and let τ be a small positive real number (the *tolerance level*).

Hypothesis testing

1. Select a model p of the opponent's strategy uniformly at random.
2. Play a smoothed best response q to the current model p.
3. Start a test phase with probability 1/s.
4. Once a test phase begins, compute the opponent's empirical frequency distribution p′ over the next s periods. If $\|p' - p\|$ exceeds τ go to step 1; if the difference does not exceed τ go to step 2.

It can be shown that, given any game G on $A_1 \times A_2$ and any $\varepsilon > 0$, if s is large enough and τ is small enough the players' behaviors in period t constitute an ε-equilibrium of G in at least $1 - \varepsilon$ of all periods. Further, if the players gradually increase s and decrease τ, we obtain the following.

Theorem 2 (Foster and Young, 2003). *Given any n-person game G on a finite action space A, if the hypothesis testing parameters are annealed sufficiently slowly, the players' period-by-period behaviors converge in probability to the set of Nash equilibria of G.*

3. Model-free learning

The second class of learning rules identified by SPG do not rely on prediction of the opponent's behavior, but on some form of heuristic adjustment to previous experience. Unfortunately, in this setting it is not so clear what is meant by "learning." Players are obviously not learning to predict, because they are not predicting. SPG suggest that, in analogy with single-agent Markov decision problems (MDP's), we could say that agents "learn" if their adaptive rules lead to high average payoffs. (In a single-agent context such rules are said to be "effective.")

The difficulty is that, whereas high payoffs are well-defined in single-agent MDP's, they are usually not well-defined in games. In this setting, agents can only optimize given what the other agents are doing. Of course, when everyone optimizes conditional on the others optimizing, the players are in some form of Nash equilibrium, either with respect to the stage game or the repeated game. This suggests one possible definition of "learning" in a model-free environment, namely, that agents' average payoffs converge to the payoffs corresponding to some Nash equilibrium. Alternatively, we might say that they learn if their behaviors come into Nash equilibrium more or less by accident: they act *as if* they were predicting and optimizing, even though they are actually using nonpredictive methods. A third possibility is that Nash equilibrium is not the appropriate solution concept in this setting. We summarize these possibilities as follows.

Learning criteria in model-free environments

I. Payoffs converge to Nash equilibrium payoffs.
II. Behaviors converge to Nash equilibrium.
III. Behaviors and/or payoffs converge to a subset that has some other normative interpretation, even though it may not correspond to a Nash equilibrium.

Let us consider these in turn. The first criterion is very easy to satisfy if by "Nash equilibrium payoffs" we mean payoffs in some repeated-game Nash equilibrium. The reason is that, by the Folk Theorem, virtually all payoff combinations can be realized in a repeated-game equilibrium, provided that each player gets at least his maximin payoff. Many adaptive model-free procedures can achieve the same thing.

The second criterion is much more difficult to achieve. Indeed, until recently it was not known whether there exist *any* model-free learning rules that cause

behaviors to converge to Nash equilibrium, except in special cases. Here is one example of a model-free rule, called *regret testing*, that achieves criterion II for any finite two-person game. Let Δ_d the set of all probability mixtures on the agent's actions that can be expressed in d or fewer decimal places.

Regret testing

1. Choose $q \in Q_d$ uniformly at random.
2. Play q for s periods in succession.
3. For each action a, compute the regret $r(a)$ from not having played action a over these s periods.
4. If $max_a\, r(a) > \tau$, go to step 1; otherwise retain the current q and go to step 2.

Given any two-person game G and any $\varepsilon > 0$, if both players use regret testing with sufficiently large s and d and sufficiently small τ, their behaviors constitute an ε-equilibrium of G in at least $1 - \varepsilon$ of all play periods (Foster and Young, 2006). By annealing the parameters sufficiently slowly one can obtain convergence in probability to the set of Nash equilibria for any finite two-person game G. With some further modifications almost-sure convergence is achievable (Germano and Lugosi, 2007).

A crucial feature of both regret testing and hypothesis testing is the random search that occurs whenever a "test" fails. What happens if we drop this aspect of the learning process? To be quite general, let G be an n-person game with finite action space A. Let s be a positive integer and let the *state* of the learning process be the last s plays of the game. Thus there are $|A|^s$ states. A learning rule for i, say $f_i(z)$, maps each state z to a probability distribution over i's actions next period. The rule f_i is *uncoupled* if it does not depend on the opponents' payoffs (Hart and Mas-Colell, 2006).

Theorem 3 (Hart and Mas-Colell, 2006). *Given a finite action space A and positive integer s, there exist no uncoupled rules $f_i(z)$ whose state variable z is the last s plays, such that, for every game G on A, the period-by-period behaviors converge almost surely to a Nash equilibrium of G, or even to an ε-equilibrium of G, for all sufficiently small $\varepsilon > 0$.*

Note that regret testing is uncoupled, yet we claimed earlier that an annealed version of it converges almost surely to the set of Nash equilibria for any finite two-person game G. This does not contradict theorem 3, however, because in the annealed version the value of s grows, hence the state variable does not consist of histories of bounded length. Furthermore, even in the non-annealed version (where s is fixed), the state variable consists of more than the last s plays of the game; it also includes the realization of a random variable, namely, the new choice of q that occurs whenever someone's regret exceeds his tolerance level τ. Hence it does not satisfy the conditions of theorem 3. (Hypothesis testing does not satisfy the state variable requirement for the same reason.) This illustrates the rather fine line that separates the possible from the impossible in multi-agent learning theory.

We conclude with an example that illustrates why Nash equilibrium is not the only way of evaluating whether agents are "learning" (criterion III). Consider the following simple adaptive procedure first proposed by Hart and Mas-Colell (2000, 2001).

Unconditional regret matching

1. Choose an action uniformly at random.
2. For each action $a \in A$, compute the regret $r(a)$ from not having played a in all previous periods.

3. Among all those actions with positive regrets, choose among them with probabilities proportional to their regrets, then go to step 2; if there are no such actions go to step 1.

When a given player uses this rule in a finite game G, his regrets r(*a*) become nonpositive almost surely no matter what the other players do (Hart and Mas-Colell, 2001). When all players use the rule, the empirical distribution of their joint behaviors converges almost surely to a convex set that contains all of the correlated equilibria of G, and therefore all of the Nash equilibria. (And the average payoffs converge to the set of expected payoffs generated by these distributions.) We shall call this the *coarse correlated equilibrium set* (Young, 2004).[1] It has a simple equilibrium interpretation, namely, it is the set of all joint probability distributions ϕ on A such that each player's expected payoff (under ϕ) is at least as high as his expected payoff if he were to deviate and play an arbitrary action, while the other players adhere to the outcome prescribed by ϕ. (In a correlated equilibrium, by contrast, no player wishes to deviate *after* his prescribed action via ϕ is revealed, which is a more restrictive condition.)

The significance of this solution concept is that it describes average behavior under a wide variety of adaptive rules, including smoothed fictitious play, calibrated forecasting with best responses (Foster and Vohra, 1997), and a number of variants of regret matching (Hart and Mas-Colell, 2000, 2001). (Some of these rules actually converge to the set of *correlated* equilibria.) It seems reasonable to conjecture that this "coarse" notion of correlated equilibrium may prove useful in describing the behavior of experimental subjects, though to my knowledge this point has never been investigated systematically.

---

[1] It was first defined, though not named, in an early paper by Moulin and Vial (1978); Hart and Mas-Colell (2001) call it the *Hannan set*.

I conclude that there is a fine line between the possible and the impossible in multi-agent learning situations. It depends on subtle differences in assumptions about the amount of information that agents have, the extent to which they optimize, the desired form of convergence, and the target set. In the preceding I have identified some prominent landmarks on either side of the dividing line. Its precise course remains to be charted.

References

Binmore, Ken (1987), "Modelling rational players (part I)," *Economics and Philosophy*, 3, 179-214.

Blackwell, David, and Lester Dubins (1962), "Merging of opinions with increasing information," *Annals of Mathematical Statistics*, 38, 882-886.

Foster, Dean P., and Rakesh Vohra (1997), "Calibrated learning and correlated equilibrium," *Games and Economic Behavior*, 21, 40-55.

Foster, Dean P., and Rakesh Vohra (1999), "Regret in the on-line decision problem," *Games and Economic Behavior*, 29, 7-35.

Foster, Dean P., and H. Peyton Young (2001), "On the impossibility of predicting the behavior of rational agents," *Proceedings of the National Academy of Sciences of the USA*, 98, no.222, 12848-12853.

Foster, Dean P., and H. Peyton Young (2003), "Learning, hypothesis testing, and Nash equilibrium," *Games and Economic Behavior*, 45, 73-96.

Foster, Dean P., and H. Peyton Young (2006), "Regret testing: learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, 1, 341-367.

Fudenberg, Drew, and David Kreps (1993), "Learning mixed equilibria," *Games and Economic Behavior*, 5, 320-367.

Germano, Fabrizio, and Gabor Lugosi (2007), "Global convergence of Foster and Young's regret testing," *Games and Economic Behavior*, forthcoming.

Hart, Sergiu, and Andreu Mas-Colell (2001), "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, 68, 1127-1150.

Hart, Sergiu, and Andreu Mas-Colell (2001), "A general class of adaptive strategies," *Journal of Economic Theory*, 98, 26-54.

Hart, Sergiu, and Andreu Mas-Colell (2006), "Stochastic uncoupled dynamics and Nash equilibrium," *Games and Economic Behavior*, 57, 286-303.

Jordan, James S. (1991), "Bayesian learning in normal form games," *Games and Economic Behavior*, 3, 60-91.

Jordan, James S. (1993),"Three problems in learning mixed-strategy equilibria," *Games and Economic Behavior*, 5 (1993), 368-386.

Kalai, Ehud, and Ehud Lehrer, Rational learning leads to Nash equilibrium, *Econometrica*, 61, 1019-1045.

Moulin, Herve, and J. P.Vial (1978), "Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon," *International Journal of Game Theory*, 7, 201-221.

Shoham, Yoav, Rob Powers, and Trond Grenager (2006), "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*.

Young, H. Peyton, *Strategic Learning and Its Limits*, Oxford: Oxford University Press, 2004.