LEARNING TO PLAY GAMES IN EXTENSIVE FORM BY VALUATION

PHILIPPE JEHIEL AND DOV SAMET

Extended abstract

Models of learning in games fall roughly into two categories. In the first, the learning player forms beliefs about the future behavior of other players and nature, and directs her behavior according to these beliefs. We refer to these as fictitious-player-like models. In the second, the player is attuned only to her own performance in the game, and uses it to improve future performance. These are called models of reinforcement learning.

Reinforcement learning has been used extensively in artificial intelligence (AI). Samuel wrote a checkers-playing learning program as far back as 1955, which marks the beginning of reinforcement learning (see Samuel (1959)). Since then many other sophisticated algorithms, heuristics, and computer programs, have been developed, which are based on reinforcement learning. (Sutton and Barto (1998)). Such programs try neither to learn the behavior of a specific opponent, nor to find the distribution of opponents' behavior in the population. Instead, they learn how to improve their play from the achievements of past behavior.

Until recently, game theorists studied mostly fictitious-player-like models. Reinforcement learning has only attracted the attention of game theorists in the last decade in theoretical works like Gilboa and schmeidler (1995), Camerer and Ho (1997), Sarin and Vahid (1999), and in experimental works like Erev and Roth (1997). In all these studies the basic model is given in a strategic form, and the learning player identifies those of her strategies that perform better. This approach seems inadequate where learning of games in extensive form is concerned. Except for the simplest games in extensive form, the size of the strategy space is so large that learning, by human beings or even machines, cannot involve the set of all strategies. This is certainly true for the game of chess, where the number of strategies exceeds the number of particles in the universe. But even a simple game like tic-tac-toe is not perceived by human players in the full extent of its strategic form.

The process of learning games in extensive form can involve only a relatively small number of simple strategies. But when the strategic form is the basic model, no subset of strategies can be singled out. Thus, for games in extensive form the structure of the game tree should be taken into consideration. Instead of *strategies* being reinforced, as for games in strategic form, it is the *moves* of the game that should be reinforced for games in extensive form.

This, indeed, is the approach of heuristics for playing games which were developed by AI theorists.[1] One of the most common building block of such heuristics is the *valuation*, which is a real valued function on the possible moves of the learning player. The valuation of a move reflects, very roughly, the desirability of the move. Given a valuation, a learning process can be defined by specifying two rules:

- A *strategy rule*, which specifies how the game is played for any given valuation of the player;
- A *revision rule*, which specifies how the valuation is revised after playing the game.

Our purpose here is to study learning-by-valuation processes, based on simple strategy and revision rules. In particular, we want to demonstrate the convergence properties of these processes in repeated games, where the stage game is given in an extensive form with perfect information and any number of players. Converging results of the type we prove here are very common in the literature of game theory. But as noted before, convergence of reinforcement is limited in this literature to strategies rather than moves.[2] To the best of our knowledge, the AI literature while describing dynamic processes closely related to the ones we study here do not prove convergence results of this type.

[1] Perhaps the concentration of the AI literature on moves rather than strategies is the reason why there seems to be almost no overlap between two major books on learning, each in its field: *The Theory of Learning in Games*, Fudenberg and Levine (1998) and *Reinforcement Learning: An Introduction*, Sutton and Barto (1998).

[2] There is no obvious way to define an assessment for a strategy from a system of node valuations. Therefore, a simple translation of our learning model in terms of strategies is not straightforward. One fundamental difficulty is that the node valuation treatment does not impose that a strategy be assessed in the same way throughout the play of the game. Also, two strategies involving the same first move should be assessed in the same way initially (a condition which does not make much sense in the reinforcement learning based on the strategic form.

First, we study stage games in which the learning player has only two payoffs, 1 (win) and 0 (lose). Two-person win-lose games are a special case. But here, there is no restriction on the number of the other players or their payoffs.

For these games we adopt the simple *myopic strategy rule.* By this rule, the player chooses in each of her decision node a move which has the highest valuation among the moves available to her at this node. In case there are several moves with the highest valuation, she chooses one of them at random.

As a revision rule we adopt the simple *memoryless revision*: after each round the player revises only the valuation of the moves made in the round. The valuation of such a move becomes the payoff (0 or 1) in that round.

Equipped with these rules, and an initial valuation, the player can play a repeated game. In each round she plays according to the myopic strategy, using the current valuation, and at the end of the round she revises her valuation according to the memoryless revision.

This learning process, together with the strategies of the other players in the repeated game, induce a probability distribution over the infinite histories of the repeated game. We show the following, with respect to this probability.

> Suppose that the learning player can guarantee a win in the stage game. If she plays according to the myopic strategy and the memoryless revision rules, then starting with any nonnegative valuation, there exists, with probability 1, a time after which the player always wins.

When the learning player has more than two payoffs, the previous learning process is of no help. In this case we study the *exploratory myopic strategy rule*, by which the player opts for the maximally valued move, but chooses also, with small probability, moves that do not maximize the valuation.

The introduction of such perturbations makes it necessary to strengthen the revision rule. We consider the *averaging revision.* Like the memoryless revision, the player revises only the valuation of moves made in the last round. The valuation of such a move is the average of the payoffs in all previous rounds in which this move was made.

> If the learning player obeys the exploratory myopic strategy and the averaging revision rules, then starting with any valuation, there exists, with probability 1, a time

> after which the player's payoff is close to her individually rational payoff (the maxmin payoff) in the stage game.

The two previous results indicate that reinforcement learning achieves learning of playing the stage game itself, rather than playing against certain opponents. The learning processes described guarantee the player her individually rational payoff (which is the win in the first result). This is exactly the payoff that she can guarantee even when the other players are disregarded.

Our next result concerns the case where all the players learn the stage game. By the previous result we know that each can guarantee his individually rational payoff. But, it turns out that the synergy of the learning processes yields the players more than just learning the stage game. Indeed, they learn in this case each other's behavior and act rationally on this information.

> Suppose the stage game has a unique perfect equilibrium. If all the players employ the exploratory myopic strategy and the averaging revision rules, then starting with any valuation, with probability 1, there is a time after which their strategy in the stage game is close to the perfect equilibrium.

Although valuation is defined for all moves, the learning player needs no information concerning the game when she start playing it. Indeed, the initial valuation can be constant. To play the stage game with this valuation, the player needs to know which moves are possible to her, only when it is her turn to play, and then choose one of them at random. During the repeated game, the player should be able to record the moves she made and their valuations. Still, the learning procedure does not require that the player knows how many players there are, let alone the moves they can make and their payoffs.

The learning processes discussed here treat separately the valuation for every node. For games with large number of nodes (or states of the board), that may be unrealistic because the chance of meeting a given node several times is too small. In chess, for example, almost any state of the board, except for the few first ones, has been seen in recorded history only once. In order to make these processes more practical, similar moves (or states of the board) should be grouped together, such that the number of similarity classes is manageable. When the valuation of a move is revised, so are all the moves similar to it. We will deal with such learning processes, as well as with games with incomplete information, in a later paper.

References

Camerer, C. and T. Ho (1997). Experience-Weighted Attraction Learning in Games: A Unifying Approach, *Econometrica.*

Erev, I. and A. Roth (1997). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibrium. *American Economic Rev.*

Fudenberg, D. and D. Levine (1998). *The Theory of Learning in Games*, The MIT Press.

Gilboa, I. and D. Schmeidler (1999). Case Base Decision Theory. *Quart. J. Econom*, 110, pp. 605–639.

Loève, M. (1963). *Probability Theory*, D. Van Nostrand, Third ed.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers, *IBM J. Res. and Devel.*, 3, p. 210—229.

Sarin, R. and F. Vahid (1999). Payoff Assessments without Probabilities: A Simple Dynamic Model of Choice. *Games and Economic Behaviour*, 28, pp. 294–309.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*, The MIT Press.