Learning TOMs: Towards Non-Myopic Equilibria

Arjita Ghosh*, **Sandip Sen**
Department of Mathematics & Computer Science
University of Tulsa
arjita-ghosh@utulsa.edu, sandip@utulsa.edu

Abstract

In contrast to classical game theoretic analysis of simultaneous and sequential play in bimatrix games, Steven Brams has proposed an alternative framework called the Theory of Moves (TOM) where players can choose their initial actions and then, in alternating turns, decide to shift or not from its current action. A backward induction process is used to determine a non-myopic action and equilibrium is reached when an agent, on its turn to move, decides to not change its current action. Brams claims that the TOM framework captures the dynamics of a wide range of real-life non-cooperative negotiations ranging over political, historical, and religious disputes. We believe that his analysis is weakened by the assumption that a player has perfect knowledge of the opponent's payoff. We present a learning approach by which TOM players can learn to converge to Non-Myopic Equilibria (NME) without prior knowledge of its opponent's preferences and by inducing them from past choices made by the opponent. We present experimental results from all structurally distinct 2-by-2 games without a common preferred outcome showing convergence of our proposed learning player to NMEs. We also discuss the relation between equilibriums in sequential games and NMEs of TOM.

Introduction

Learning and reasoning in single or multistage games have been an active area of research in multiagent systems (Banerjee, Sen, & Peng 2001; Bowling & Veloso 2001; Claus & Boutilier 1998; Hu & Wellman 1998; Littman 1994; 2001; Littman & Stone 2001). Most of this research has concentrated on simultaneous move games with solution concepts like Nash equilibria (Myerson 1991; Nash 1951). Though dynamic or extensive-form games have also received attention in game theory, Brams have argued for an alternative framework called Theory of Moves (TOM) and its logical extension to anticipation games.

In the TOM framework, agents have knowledge of the starting state of the game and make sequential and alternate, rather than simultaneous moves. The basic TOM formulation involves a complete information 2x2 bimatrix game where players have perfect information about both payoff matrices and know the starting strategy profile. This means that the play starts at a certain state from which the two players can move in alternate turn. The players decide their move based not only the feedback they will receive if they change their current strategy and hence move to a new state, but also on the possible counter-move of the opponent, its own counter-move to the opponent's counter-move, and so on. With two moves per player, a cycle can result in at most four moves. The rules followed by a TOM player are presented in Theory of Moves Framework Section. The basic idea is that both players make moves projecting sufficiently ahead into the future but assuming that cycles should be avoided. From the starting state both players are asked if they want to move. As we are dealing with the basic TOM framework of 2x2 games, i.e., each player has two actions (pure strategies or strategies, in short), moving corresponds to changing the current strategy and not moving corresponds to continue using the current strategy. To make this decision, the player looks three moves ahead and uses backward induction to decide whether moving will be beneficial or not. If both players decide not to move, the starting state is the equilibrium. If only one player decides to move, the state changes and it is the other player's turn to move who will use a two-move lookahead to decide its move, and so on. The resulting state where a player decides not to move is an equilibrium. If both players decide to move, we have an indeterminate outcome which can produce two different equilibrium states depending on which player moves first from a particular starting state. These equilibria are called non-myopic equilibria (NME) as player uses look-ahead to select equilibrium states.

It is worth noting that with perfect information and both players following TOM rules, it is not actually necessary to carry out the actual play. Given a starting state, each player calculates the equilibrium state that would result if it was to move first. If the two states are equal or if one of the players decide not to move, then there is a unique (NME) given the starting state. If, however, both players decide to move and their respective first move will result in different equilibrium states, we have an indeterminate situation with multiple NMEs given the initial state. The players can calculate the different NMEs resulting from each of the four initial states.

The respective payoffs to the players from the NMEs

*Primary author is a student.

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

given each of these states can then be used as the payoff matrix of an anticipatory game (AG). The Nash equilibrium of this AG, can be used to choose the starting state of the play in TOM. In this paper, we do not concern ourselves with how the initial state is chosen. Rather, we focus on learning to converge to NMEs given any starting state and without the knowledge of the opponent's payoff matrix or preference over states.

It should also be noted that since TOM assumes alternate moves, only relative values of the payoffs are required and not absolute values. To see why this is the case, note that at every stage of the game, an agent is deciding whether to move or not, i.e., choosing between two adjacent cells in the payoff matrix. Thus a total preference order over the states is sufficient to make this decision. As such we can work with only structurally distinct ordinal games. In a later section we outline how to construct this set of games.

Brams argues convincingly that TOM play can be used to model a wide range of real-life scenarios (Brams 1994). While we agree with most of his arguments, we believe that a serious weakness in the basic TOM framework is the assumption of complete information. Brams discusses the issue of incomplete information in large games incorporating voting among players, but does not give a complete treatment that this topic deserves. We believe that in real-life scenarios, a player is unlikely to have access to the payoff of the other player, and has to make its decisions based on only its own payoff matrix and that of the observed decisions of the other player in past scenarios and it also may not able to negotiate with others. This motivates us to develop a learning approach that can approximate the preferences of the opponent, and using that, decide on own actions that will consistently produce outcomes identical to TOM players with complete information.

Structurally distinct 2x2 ordinal games

In the current paper, we will only consider a subset of the possible 2x2 payoff matrices where agents have a total preference order over the four possible states. We will use the numbers 1, 2, 3, 4, as the preference of an agent for a state in the 2x2 matrix, with 4 being the most preferred. The following discussion allows us to count the number of structurally distinct matrices. For a more elaborate discussion see (Rapoport & Guyer 1966).

Agent A's lowest payoff can be combined with four possible payoffs for agent B. For each such combination, there are three payoffs to agent B that can be combined with the next-to-lowest payoff, and two payoffs to be combined with agent A's second most-preferred payoff, and the remaining one to be paired with agent A's highest payoff. This results in 4!=24 sets of four pairs of numbers. To generate a bimatrix game, we have to distribute a given set of four pairs over the four states of the matrix. This can be done in 3!=6 ways: put the first pair in one corner, one of the remaining three in the opposite, and then there will be two ways of placing the last two pairs. Though this results in $24 * 6 = 144$ possible games, they are not all distinct. Some pairs of these matrices are identical if the row and column players are renamed in one of the pair, i.e., the payoff matrices in one are transposes of the payoff matrices in the other. However, the payoff matrices where the off-main-diagonal payoffs to the two players are identical, do not have corresponding matching matrices. There are 12 such matrix pairs. Hence the total number of distinct 2x2 ordinal games is $\frac{144}{2} + \frac{12}{2} = 78$. Out of these, there are 57 games in which there are no mutually preferred outcomes. These are often referred to as nonconflicting games. Brams lists all of these games, together with their NMEs (Brams 1994) and the matrix numbers used in this paper correspond to those used by Brams. Of these 57 games, 31 have a single NME, 24 have two NMEs, and only 2 have three NMEs.

Theory of Moves Framework

In the TOM framework, players alternate in making moves and think ahead not just to the immediate consequences of making moves but also to the consequences of countermoves to these moves, counter-counter-moves, and so on. TOM extends strategic thinking into the more distant future than most other dynamic theories do. To incorporate this concept TOM has specific rules. The rules of play of TOM for two-person games, which describe the possible choices of the players at each stage of play, are as follows (Brams 1994):

1. Play starts at an outcome, called the *initial state*, which is at the intersection of the row and column of a 2×2 payoff matrix.
2. Either player can unilaterally switch its strategy, and thereby change the initial state into a new state, in the same row or column as the initial state. The player who switches is called player 1 (P1).
3. Player 2 (P2) can respond by unilaterally switching its strategy, hereby moving the game to a new state.
4. The alternating responses continue until the player (P1 or P2) whose turn it is to move next chooses not to switch its strategy. When this happens, the game terminates in a *final state*, which is the *outcome* of the game.
5. A player will not move from an initial state if this move
   - leads to a less preferred final state (i.e., outcome); or
   - returns play to the initial state (i.e., makes the initial state the outcome).
6. Given that players have complete information about each other's preferences and act according to the rules of TOM, each takes into account the consequences of the other player's rational choices, as well as its own, in deciding whether to move from the initial state or later, based on backward induction. If it is rational for one player to move and the other player not to move from the initial state, then the player who moves takes precedence: its move overrides the player who stays, so the outcome will be induced by the player who moves.

Let's take the pay-off matrix as follows:

$$\begin{array}{llll} \text{Matrix 13 :} & & Cplayer & \\ & & c_1 & c_2 \\ Rplayer & r_1 & (3,4) & (4,1) \\ & r_2 & (1,2) & (2,3) \end{array}$$

According to TOM, play may begin at any state and any one of the two players can start the game. To explain the working of TOM we assume (a) it is a complete-information game and (b) each player knows that the other player plays according to TOM.

Initial State: (3,4)

- Suppose R moves first. The counter-clockwise progression from (3,4) back to (3,4) is as follows:

$$\begin{array}{cccccc} & R & C & R & C & R \\ & (3,4) \rightarrow\mid\text{c} & (1,2) \rightarrow & (2,3) \rightarrow & (4,1) \rightarrow & (3,4) \\ \mathcal{S} & (3,4) & (3,4) & (3,4) & (3,4) & \end{array}$$

  We will illustrate the backward induction process for this case only and use it without further explanation in following cases. S denotes the state by which R or C can reach by estimating backward induction. R looks ahead 4 states and finds that C will move from state (4,1) to (3,4) as it gains more by doing so. Following backward induction, R reasons that if it is put in state (2,3) it can expect to get (3,4) by moving (as C will also move from (4,1)). Following the same reasoning, R believes C will move from state (1,2). Hence it concludes that if it were to move from state (3,4) the play will cycle. Therefore, R will stay at (3,4) according to rule 5. This special type of blockage is indicated by "|c" (for cycling) following the arrow.

- Suppose C moves first. The clockwise progression from (3,4) back to (3,4) is as follows:

$$\begin{array}{cccccc} & C & R & C & R & C \\ & (3,4) \rightarrow\mid & (4,1) \rightarrow\mid & (2,3) \rightarrow & (1,2) \rightarrow & (3,4) \\ \mathcal{S} & (3,4) & (4,1) & (3,4) & (3,4) & \end{array}$$

  If C starts there is blockage at the start. The usual blockage is indicated by "|" following the arrow.That means, if C moves from this state it will get lesser payoff and for this C prefers to stay at initial state.

So, if the game starts from the state (3,4) none of the players are interested in moving and the outcome (3,4) is an NME.

Initial State: (4,1).

- Suppose R moves first. The clockwise progression from (4,1) back to (4,1) is as follows:

$$\begin{array}{cccccc} & R & C & R & C & R \\ & (4,1) \rightarrow\mid & (2,3) \rightarrow & (1,2) \rightarrow & (3,4) \rightarrow\mid & (4,1) \\ \mathcal{S} & (4,1) & (3,4) & (3,4) & (3,4) & \end{array}$$

  There is blockage at first. So R prefers to stay at initial state.

- Suppose C moves first. The counter-clockwise progression from (4,1) back to (4,1) is as follows:

$$\begin{array}{cccccc} & C & R & C & R & C \\ & (4,1) \rightarrow & (3,4) \rightarrow\mid & (1,2) \rightarrow\mid & (2,3) \rightarrow & (4,1) \\ \mathcal{S} & (3,4) & (3,4) & (1,2) & (4,1) & \end{array}$$

  According to TOM rule, C wants to go to state (3,4) and hence it prefers to move.

So, if play starts at state (4,1), there is a conflict: R wants to stay but C wants to move. But because C's move takes precedence over R's staying, the outcome is that which C can induce, namely, (3,4), which is the NME.

Following the procedures as described above, it can be observed that if game starts at state (2,3), both player will prefer to move and state (3,4) will be achieved as terminal. So, the NME is (3,4). Similarly, if game starts at state (1,2), both player will again prefer to move and hence the induced state will be (3,4) making the state NME again. So, this payoff matrix has only one Non-Myopic Equilibrium (NME) at state (3,4).

All the outcomes shown above are derived from a complete-information game. But, in real-life problems it is more likely that a player only knows its own payoff and not that of the opponent. The basic TOM methodology cannot be applied in this situation. We will address this important and highly relevant problem by using a learning approach.

Learning TOM players

We consider the case where two players, without knowledge of opponent payoffs, are playing according to the TOM framework. The goal of our learning process is to infer opponent preferences from repeated play of the game starting at randomly selected states. By inducing minimally sufficient knowledge, the learning TOM players should be able to converge to equilibrium from arbitrary starting states.

To facilitate learning, we approximate conditional probability of the opponent moving from a state given the starting state of play and the player who is to make the first move. Conditional probability is essential here because opponent's movement from a fixed state may vary depending upon how far the game will continue. We use uniform priors as starting points, i.e., all probabilities are initialized to 0.5 in 2x2 games. The states of games have been considered in following order:

$$\begin{array}{cccc} & & Cplayer & \\ & & c_1 & c_2 \\ Rplayer & r_1 & S0 & S1 \\ & r_2 & S3 & S2 \end{array}$$

The algorithm that we have used is described as follows:

Using Probability: A player calculates its probability of changing the state by taking the product of conditional probability of the states that will come next in the sequence of play, e.g.,

$$\begin{array}{ccccc} C & R & C & R & C \\ S0 \rightarrow & S1 \rightarrow & S2 \rightarrow & S3 \rightarrow & S0 \\ & P_0^{0,c} & P_1^{0,c} & P_2^{0,c} & \end{array}$$

$P_0^{0,c}$, $P_1^{0,c}$ and $P_2^{0,c}$ are the conditional probabilities of moving for player R at state S1, player C at state S2 and player R at state S3 respectively given that the starting state was S0 and C was to make the first move (for brevity of presentation, we will drop the superscript where the starting state and the first player is specified). To make its move decision, player C will look up P_0 and P_2 for player R from past frequencies of moves and non-moves by player R from the respective states given the starting state S0. And depending on P_2 player C calculates its own conditional probability P_1. In the following we assume that for $\forall i, Q_i = 1 - P_i$ and $U_x(y)$ is the payoff received by player x in state y.

Probability Calculation: The goal is to calculate the probability of moving by C at state S0, $P_C(S0)$. But first, in the process of backward induction, it has to calculate its probability of moving at state S2, $P_C(S2)$ (which is the same as P_1) as follows:

$P_C(S2) \leftarrow 0$
If $U_C(SO) > U_C(S2)$ then $P_C(S2) += P_2$. {C benefits if R moves from S3 to S0, play results in cycle}
If $U_C(S3) > U_C(S2)$ then $P_C(S2) += Q_2$.{C benefits if R stays at S3, play stops at S3}

As TOM does not allow cycle, and $P_0 \times P_1 \times P_2$ is the probability of creating a cycle, the probability of not changing should be at least this value. The process for calculating $P_C(S0)$ is then as follows:

$P_C(S0) \leftarrow 0$
If ($U_C(S3) > U_C(S0)$) then $P_C(S0) += P_0 \times P_1 \times Q_2$ {C benefits if play stops at S3.}
If ($U_C(S2) > U_C(S0)$) then $P_C(S0) += P_0 \times Q_1$. {C benefits if play stops at S2.}
If ($U_C(S1) > U_C(S0)$) then $P_C(S0) += Q_0$. {C benefits if play stops at S1.}

Making the decision: After these probabilities are calculated, play proceeds by every player throwing a biased coin with the calculated probability to make a move from its current state. An iteration stops when a player does not move or if a cycle is created. Cycles can be created initially because of uninformed priors. Note that if C decides to move from S0, R has to calculate $P_R(S1)$ based on its estimates of $P_1^{0,c}$, and its decision to move or not at S3, which it can calculate in a straightforward manner. Also, if R decides to move at S1, then C can reuse its previous calculation of P_1 to choose its move at S2.

Convergence to NMEs: Over time the backward induction procedure combined with the above decision mechanism, will eliminate cycles. To see why this will happen, notice that, in the above scenario, R's decision at state S3 is deterministic. i.e., it changes if and only if U_R(S0)$>$ U_R(S3). Initially, C is uncertain of this decision and assumes P_2 is 0.5. Once it repeatedly observes R's decision at state S3, C's estimation of P_2 will converge to 1 or 0. This, in turn will lead to P_1 converging to 1 or 0 depending on the utilities C receive at states S2, S3, and S0. The updates are reflected in the subsequent plays by each player, which in turn, allows the other player to get an accurate estimate of their relative preferences. Since the learning is based on backward induction, accurate deterministic choices are used to update less accurate estimates and so on. Over time and repeated plays, the probabilities will become small or large enough to produce almost deterministic actions reflecting actual preferences. As a result, the players will converge on NMEs.

Theorem: *Learning TOM players converge to NME.*

Proof: Without lack of generality we consider the case of player C starting play from state $S0$ (all other starting state, player combinations can be handled in an analogous manner). We can start our proof in one of two cases: i) when an agent reaches a state with its maximum payoff, 4, and ii) when an agent reaches to the state just one step behind the terminal state (here, S_3). In all such cases, agent takes its decision deterministically. Let's study the case where S_3 is reached. The deterministic decision by a player P while considering move from state Si to Sj is defined as δ_{ij}^P = 1, **if** $U_C(S_i) > U_C(S_j)$; **else** 0. Also, the observed ratio of the number of times a player P moving from state Si to the number of times it made a decision in that state is designated r_i^P. Hence, $P_C(S_2) = \delta_{02}^C r_{S_3}^R + \delta_{32}^C(1 - r_{S_3}^R)$. This value will remain constant since at this state R's behavior is deterministic. Consequently, C'behavior is now deterministic and $r_{S_2}^C$ will tend to 0 or 1. Likewise, we can also calculate

$$P_R(S_1) = \delta_{01}^R \underbrace{r_{S_3}^R}_{=0\ or\ 1} r_{S_2}^C + \delta_{31}^R(1 - r_{S_3}^R) \underbrace{r_{S_2}^C}_{\to 0\ or\ 1} + \delta_{21}^R(1 - r_{S_2}^C)$$

As the δ terms are 0 or 1 and the ratios converge to 0 or 1, $P_R(S_1)$ will converge to a deterministic choice. Eventually, then all the three probabilities, $P_R(S_3)$, $P_C(S_2)$ and $P_R(S_1)$ will have values of 0 or 1. As the probability calculations follow the same backward induction used in TOM, when the probabilities converge, the agent decisions will coincide with the moves made by TOM players with complete information. Hence, learning TOM players will converge to NMEs under TOM play rules.

Now, we will illustrate the working of the above procedure using the example in the Theory of Moves Framework Section (in the following, we assume a player can only observe its own payoffs even though we show both payoffs for helping the discussion):

Initial State, S0, R to move: (3,_)[1].

The counter-clockwise progression from (3,_) back to (3,_) is as follows:

$$\begin{array}{ccccccccc} R & & C & & R & & C & & R \\ (3,4) & \rightarrow & (1,2) & \rightarrow & (2,3) & \rightarrow & (4,1) & \rightarrow & (3,4) \\ S_0 & & S_3 & & S_2 & & S_1 & & S_0 \\ & & P_0 & & P_1 & & P_2 & & \end{array}$$

From our previous discussion we can calculate the probability that R will move, P_R(3,_) = $P_0 \times P_1 \times Q_2$, and that of its not moving to be $P_0 \times P_1 \times P_2 + P_0 \times Q_1 + Q_0$.

Following our procedure, R can calculate $P_1 = P_2 + Q_2$ (as payoff at S0 and at S3 both are larger than that at S2) = 1.0. Hence, P_R(3,_) is calculated to be 0.25 as initial estimates of P_0 and P_2 are 0.5. Now, player R may move or not based on the biased coin toss with this probability. If R does not move, the starting state is the outcome of this iteration which is also the outcome obtained if the TOM player had complete information. But if the biased coin toss results in a move at this state, the following scenario unfolds:

- C player will play from S3 and will look ahead two moves. It assumes P_1=0.5 and calculates P_2 to be 1.0 (based on its preference of (_,4) over (_,1)) and P_C(_,2)

[1] We use an _ to signify unknown payoff of the opponent.

$= P_1 \times P_2 + Q_1 = 0.5 + 0.5 = 1.0$. As C will move at this state, play will continue.

- R player will play from S2 and will look ahead 1 move. Here, $P_R(2,\_) = P_2 + Q_2 = 1.0$. So, R will move and play will continue.
- C will move from (_,1) to (_,4) by a deterministic choice.

This results in a cycle and stops the iteration. The cycle is a violation of TOM rules. But this iteration allows R to update P_2 to 1. If the same starting state and player is repeated, $P_R(3,\_) = P_0 \times P_1 \times Q_2$ becomes 0. As a result, player R will not move from the starting state. Thus if (3,4) is chosen to be the starting state, the outcome is consistent with complete-information TOM play.

By analyzing the move of C from this state we can show that it will not move from (_,4) and hence if the play starts at (3,4) the outcome is an NME with learning TOMs.

Initial State, S1, C to move: (_,1).

Let's consider another situation. Suppose C moves first. The clockwise progression from (_,1) back to (_,1) is as follows:

$$\begin{array}{ccccc} C & R & C & R & C \\ (4,1) \rightarrow & (3,4) \rightarrow & (1,2) \rightarrow & (2,3) \rightarrow & (4,1) \\ S_1 & S_0 & S_3 & S_2 & S_1 \\ & P_0 & P_1 & P_2 & \end{array}$$

- Player C will play from (_,1): $P_C(\_,1) = P_0 \times P_1 \times Q_2 + P_0 \times Q_1 + Q_0 = 0.125 + 0.25 + 0.5 = 0.875$, where $P_1 = Q_2 = 0.5$. If the biased coin toss results in no move, the outcome is not consistent with TOM player under complete information. If C moves, play continues as follows.
- Player R will play from state (3,_): $P_R(3,\_) = P_1 \times P_2 = 0.5$, where $P_2 = 1.0$ (chosen deterministically). If R does not move output is consistent with complete-information TOM play, but if R moves result will be inconsistent and play continues as follows.
- Player C will play from state (_,2): $P_C(\_,2) = Q_2 = 0.5$. Now, if it does not move the output will be (1,2) which is erroneous. Moreover, R will have an erroneous estimate of P_1. But if C moves, play continues as follows.
- Player R will play from (2,_) to (4,_). This will change P_2 to 1, which results in a reduction of $P_R(3,\_)$. Over time then R will not move from (3,4) resulting in an outcome consistent with perfect-information TOM.

Comparison of NMEs with other equilibrium concepts

As TOM proceeds sequentially, it is instructive to compare this framework with the concept of dynamic games from classical game theory. We start the discussion by reviewing a popular equilibrium concept Nash equilibrium (NE) for simultaneous-move games, which is defined as follows: A *Nash equilibrium* (Nash 1951) is an outcome from which neither player would unilaterally depart because it would do worse, or at least not better, if it did. For example, (r_1, c_1) is a Nash equilibrium in the following matrix with payoffs of

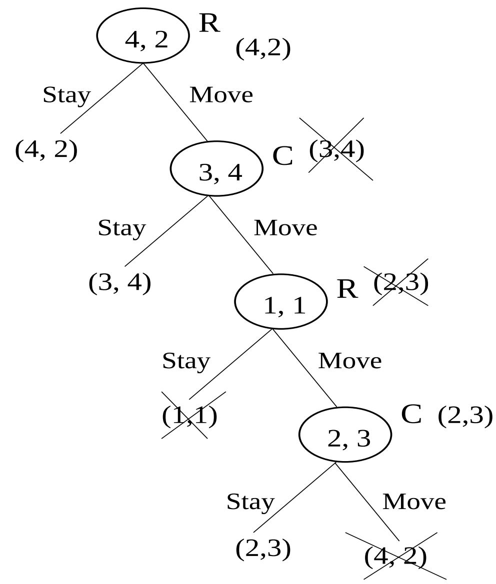

Figure 1: Backward Induction in TOM: R player wants to stay.

2 and 3 to the row and column player respectively:

Matrix48		*Cplayer*	
		c_1	c_2
Rplayer	r_1	$(2,3)$	$(4,2)$
	r_2	$(1,1)$	$(3,4)$

But NE is calculated on the basis of immediate payoff. It is instructive to evaluate the hypothesis whether it is beneficial for a player to depart from an NE strategy when considering not just immediate payoff but also those received from future moves and countermoves. TOM adopts this line of reasoning and may achieve different equilibrium states.

Dynamic games are the form of games studied in classical game theory that has somewhat similar motivations to TOM. A dynamic game consists of alternating moves by the players where the starting player and the depth of the game tree is pre-determined. Along with payoffs of players, dynamic games provide the sequence of play. And as there is a sequence, all actions are not credible. The equilibrium concept in dynamic games is that of subgame perfect Nash equilibrium (SPNE), which can be calculated by backward induction on the game tree. Game theory states that any dynamic game can be represented by a corresponding simultaneous-move game. Any SPNE in a dynamic game corresponds to a NE of the corresponding simultaneous move game, but not vice versa.

The common aspect for calculating equilibria in TOM and dynamic game is the backward induction process. The figure 1 shows the backward induction process used by TOM on matrix 48 considering (4,2) as starting state and R player as starting player. From the figure we can see that R player will decide to stay at the current state.

There are, however, fundamental differences between dynamic games and TOM play. The first difference is in the representation of the game trees. In contrast to TOM play where play commences from a state in the game matrix, i.e., the players have already chosen a strategy profile[2], there is no concept of a physical state at the root of a game tree corresponding to a dynamic game. The starting player in a dynamic game chooses from one of its possible strategies. For each such strategy choice, the other player can respond with one of its strategies and so on. Payoffs to the players at a leaf of the game tree are based on the combination of strategies played by the players from the root to the corresponding leaf. So, the state information in dynamic games is captured by the path from root to a leaf of the tree and not at each tree node.

To further illustrate the difference, consider the 2-by-2 payoff matrix of the simultaneous-move equivalent of a dynamic-form game where each player has only two strategies. For this one dynamic game, there are eight different TOM game trees depending on which of the two players make the first move and which of the four states is chosen as the starting state. As a result there can be more equilibria, i.e., NME, for this matrix when using TOM than there are SPNEs. Besides this, according to TOM rule, given a starting state, if one player decides to move (based on backward induction on a game tree where it is the starting player) and the other does not (based on a game tree with the same initial state but where this second player was the starting player), then the game proceeds according to the first game tree. Hence TOM framework provides a different set of equilibria, known as NMEs that may or may not contain the NEs of the corresponding game matrix. Usually, the number of NMEs are more than that of NEs because, here for calculating NMEs each of the combination of starting state and starting player has been considered. In case of matrix 48, there are two NMEs: (4,2) and (3,4); none of them are NE. So, we can say that TOM is alternative approach of standard game theory.

We emphasize, however, that these difference in the set of equilibria in TOM play and for dynamic games for the same payoff matrix stems from TOM assuming a starting state from which players moves, which is not the case with dynamic games. In particular, TOM play does not violate any basic rationality premise. More specifically, the backward induction used in TOM is functionally identical to the backward induction process used in dynamic games to calculate SPNEs. In this sense, the two equilibrium concepts are identical.

Experimental Results

We have run experiments with all 57 non-conflicting, structurally distinct 2x2 games. For each game, we run several epochs, where each epoch consists of play starting from each of the 4 states, and each player getting the first move from a state. In one iteration, one player gets the choice to make the first move starting from a given state. Play continues until one player chooses not to move or if a cycle is created. Probabilities are updated and actions are chosen as per our procedure outlined in the Learning TOM players Section. We plot how the terminal state has been achieved from a particular state for a particular play. We observe that over the iterations the correct terminal states from each of 4 states have been reached in all 57 matrices. Hence, we can say that our learning TOM players accurately converge to the NMEs of the corresponding game without prior knowledge of the opponent's preferences or payoffs.

As an example, Figure 2 depicts the result of an experiment with Matrix 13 having state 1 as starting state. In this figure and the following, Rtom and Ctom (Rlearn and Clearn) corresponds to states reached when the row and column TOM (learning TOM) player moves first from the starting state. Here, the learning curve of R player has quickly converged to the equilibrium state chosen by TOM players with complete information, whereas the C player took more time to reach that state. Figure 3 depicts the result on a matrix corresponding to the well-known Prisoners' Dilemma problem:

$$\begin{array}{llll} \mathbf{Prisoners'Dilemma}: & & Cplayer & \\ & & c_1 & c_2 \\ Rplayer & r_1 & (2,2) & (4,1) \\ & r_2 & (1,4) & (3,3) \end{array}$$

There are two NMEs in this matrix: S0 and S2. In this figure, the terminal state obtained from state 1 considering R and C player respectively as the starting players, is the state S2. Although learning TOM players choose non-terminal states S1 or S0 in the first few iterations, the desired state has been learned over time.

Similar results are observed on all four states of the remaining matrices. So, we conclude that our learning approach results in consistent convergence to NMEs in TOM play with limited information.

Conclusions

In this paper, we have presented a learning approach to TOM play in 2x2 games which does not require the knowledge of the opponent's preferences or payoffs. As TOM results in alternating plays, moves taken by opponents can be used to approximate their preferences, and this in turn can lead to a decision procedure, the outcome of which is consistent with complete-information TOM play. As it is unlikely for an agent to be able to observe the opponent's payoffs in most real-life situations, our learning approach to TOM play that does not require such payoff information provides a valuable contribution to the literature on TOM. Combined with the fact that the TOM framework has been used to model a wide range of political, diplomatic, and historical conflicts, our learning procedure provides an effective contribution to game playing with alternating moves.

We have proved the convergence of the learning TOM players to NMEs that result from TOM play with complete information. We have also discussed the relationship of NME with the concept of subgame perfect Nash equilibrium as discussed in classical game theory literature for dynamic games.

[2] Brams argue that in real-life situations often the starting point or the context of negotiation between negotiating parties already exist, and the negotiators argue over how to change the current context to another, more desirable state.

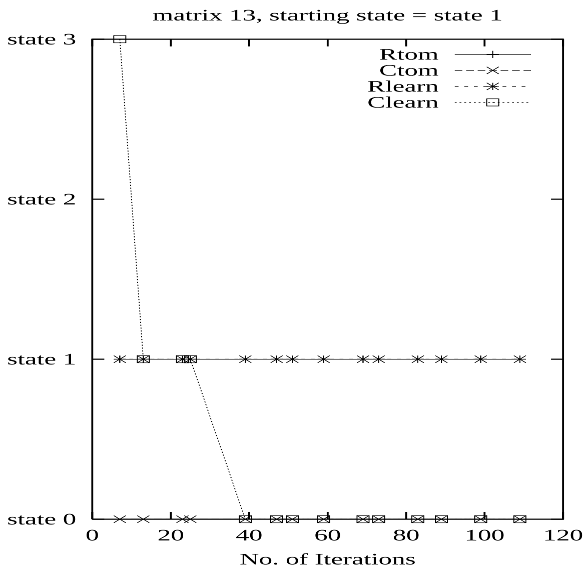

Figure 2: Outcome from state 1 by learning TOMs in Matrix 13. States reached considering R as starting player: State 1; C as starting player: State 0

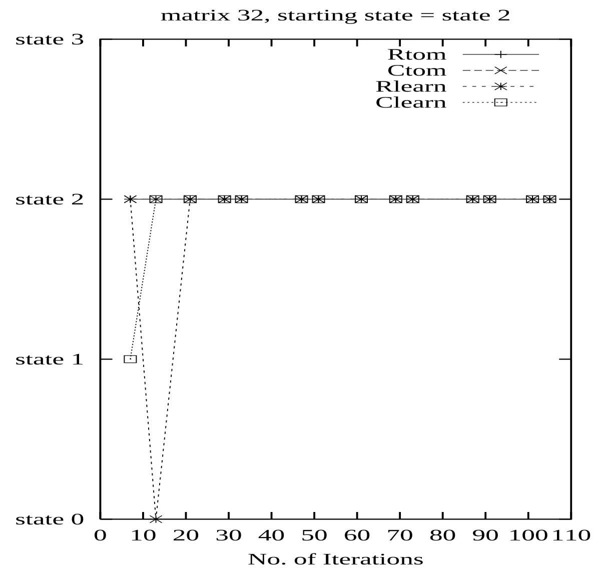

Figure 3: Outcome from state 2 by learning TOMs in Prisoners' Dilemma. State reached considering R and C respectively as starting player: State 2

We plan to scale this approach up to larger matrices. Intuitively speaking, our learning framework is capable to deal with many (more than two) players and multiple payoffs. In case of bimatrix game, each player has to estimate one opponent's move probabilities, whereas in a multi-player game, it has to store these probabilities of all other players. The basic decision mechanism presented here can be applied in multi-player cases as well. We have to experimentally evaluate the scale up properties for more agents.

Acknowledgments: This work has been supported in part by an NSF award IIS-0209208.

References

Banerjee, B.; Sen, S.; and Peng, J. 2001. Fast concurrent reinforcement learners. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 825–830.

Bowling, M., and Veloso, M. 2001. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 1021–1026.

Brams, S. J. 1994. *Theory of Moves*. Cambridge University Press, Cambridge: UK.

Claus, C., and Boutilier, C. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 746–752. Menlo Park, CA: AAAI Press/MIT Press.

Hu, J., and Wellman, M. P. 1998. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik, J., ed., *Proceedings of the Fifteenth International Conference on Machine Learning*, 242–250. San Francisco, CA: Morgan Kaufmann.

Littman, M. L., and Stone, P. 2001. Implicit negotiation in repeated games. In *Intelligent Agents VIII: AGENT THEORIES, ARCHITECTURE, AND LANGUAGES*, 393–404.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, 157–163. San Mateo, CA: Morgan Kaufmann.

Littman, M. L. 2001. Friend-or-foe q-learning in general-sum games. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 322–328. San Francisco: CA: Morgan Kaufmann.

Myerson, R. B. 1991. *Game Theory: Analysis of Conflict*. Harvard University Press.

Nash, J. F. 1951. Non-cooperative games. *Annals of Mathematics* 54:286 – 295.

Rapoport, A., and Guyer, M. 1966. A taxonomy of 2x2 games. *General Systems* 11:203–214.