Modeling Opponent Actions for Table-Tennis Playing Robot

Zhikun Wang and **Abdeslam Boularias** and **Katharina Mülling** and **Jan Peters**
Max Planck Institute for Intelligent Systems
Spemannstr 38, 72076 Tübingen, Germany
{firstname.lastname}@tuebingen.mpg.de

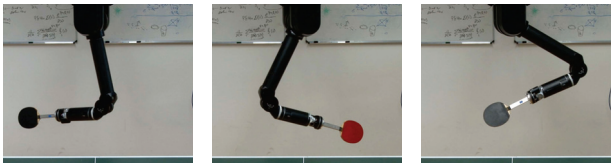
Abstract

Opponent modeling is a critical mechanism in repeated games. It allows a player to adapt its strategy in order to better respond to the presumed preferences of its opponents. We introduce a modeling technique that adaptively balances safety and exploitability. The opponent's strategy is modeled with a set of possible strategies that contains the actual one with high probability. The algorithm is safe as the expected payoff is above the minimax payoff with high probability, and can exploit the opponent's preferences when sufficient observations are obtained. We apply the algorithm to a robot table-tennis setting where the robot player learns to prepare to return a served ball. By modeling the human players, the robot chooses a forehand, backhand or middle preparation pose before they serve. The learned strategies can exploit the opponent's preferences, leading to a higher rate of successful returns.

Introduction

Opponent modeling allows to exploit the opponents' preferences or weaknesses in repeated games. It has been successfully used for computer poker games (Saund 2006), soccer robot games (Butler and Demiris 2009), etc. In practice, inaccurate models are inevitable for a limited number of observations, which exposes the player to the risk of adopting hazardous strategies. To address the safety issue, Markovitch and Reger (2005) proposed to infer a weakness model instead of estimating the precise model. Johanson et al. (2008) proposed an ϵ-safe learning algorithm that chooses the best counter-strategy from a set of safe strategies. Strategies in the set do not lose more than ϵ in the worst case.

In this paper, we propose a different idea that models an opponent's strategy with a set of possible strategies that contains the actual one with high probability. We apply this modeling idea to repeated two-player games. Given a parameter δ that controls the trade-off between safety and exploitability, the proposed algorithm can provide a counter-strategy whose expected payoff is lower-bounded by the minimax payoff with probability no less than $1-\delta$. Therefore, the learned strategy can mildly exploit the observed preferences, and converge to the best-response if the opponent uses a stationary strategy for all games. To cope with non-stationary opponent's strategies, statistical hypothesis tests are used to detect changes in observed preferences.



(a) forehand pose (b) middle pose (c) backhand pose

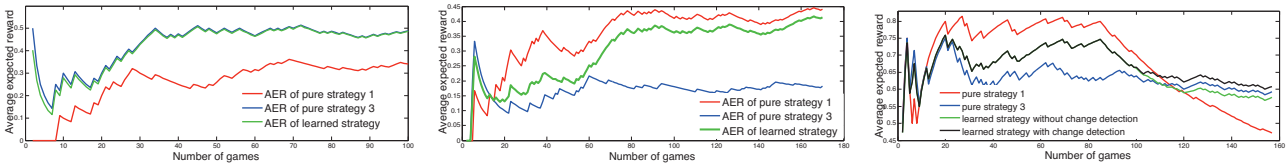
Figure 1: Three pre-defined preparation poses. They are optimized for hitting points in different regions.

The proposed modeling technique allows a table-tennis playing robot to improve its response to balls served by human opponents. The used robot setting (Muelling, Kober, and Peters 2010) has three possible high-level actions, i.e. setting to one of its forehand, backhand and middle preparation poses before the opponent starts serving. Each action has a relatively high success rate when the ball is served to its corresponding region (Figure 1). However, the robot is limited in its acceleration, resulting in low success rate for incoming balls far away from the preparation pose. Assuming that the opponent uses a stationary strategy, this algorithm generates counter-strategies such that the robot can be more likely to successfully return the served ball. We use the low-level planner to verify the feasibility of the plans.

Opponent Modeling and Strategy Learning

We consider two-player normal-form games. The two participants are indicated by player i and player j, and the algorithm learns the strategy for player i. The reward matrix for player i is denoted by $\mathbf{R}$. In each game, the two players choose their own actions a_i, a_j independently from action spaces A_i, A_j according to their respective strategies. The strategies $\boldsymbol{\pi}_i \in \Delta^{|A_i|}$ and $\boldsymbol{\pi}_j \in \Delta^{|A_j|}$ are probability distributions over all possible actions, where Δ^n is the n-simplex set. Thus, the expected reward for player i is $\boldsymbol{\pi}_i^T \mathbf{R} \boldsymbol{\pi}_j$.

Assume the opponent's strategy $\boldsymbol{\pi}_j^*(a_j)$ is stationary during recent N games. With probability no less than $1-\delta$, the Kullback-Leibler (KL) divergence between the empirical distribution $\tilde{\boldsymbol{\pi}}_j$, which is obtained from observed opponent's actions, and $\boldsymbol{\pi}_j^*$ is bounded by $\varepsilon(\delta) = ((|A_j|-1)\ln(N+1) - \ln(\delta))/N$. Using this inequality, we model the opponent's



(a) Opponent with an uniform strategy. (b) Opponent with a strong preference. (c) Opponent with changed strategies.

Figure 2: The curves show the average expected reward, where the pure strategy 1 and 3 always choose the forehand and middle pose. They are the optimal strategy respectively in (b) and (a). The learned strategies converge to them for stationary opponent's strategy, and can adapt to changes of the opponent's strategy.

strategy by $\Omega(\delta) = \{\boldsymbol{\pi}_j \in \Delta^{|A_j|} | \mathrm{KL}(\tilde{\boldsymbol{\pi}}_j || \boldsymbol{\pi}_j) \leq \varepsilon(\delta)\}$, which contains the actual strategy with probability no less than $1 - \delta$. However, the player has no additional information to choose the actual strategy among all possible strategies in this set. To ensure safety of the counter-strategy, we aim for the maximal expected payoff in the worst case: $\boldsymbol{\pi}_i^* = \arg\max_{\boldsymbol{\pi}_i \in \Delta^{|A_i|}} \min_{\boldsymbol{\pi}_j \in \Omega(\delta)} \boldsymbol{\pi}_i^T \mathbf{R} \boldsymbol{\pi}_j$. Finding the counter-strategy can be solved efficiently using sub-gradient methods.

When the model Ω contains the actual strategy $\boldsymbol{\pi}_j^*$, the expected payoff of the best-response strategy has a lower-bound above the minimax payoff. Therefore, the counter-strategy is safe with probability no less than $1 - \delta$. The learned strategy will converge to the best-response counter-strategy if the opponent's strategy converges. In this case, the cumulative expected regret bound with respect to the best-response strategy has a growth rate of $O(\sqrt{T \ln T})$, where T is the number of played games so far.

The learned strategy is δ-safe only if observed N actions are selected from the same strategy. However, the opponent may change its strategy during the games. Therefore, an adaptive learning algorithm is required to deal with the strategy changes. The proposed algorithm maintains two sets of samples: a set X that contains observed actions for learning a counter-strategy, and a set Y that serves as a validation set. We test the hypothesis that the probability of executing any action a_j is the same in the local strategies that generated X and Y. Therefore, the changes in the strategy can be detected with high probability.

Robot Table-Tennis

We consider the problem of choosing the preparation pose for a served ball as a repeated two-player game. The robot has three possible actions, namely, choosing the forehand, backhand or middle preparation poses before the opponent serves. In the meantime, the opponent can choose to serve the ball to the right, left or middle region. We have no knowledge whether the opponent is competitive or cooperative as the opponent's reward is not available. Whereas, an empirical reward matrix for the robot can be computed as the success rates in previous games. Therefore, we roughly know how well a robot's preparation pose can return balls in those three regions. The minimax play will choose the middle preparation pose with probability around 0.8 and the forehand pose with probability 0.2. However, it does not exploit the opponents if they tend to serve the ball to a specific region more frequently.

	Right	Left	Middle
Forehand	0.6506	0.0648	0.5222
Backhand	0.0449	0.3889	0.1222
Middle	0.4103	0.5648	0.7444

We recruited three volunteers to repeatedly serve the ball, and analyze the performance of the algorithms. We measure the performance by the Average Expected Reward (AER), which is the sum of its expected reward divided by the number of trials. The first volunteer served the balls with approximately a uniform distribution over the three regions. Therefore, the *pure strategy* 3 that always chooses the middle preparation pose leads to the maximal AER. As shown in Figure 2(a), the learned strategies preform slightly worse than it in the beginning as the algorithm starts from the minimax strategy, yet quickly converge to the optimal payoff. The second volunteer had a significant preference to serve to the right region. As shown in Figure 2(b), the learned strategies move gradually towards the optimal counter-strategy.

The third volunteer started with the preference of serving the ball to the middle/right side, and then intentionally switched to favoring the middle/left side. Pure strategy 1 and 3 are optimal before and after the switch. We compare the adaptive algorithm to the version without change detection in this case. As shown in Figure 2(c), both algorithms have decreased performance right after the switch. As the hypothesis test failed, the adaptive algorithm successfully detected the strategy switch and adapted to it. Therefore, it achieved the best overall performance.

Future Work

We will extend this modeling idea to more complicated games, e.g. stochastic games, and use it for other robotic applications.

References

Butler, S., and Demiris, Y. 2009. Predicting the movements of robot teams using generative models. *Distributed Auton. Robotics Systems* 8:533–542.

Johanson, M.; Zinkevich, M.; and Bowling, M. 2008. Computing robust counter-strategies. *Advances in Neural Information Processing Systems* 20:721–728.

Markovitch, S., and Reger, R. 2005. Learning and exploiting relative weaknesses of opponent agents. *Autonomous Agents and Multi-Agent Systems* 10(2):103–130.

Muelling, K.; Kober, J.; and Peters, J. 2010. A Biomimetic Approach to Robot Table Tennis. In *Proceedings of IROS*.

Saund, E. 2006. Capturing The Information Conveyed By Opponents' Betting Behavior in Poker. In *IEEE Symposium on Computational Intelligence and Games*.