

# Transfer of Learned Opponent Models in Zero Sum Games

Ismail Guennouni, Maarten Speekenbrink

December 11, 2019

## Introduction

Being able to rapidly learn a model of the environment and successfully transfer previously acquired knowledge to a new domain is one of the hallmarks of human intelligence (Taylor et al., 2008). Humans are naturally endowed with the ability to extract relevant features from a situation, identify the presence of these features in a novel setting and use previously acquired knowledge to adapt to previously unseen challenges using acquired knowledge (Lake et al., 2017). More formally, Perkins (1992) defines transfer of learning as the application of skills, knowledge, and/or attitudes that were learned in one situation to another learning situation. This is a major theme in current research on artificial intelligence, and neural networks in particular. While these architectures have made important strides in various challenges from machine vision to game play, we have not been able to build AI systems that learn as fast and as efficiently as humans do, and more importantly that have the ability to generalize and transfer knowledge to new domains. Therefore, being able to capture mechanisms that underpin how humans learn and transfer models of the environment is likely to be a key component towards building a general artificial intelligence.

In this paper, we are specifically interested in the way in which people build and use models of their opponent to facilitate learning transfer, when engaged in situations involving an interaction with strategic considerations. These situations arise frequently such as in negotiations, auctions, strategic planning and all other domains in which theory of mind abilities (Premack Woodruff, 1978) play a role in determining human behaviour.

In order to explore learning transfer in strategic settings, it is generally useful to reduce the complexity of real life interactions by using an experimental paradigm that reduces non relevant confounding effects. In the context of inter-dependent decision making, the study of simple games as a model of more complex interactions plays this essential role. More specifically, when studying learning, we need a framework that allows the study of whether and how a player takes into consideration, over time, the impact of its current and future actions on the future actions of the opponent and the future cumulative rewards. Repeated games, in which players interact repeatedly and have the ability to learn about the opponent's strategies and preferences (Mertens, 1990) are therefore used whenever we are interested in the evolution of learning.

Early literature on learning transfer in games has mostly focused on measuring the proportion of people who play normatively optimal or salient actions in later games, having had experience with a similar game environment initially. For instance, Ho et al. (1998) measure transfer as the proportion of players who choose the Nash Equilibrium in later p-beauty contest games having had experience with similar games. They find there is no evidence of immediate transfer (Nash equilibrium play in the first round of the new game) but positive structural learning transfer as shown by the faster convergence to equilibrium play by experienced vs non experienced players. Camerer Knez (2000) test learning transfer in players exposed to two games with multiple equilibria sequentially and explore the ability of players to conordinate their actions to choose a particular equilibrium in subsequent games having reached it in prior ones. They distinguish between games that are similar in a purely descriptive way, meaning similar choice labels, identity of players, format and number of action choices; and games that are similar in a strategic sense, meaning similar payoffs from combination of actions, identical equilibrium properties or significant social characteristics of payoffs such as possibility of punishment, need for fairness and cooperative vs competitive settings. They find that transfer of learning (successful coordination) occurs more readily in the presence of both descriptive and strategic similarity. If the games were only strategically similar, then the transfer was much weaker.

Juvina et al.(2014) made a similar distinction between what they deemed surface and deep similarities and find that both contribute to positive learning transfer. However, they show that surface similarity is not necessary for deep transfer, and can either aid or block this type of transfer depending on whether it leads to congruent or incongruent actions in later games. In a series of experiments using economic signalling games, Cooper Kagel (2003, 2004, 2008) find that participants who have learned to play according to Nash Equilibrium in one game can transfer this to subsequent games, even though the actions consistent with Nash Equilibrium in later games are different. They show that this transfer is driven by the emergence of sophisticated players who are able to represent the strategic implications of their actions and reason about the consequences of changed opponent payoffs.

Most of these studies fail to offer a formal explanation of this transfer or a modelling framework that can explain the experimental observation of transfer between games and generalise it to extensive classes of games. A notable exception is the effort by Haruvy and Stahl (2012) to specify a model of learning where players learn abstract rules that they can generalise and transfer across dissimilar games, rather than action choices that can only be used within the same game. Participants played ten games, presented as 4x4 normal form (matrix payoffs). Their results suggest that subjects do transfer learning over descriptively similar but strategically dissimilar games and that this learning transfer is significant. They also showed that players learn abstract aspects of the game that are then transferred to new settings. Their rule learning model, based on Stahl (1996) was able to capture participants dynamic behavior and show that the propensity to select particular rules is perfectly transferred across games.

In most of these studies, participants were playing other human players either one on one or as a group. In this case, players are learning about the game but so are their opponents. While the approach of matching groups of human players repeatedly to play economic games has ecological validity in recreating environments where social decision making can be studied, it complicates the detection and modelling of strategies used by participants. It

is harder to focus on an individual and how her strategies are changing and adapting to the opponent’s play if we cannot experimentally control the behaviour of the opponent. It is also difficult to disentangle the process of learning about the opponent from that of learning about the game structure and payoffs.

Therefore, studies seeking to model human strategy learning have replaced the human opponent with a computerised agent, allowing researchers to manipulate the opponent’s behaviour and observe how the human participants adapt their behaviour. For instance, Spiliopoulos (2013) made humans play constant sum games against 3 computer opponents programmed to take advantage of known patterns in human play such as imperfect randomization and heuristics use, and found that human participants do adapt to the opponent they are facing. Shachat Swarthout (2004) made human participants face computer opponents playing various mixed strategies in a zero-sum asymmetric matching pennies game. They found that the players changed their strategies towards exploiting the deviations from the MSNE, and that this exploitation was very likely if the deviation from the MSNE play was high. While useful in showing how adaptable humans are to variance in opponent strategies, the computerised opponents used did not play the way humans participants would. After all, we are not very good at playing specific mixed strategies with any precision or at detecting patterns from long sequences of past play due to cognitive constraints. Also, none of these studies looked at the learning transfer aspect.

In this study, we aim to contribute to this under-explored literature through testing opponent modelling and its transfer with the use of both similar and dissimilar games played against computer opponents implementing fixed behavioral rules with differing degrees of sophistication. Unlike previous studies, our contribution is to focus on purely competitive situations through the use of zero sum games. Indeed, previous studies have used games that can give rise to confounding factors. For instance, social dilemma games and coordination games crystallise the conflict humans deal with between self-interest and cooperation. Playing these games repeatedly implicates important social constructs such as reputation building, trust and other psychological attributes such as cooperativeness. Coordination games also test the ability of choosing the safe self-interested choice compared to the risky cooperative choice and may depend on similar latent factors. As such, these games are not as helpful as purely competitive settings in isolating how humans model their opponents and the mechanism through which these models are updated, learned and eventually transferred. Zero sum games for instance do not incentivize any cooperation since one player’s gain is necessarily the opponent’s loss. As such, they encourage learning about the opponent play and exploiting that in order to maximise cumulative.

Moreover, studies that looked at how players adapt to their opponent strategies when facing computer agents have mostly looked at the ability of players to detect and exploit action-based learning rules. They either used deterministic strategies that did not consider the human participant’s play, as in the case of pre-determined mixed strategy algorithms, or strategies that only depended on the player’s frequency or patterns of past plays, not considering the computer’s own past plays, or the degree of strategic reasoning of the human player about the computer. We aim to use computer agents who mimic human theory of mind abilities and as such, play in a more "human" way.

In order to build agents with human like characteristics, we use the idea that humans possess a 'bounded rationality'. Simon (1972) explains that humans have limited cognitive

capacities and as such cannot be expected to solve computationally intractable problems such as finding Nash equilibria. Instead, they will try to 'satisfice' by choosing a strategy that is adequate in a simplified model of the environment, rather than an optimal one. This concept finds its natural application in 'level-k' theory, first adopted by Stahl & Wilson (1995) which posits that deviations from Nash equilibrium solutions in simple games are explained by the fact that humans have a heterogeneous degree of strategic sophistication. At the bottom of the ladder, level-0 players are non-strategic and play either randomly or use a salient strategy in the game environment (Ariad & Rubinstein, 2012). Level-1 players are next up the ladder of strategic sophistication and will assume all their opponents belong to the level-0 category and as such will best respond to them given this assumption. Likewise, a level-2 player will choose actions that are the best response given the belief that all opponents are exactly one level below and so on. In this study, our computer agents will play either as a level-1 or level-2 human player would, mimicking human theory of mind abilities and the limited recursion depth they exhibit (Goodie et al., 2012).

Finally, unlike the vast majority of experimental paradigms, we make participants play the games in an engaging, interactive and ultimately more ecological way rather than simply provide the matrix form of the game. We use visual aids representing the player and their computer opponent as well as the various actions that people can select built on a highly interactive interface providing image, audio and sometimes video feedback at each step, from choosing actions, to presenting outcomes of the rounds and players scores.

## Methods

We ran two experiments where human participants played against computer opponents a total of 3 different games. In the first experiment, participants played 3 games sequentially against the same computer opponent. The computer opponent either used a level-1 or level-2 strategy. The three games were rock-paper-scissors, fire-water-grass, and the numbers game. A typical Rock-paper-scissors game (hereafter RPS) is a 3x3 zero sum game, with a cyclical hierarchy between possible actions: rock blunts scissors, paper wraps rock, and scissors cut paper. If one player chooses an action which dominates their opponent's action, the player wins (receives a reward of 1) and the other player loses (receives a reward of -1). Otherwise it is a draw and both players receive a reward of 0. It has a unique MSNE consisting of randomly playing one of the three options each time.

The second game is identical to Rock-Paper-Scissors in all but action labels. We call it Fire-Water-Grass (FWG): Fire burns grass, water extinguishes fire, and grass absorbs water. We are interested in exploring whether learning is transferred in a fundamentally similar game where the only difference is in the description of the choice actions. Finally, the numbers game is a generalization of rock-paper-scissors. In the variant we use, 2 participants concurrently pick a number between 1 and 5. To win in this game, a participant needs to pick a number exactly 1 higher than the number chosen by the opponent. For example, if a participant thinks their opponent will pick 3, they ought to choose 4 to win the round. To make the strategies cyclical as in RPS, the game stipulates that the lowest number (1) beats the highest number (5), so if the participant thinks the opponent will play 5, then the winning choice is to pick 1. This game has a structure similar to RPS in which every action

is dominated by exactly one choice. All other possible combination of choices that are not consecutive are considered ties. A win would add 1 point to the score of the player, while a loss deduces one point and a tie does not affect the score. Similar to RPS, the MSNE is to play each action with equal probability in a random way.

Participants were informed they would play three different games against the \*same\* computer opponent, namely: Rock-Paper-Scissors, Fire-Water-Grass and the numbers game. Each participant plays all three games consecutively and in the same order described above. Participants were told that the opponent cannot cheat and will choose its actions simultaneously without knowledge of the participant's choice. A total of 50 rounds of each game was played with the player's score displayed at the end of each game. The score was calculated as the number of wins minus the number of losses. Ties did not affect the score. In order to incentivise the participants to maximise the number of wins against the opponents, players were paid a bonus at the end of the experiment that was proportional to their final score. At the end of the experiment, the overall score of the players was translated into the bonus by making each point worth  $\pounds 0.02$ . This bonus is significant as players could increase the total payoff from the experiment by up to 60% assuming they'd won all rounds against the computer opponent.

In the second experiment, participants each played 3 games sequentially against [both] computer opponents, rather than just one like in the first experiment. The computer opponents either used a level-1 or level-2 strategy throughout the three games. The three games were Rock-Paper-Scissors, Fire-Water-Grass, and the penalty shootout game. The first two games were identical to the ones used in the first experiment. In the final game (shootout) the participants played the role of the player shooting a football (soccer) penalty, and the AI opponent was the goalkeeper. Players had the choice between three actions, like in the first two games: Shooting the football to the left, right or centre of the goal. If the player shoots in a direction different from that of where the goalkeeper dives, they win the round and the goalkeeper loses. Otherwise, the goalkeeper catches the ball and the player loses the round. There is no possibility of ties in this game. We can easily check that there is no pure strategy Nash equilibrium in this game. However, like the first 2 games, there is a unique mixed strategy Nash equilibrium consisting of randomising over the three actions for both players (Probability of shooting/diving in any direction is  $1/3$ ). What makes this game different however is that there are two ways to beat the opponent in each round: if we think the opponent is going to choose "right" in the next round, we can win by both choosing "left" and "centre". A level-1 player (thinks that his opponent will repeat his last action) has two ways to win the next round. As such, we have programmed the level-2 computer program to choose randomly between the two possibilities that a level-1 player may choose.

Like in the first experiment, the computer opponents retained the same strategy throughout the 3 games, however the participants faced each opponent twice in each game. Specifically, each game is divided into 4 stages numbered 1 to 4, consisting of 20, 20, 10, and 10 rounds respectively for a total of 60 rounds per game. Participants start by facing one of the opponents in stage one, then face the other in stage two. This is repeated in the same order in stages 3 and 4. Which opponent they faced first was counterbalanced. All participants engage in the same three games (namely RPS, FWG and Shootout) in this exact order, and were aware that the opponent was not able to know their choices beforehand but was

choosing simultaneously with the player.

To summarise, all the games used in both experiments have a unique MSNE consisting of randomising across actions. If participants follow this strategy, or simply don't engage in learning how the opponent plays, they would score 0 on average against both level-1 and level-2 players. Evidence of sustained wins would indicate that participants have learned to exploit patterns in the opponent play.

## **Results**

## **Discussion**

## **Conclusion**

## **1 References**

Goodie, A. S., Doshi, P., Young, D. L. (2012). Levels of theory-of-mind reasoning in competitive games. *Journal of Behavioral Decision Making*, 25(1), 95–108. Return to ref 34 in article