Word count: 1006

Introduction

The aim of this paper is to understand the effectiveness of learning in repeated, finite and infinite, stage games. The strategy profile of agents, aiming at maximising their utility, converge eventually to the Nash Equilibrium (NE) of the repeated game, independently on information being incomplete or not (Kalai & Lehrer,1993). Furthermore, forgivingness can win you tournaments, as is the case with the strategy profile *Tit-for-tat* (Simpson,2016), proposed by one of the greatest contributors to modern game theory, namely Anatol Rapoport. This paper compares *tit-for-tat* with two learning-based strategies.

It is well-known from evolutionary game theory, that a strategy is dominating (evolutionary stable) if a strategy resists the invasion of invaders, deploying another strategy, over time (Smith, 1982).

This paper intends to explore the implications of combining the insights from the three citations, and thereby gain understanding of the effectiveness of learning in repeated games. See the ODD, 1, for an in-depth description of the *purpose* of this paper.

Note on Multi Agent Fictitious Play

The idea behind MA-FP is to counteract any irrationality inherent in the opponent, as described in 4a and 7 (MA-FP) of the ODD. This counteracting procedure can be implemented in at least two slightly different ways, which can be expected to influence the effectiveness of the procedure.

The foundation of MA-FP is that local characterisations in dynamic environments (here local beliefs in a repeated stage game) not necessarily are accurate representations, to derive optimal responses based on. This is a well-known fact in optimisation, where local optimums not necessarily are equivalent to global optimums.

How to obtain a correct view, which in this setup constitutes to how to sample the global beliefs, is a reasonable question.

Two possible ways are, to either sample all beliefs present in the stage space or only to sample those beliefs belonging to agents with same strategy profile, as the opponent. The former is implemented in this model, but it seems reasonable to expect that this causes too much smoothing of the optimal behaviour, and reduces the effectiveness of MA-FP. The latter way can be expected to be unstable and ineffective for a low number of games/agents and replies on the assumption of complete information. However, it is subject to future investigation.

Notes on the setup of the games

Most of the setup of games are shown in table-1, and further discussed in the next section. However, the static variables of the games are the number of agents, the bounds of the beliefs and the pay-off matrix. The number of agents is 10 and the upper bounds of the beliefs are chosen to be six. The bounds are chosen as they are, because an upper bound of six allows for some variation but not too

much, which is desirable because the learning-based strategies are sensitive to the initial beliefs. However, the actual effect of the upper bounds is something, which are subject to future investigation. The pay-off matrix is symmetric, same for row and column player, and both players receive payoff at each interaction. The payoff matrix is seen in table-1 below.

Table-1: ***Payoff-matrix***

	0	1	2
0	2/2	1/1	0/0
1	3/3	1/1	0.3/0.3
2	4/4	1/1	0.6/0.6

Results

Having a model like this fosters many interesting things to investigate, yet limitations is often shown to be key when understanding is the aim, which is the case with this paper. The explored combinations of this paper are seen from table-2, which contains 16 combinations. Table-2 contains variation in five key variables, namely the three strategies (*Random* is a residual from main strategies), *Rationality* and *Decision*. The chosen variables enable the investigation of the research question formulated in the ODD, see 1. Two horizons are considered, a finite repeated stage game of 50 games and 2000 games, which is meant to simulate an infinite repeated stage game. The results are seen from table-3 and table-4 respectively. The results are average results over 500 runs, and *Total Utility* is expressed as the amount of the specific combination normalised by the maximum over all combinations within that horizon (over all columns). Furthermore, the total amounts of utility obtained by each group of agents, employing the different strategies, are normalised by the total amount of utility within that combination (over the four strategy-rows). Normalisation is done to ease comparison, as actual levels, as such, aren't of interest.

Table-2: ***Investigated Combinations***

	Tit-for-tat	Fictitious Play	MA Fictitious Play	Random	Rationality	Decision	Combinations
Equal Split	0.25	0.25	0.25	0.25	1.0	1.0	0
High Tit-for-tat	0.50	0.17	0.17	0.16	1.0	1.0	1
High Fictitious Play	0.17	0.50	0.17	0.16	1.0	1.0	2
High MA Fictitious Play	0.17	0.17	0.50	0.16	1.0	1.0	3
Equal Split	0.25	0.25	0.25	0.25	1.0	0.5	4
High Tit-for-tat	0.50	0.17	0.17	0.16	1.0	0.5	5
High Fictitious Play	0.17	0.50	0.17	0.16	1.0	0.5	6
High MA Fictitious Play	0.17	0.17	0.50	0.16	1.0	0.5	7
Equal Split	0.25	0.25	0.25	0.25	0.5	0.5	8
High Tit-for-tat	0.50	0.17	0.17	0.16	0.5	0.5	9
High Fictitious Play	0.17	0.50	0.17	0.16	0.5	0.5	10
High MA Fictitious Play	0.17	0.17	0.50	0.16	0.5	0.5	11
Equal Split	0.25	0.25	0.25	0.25	0.5	1.0	12
High Tit-for-tat	0.50	0.17	0.17	0.16	0.5	1.0	13
High Fictitious Play	0.17	0.50	0.17	0.16	0.5	1.0	14
High MA Fictitious Play	0.17	0.17	0.50	0.16	0.5	1.0	15

Table-3: ***Average results after 50 games***

Combinations	0	1	2	3	4	5	6	7
Total Utility	0.815	0.983	0.702	0.687	0.826	0.989	0.696	0.691
totalRandomUtility	0.222172	0.147861	0.131218	0.132039	0.226514	0.15332	0.133424	0.126372
totalTFTUtility	0.256426	0.548774	0.195864	0.169986	0.270982	0.549092	0.183641	0.182752
totalFPUtility	0.257159	0.144336	0.505957	0.186057	0.255901	0.157037	0.51457	0.172664
totalMAFPUtility	0.264242	0.159029	0.166962	0.511918	0.246604	0.140551	0.168364	0.518212
averageGlobalBeliefsFP	[0.219 0.188 0.536]	[0.236 0.154 0.458]	[0.202 0.176 0.622]	[0.163 0.146 0.527]	[0.221 0.188 0.537]	[0.248 0.162 0.477]	[0.199 0.177 0.621]	[0.166 0.148 0.53]
averageGlobalBeliefsMAFP	[0.221 0.192 0.547]	[0.247 0.157 0.466]	[0.169 0.149 0.522]	[0.197 0.177 0.624]	[0.222 0.187 0.533]	[0.232 0.152 0.45]	[0.164 0.146 0.526]	[0.2 0.176 0.624]
latestMostPlayedAction	2	2	2	2	2	2	2	2
Combinations	8	9	10	11	12	13	14	15
Total Utility	0.822	0.98	0.691	0.698	0.814	1	0.934	0.691
totalRandomUtility	0.228985	0.14678	0.134495	0.135581	0.228666	0.154601	0.457703	0.130668
totalTFTUtility	0.260146	0.54823	0.172756	0.179414	0.250171	0.560789	0.163646	0.173654
totalFPUtility	0.248405	0.148303	0.519072	0.178876	0.259997	0.135465	0.183517	0.175004
totalMAFPUtility	0.262464	0.156687	0.173677	0.506128	0.261166	0.149144	0.195135	0.520675
averageGlobalBeliefsFP	[0.221 0.189 0.532]	[0.239 0.157 0.462]	[0.198 0.178 0.622]	[0.167 0.151 0.532]	[0.218 0.189 0.541]	[0.235 0.151 0.444]	[0.222 0.209 0.406]	[0.165 0.148 0.525]
averageGlobalBeliefsMAFP	[0.218 0.189 0.532]	[0.239 0.157 0.466]	[0.165 0.149 0.53]	[0.199 0.178 0.62]	[0.22 0.19 0.54]	[0.241 0.157 0.455]	[0.227 0.212 0.411]	[0.197 0.178 0.625]
latestMostPlayedAction	2	2	2	2	2	2	2	2

The results in table-3 and table-4 are similar, and three interesting insights emerge. *Tit-for-tat* is the dominating strategy, in terms of achieving highest *total utility*, consisting with (Simpson, 2016). Secondly, all learning strategies favours to play *Hostile*, which is the NE of the stage game, consistent with (Kalai & Lehrer,1993).

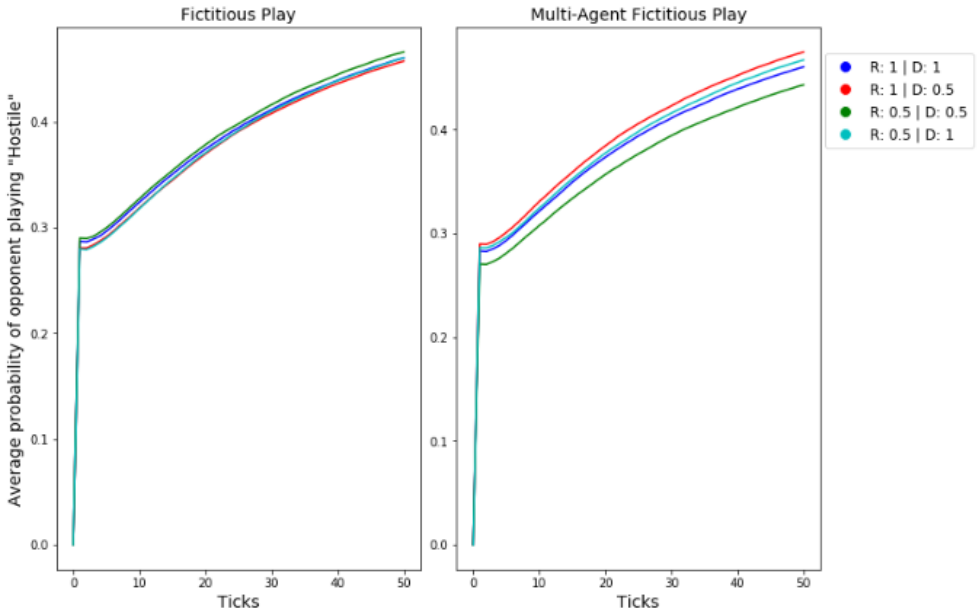
Table-3: ***Average results after 2000 games (approximating infinity)***

Combinations	0	1	2	3	4	5	6	7
Total Utility	0.814	0.977	0.686	0.679	0.82	1	0.691	0.693
totalRandomUtility	0.22323	0.142467	0.130523	0.124844	0.232577	0.145422	0.136244	0.137133
totalTFTUtility	0.298519	0.620678	0.210253	0.203526	0.293217	0.636282	0.206869	0.209146
totalFPUtility	0.240729	0.116939	0.494441	0.178908	0.230656	0.104593	0.499043	0.164708
totalMAFPUtility	0.237521	0.119916	0.164783	0.492723	0.24355	0.113703	0.157844	0.489013
averageGlobalBeliefsFP	[0.13 0.099 0.705]	[0.167 0.061 0.6]	[0.086 0.068 0.844]	[0.067 0.054 0.733]	[0.131 0.1 0.693]	[0.172 0.062 0.576]	[0.088 0.07 0.84]	[0.068 0.055 0.696]
averageGlobalBeliefsMAFP	[0.128 0.097 0.704]	[0.177 0.065 0.612]	[0.067 0.054 0.707]	[0.083 0.066 0.851]	[0.138 0.105 0.713]	[0.189 0.063 0.604]	[0.073 0.059 0.716]	[0.089 0.071 0.84]
latestMostPlayedAction	2	2	2	2	2	2	2	2
Combinations	8	9	10	11	12	13	14	15
Total Utility	0.817	0.985	0.688	0.688	0.804	0.979	0.967	0.683
totalRandomUtility	0.227044	0.146273	0.13679	0.137674	0.219063	0.152042	0.447392	0.133766
totalTFTUtility	0.2986	0.621334	0.200169	0.201067	0.289399	0.609148	0.194872	0.196239
totalFPUtility	0.236497	0.112486	0.498799	0.160597	0.242423	0.123763	0.172166	0.167663
totalMAFPUtility	0.23786	0.119908	0.164242	0.500661	0.249115	0.115047	0.18557	0.502332
averageGlobalBeliefsFP	[0.135 0.102 0.718]	[0.174 0.064 0.598]	[0.087 0.071 0.842]	[0.069 0.057 0.702]	[0.126 0.096 0.717]	[0.175 0.067 0.606]	[0.176 0.162 0.5]	[0.069 0.056 0.723]
averageGlobalBeliefsMAFP	[0.131 0.099 0.708]	[0.179 0.065 0.608]	[0.069 0.057 0.702]	[0.086 0.07 0.84]	[0.131 0.1 0.725]	[0.17 0.064 0.599]	[0.179 0.165 0.514]	[0.085 0.069 0.846]
latestMostPlayedAction	2	2	2	2	2	2	2	2

Thirdly, *Rationality* does not seem to affect the results, at first, visible by the fact that the lower parts of table-3 and table-4 are like their respectively upper part. However, looking at the two last combinations, where *Rationality* is 0.5 and *Decision* is 1, shows that MA-FP can counteract the irrationality. It is seen by the fact that the largest group accountable for the total utility is *MA-FP* in combination 15, compared to 14 where it is *Random*. *Random* dominates 14 because the irrationality has implied that the random strategy has won most of the games.

Figure-1 shows the average learning curves for each learning-based strategy, for all *Tit-for-tat* combinations. The learning curves for the *Tit-for-tat* combinations are shown because they represent the *worst case* scenario, i.e. where the learning-based strategies are smaller proportions compared to the dominating strategy. The two average learning curves shows the same tendency, which is that the NE is quickly located, and convergence starts. A more more in-depth analysis of the learning curves is left for future investigation.

Figure-1: *Average beliefs*



Conclusion

With the research question in mind, the learning-based strategies didn't manage to dominate the old dog *Tit-for-tat*, in the settings considered. *Tit-for-tat* where dominating in terms of obtaining high *total utility*. The learning-based strategies converged to the NE strategy, as expected, and they did it rather quickly, see figure-1. All in all, is the conclusion that, the learning-based strategies isn't superior to *Tit-for-tat*, within this setup. However, this investigation has only scratched the surface, and the learning-based strategies should be studied in more complex environments

Refences

Kalai, E. & Lehrer, E. (1993). *Rational Learning Leads to Nash Equilibrium*. Econometrica, Vol. 61, No. 5 (September, 1993), 1019-1045.

Simpson, E. (2016). *The Contributions of Anatol Rapoport to Game Theory*. Political Science Publications. 135.

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge University Press.