Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210145

Alternation Measures for the Evaluation of Selfish Agents' Turn-Taking

Nikolaos Al. PAPADOPOULOS[a] and Marti SANCHEZ-FIBLA[b,1]
[a] *Department of Applied Informatics, University of Macedonia, Greece*
[b] *Department of Technology, Universitat de Pompeu Fabra, Spain*

Abstract. Multi-Agent Reinforcement Learning reductionist simulations can provide a spectrum of opportunities towards the modeling and understanding of complex social phenomena such as common-pool appropriation. In this paper, a multiplayer variant of Battle-of-the-Exes is suggested as appropriate for experimentation regarding fair and efficient coordination and turn-taking among selfish agents. Going beyond literature's fairness and efficiency, a novel measure is proposed for turn-taking coordination evaluation, robust to the number of agents and episodes of a system. Six variants of this measure are defined, entitled Alternation Measures or ALT. ALT measures were found sufficient to capture the desired properties (alternation, fair and efficient distribution) in comparison to state-of-the-art measures, thus they were benchmarked and tested through a series of experiments with Reinforcement Learning agents, aspiring to contribute novel tools for a deeper understanding of emergent social outcomes.

Keywords. Reinforcement Learning; Game Theory; Multi-agent Battle of the Exes; MBoE; Markov Games; Perfect Alternation; Fairness; Efficiency; Alternation Measures; ALT

1. Introduction

Multi-Agent Reinforcement Learning (*MARL*) can be adequate to model and explain complex conflictive situations like common-pool resource appropriation [1]. With a Markov Game [2] like modelling one can express situations that go beyond the minimal Game Theoretic frameworks. With *RL*, agents can learn selfishly and the equilibria reached can be studied [1]. Many measures (taken from Economics) have been applied to study the characteristics of such equilibria like Efficiency, Fairness and Sustainability e.g. [1],[3],[4]. These measures, although being adequate to capture the general exploitation of reward and how it is spread among agents, fail to capture the temporal dynamics of how reward is exploited, i.e. by turn taking. For this purpose we define a minimal environment based on the BoE [4] for the computational examination of turn-taking coordination among multiple selfish Q-learning agents. As it was shown in [2], the literature's Fairness and Efficiency measures were found insufficient and non-indicative for the evaluation of the fair and efficient distribution of such multi-agent systems, as mainly, they can either be "blind" to unfairness or inefficiency. Instead, we introduce the Perfect Alternation (*PA*) equilibrium notion as an optimal case where many agents acquire the full reward successively, as a point of reference to describe such systems' emergent behaviors by their discrepancy from PA. Furthermore, 6 novel Alternation measures for evaluation are shortly defined. Some results are indicatively presented at the end, to showcase the proposed measures usage.

[1] Corresponding Author: Technology Department, Universitat Pompeu Fabra, Carrer de Roc Boronat 138, 08018 Barcelona, Spain. E-mail: marti.sanchez@upf.edu.

2. Multiplayer Interpretation of the Battle of the Exes game

For the *MBoE* version, we consider that *n* agents do cooperate, if and only if they successfully alternate to acquiring the max reward, as this would be the fairest and efficient distribution of the recourses. We suggest a minimal interpretation yet other interpretations can also be considered. To define the problem, the goal is to generalize BoE to a multi-player game-theoretic scenario so that the coordination of selfish agents can be tested experimentally in a minimally dynamic environment. The basic requirement to ensure consistency with the problem definition of the 2-players version is the conceptualization of episodic non-cooperative game-theoretic scenarios where every agent moves simultaneously in each round. When only one agent reaches a terminal state, it gets the higher possible payoff and the remaining *n-1* agents get 0, yet when all *n* players reach the terminal state (tie) no one gets a payoff. The above are considered to be the basic limitations to ensure that the problem definition is consistent with BoE. It is accepted here that every agent can only move one step at a time only at his own pathway of *m* possible positions, including the initial one. The only way that an episode is terminated is when **at least one** agent reaches the end of its pathway. As in [2], an episode cannot end in a round that everyone halts.

Furthermore, special attention is required to be given regarding the reward that will be acquired, in the case that only some of the agents *k* with $0<k<n$, manage to reach their goal-states and therefore, if the game will be zero-sum or not. For example: 3 agents ($n = 3$) compete over the high payoff of 1 that can be acquired only if they reach the terminal state individually. In case that more than one agents reach the terminal state, they get only a low partial payoff, for example, $1/9$ ($p = 1/n^2$), and they get no payoff if all of them reach it together. This minimally dynamic version was carried out included 3 positions per agent. The initial, an intermediate, and a top one.

3. Perfect Alternation Measures

Before proceeding, it would be useful to define and propose Perfect Alternation (PA) Equilibrium. A Perfect Alternation (PA) in repeated games is considered the Pareto-optimal Nash Equilibrium when **all *n* players, alternate successively to the state of the highest payoff, one-by-one, and episode-by-episode, in any order**. However, this order is ideally repeated intact every *n* episodes. This equilibrium is meant to be diversified from other types of Alternation Equilibria which can include turn-taking of players every any number of episodes, in groups or even asymmetrically for each agent/ group. Of course, there can be other types of special equilibria that can be fair and efficient, however, their definition as "alternation equilibria" can be questioned, as semantically maybe "solid alternation" should imply fixed periodicity of turn-taking among agents as in [4]. More specific practical and theoretical purposes that motivated the distinction of PA from the rest, are analytically discussed in [2].

Alternation measures or *ALT*, calculate the discrepancy of a turn-taking system behaviour from the ideal case of PA. Specifically, they aim to capture the performance of agents' succession to terminal positions, measuring the weighted rate of successful alternation of winners. This repeats per all possible sequences of *n* episodes - called batches - to indicate the agents' coordination throughout all *v* episodes. Ideally, every agent should win once every *n* episodes. For this reason, the algorithm evaluates each batch of episodes *b,* which can be considered as an overlapping window of *n* (number

of agents) size. Then, the normalized accumulated evaluations of batches are averaged by the number of batches. Of course, the total number of batches is always $b=v-(n-1)$ so the number of agents n must always be at least equal to the number of episodes v.

Specifically, to evaluate each agent's alternation within each batch, first, a sub-measure/ weight β is calculated (β calculation is analytically explained below). In the best-case scenario, β is equal to *1* for each batch. Thus, for all versions of ALT that are introduced in [2], the optimal Alternation's value $\widehat{ALT}$ is equal to *1*, meaning that in such a case all n agents exclusively won in succession, one per time throughout all v episodes. However, for each agent that wasn't included in the list of winners within n episodes of a batch, the evaluation measure β of this batch is reduced. In tie cases, that more than one winners occur in one episode, the algorithm splits their evaluation weight β, in a different manner according to the version of *ALT* that is used. $ALT^n = \frac{\sum_{j=1}^{b}\beta_j}{b}$, where j is the integer id of each batch, b is the number of possible batches within ν episodes and β is the sub-measure that evaluates each batch's alternation of agents accumulatively. β weights vary depending on which the version is measured.

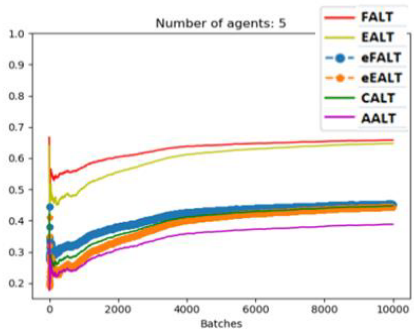
Fractional Alternation Measure (FALT) weights each batch with a fraction of the number of individual agents that managed to reach their terminal position at least once denoted, by the total of all the cases that an agent reached it, within this batch j: $\beta_j^{FALT} = \frac{f_j}{t_j}$ with f denoting how many agents out of n appeared in their terminal positions at least once within this batch of episodes, t denoting all the terminal occurrences, j the id of this batch. ***Exponential Fractional Alternation Measure (eFALT)*** uses the exponential version (square of) β_j^{FALT}. It is proposed for the cases when "stricter" evaluations of low alternation (as defined for FALT) are preferred, while more "generous" evaluations of higher alternation fit better to the given problem.

Exclusive Alternation (EALT) measures the rate which agents exclusively win within each episode of a batch. It is not that tolerant as FALT because only one agent should win at each episode. Otherwise the whole episode will be evaluated with *0* for all the batches that are included. The β value is calculated as $\beta_j^{EALT} = \frac{w_j * f_j}{n^2}$ where w_j is the number of winning episodes with an exclusive winner within the j^{th} batch of episodes. ***Exponential Exclusive Alternation Measure (eEALT)*** or β_j^{eEALT} is again the square of β_j^{EALT} for an exponential "treatment" of the evaluation as in *eFALT*.

Complete Alternation (CALT) is stricter than the last proposed versions of the ALT measure, as it assigns a weight of *0* to tie situations. Specifically, for each episode, it multiplies β^{eFALT} with the difference between the maximum number of possible agents that can reach their top position and the actual ones. Then it adds up all the weighted β^{eFALT} values of each episode to divide them with the number of episodes per batch, which is always equal to n, times the max possible weight, which is equal to *(n – 1)* when only one agent reached the top in an episode, to average each episode's performance. This way, the fewer the winners the more the weight of β^{eFALT} for each episode. In the extreme case of a tie, this episode is evaluated with 0, affecting the batch's evaluation. Its β batches weights are calculated as follows: $\beta_j^{CALT} = \frac{\sum_{k=1}^{n}\left((n-Y_k)*\beta_j^{eFALT}\right)}{n*(n-1)}$, where integer k indicates the id of an episode within batch j with $0 \le k \le n-1$ and n is the number of agents. Also, Y is an n-sized array whose each element contains the number of agents who reached their top, for every episode k of batch j.

Absolute Alternation (AALT) is the last and the most sensitive measure to Alternation, as its changes are dramatic depending on the alternation phenomenon. It is expected to always be lower or equal to the rest of the Alternation measures. It assigns any non-exclusive winning position of an agent within a batch, with a weight of *0,* and takes into account only the successful alternations of exclusive winners of a batch. It gets *1* only in the ideal case that all agents win at least and only once per batch of episodes. Its β batches weights are calculated easily as follows: $\beta_j^{AALT} = \frac{g_j}{t_j}$, where g_j is the number of unique exclusive winnings of all agents within the whole j^{th} batch.

All the ALT measures values *X* are suggesting a level of alternation *A(X)* or *AltRatio.* This ratio or percentage is calculated by a function of the number of agents and the coefficients which are estimated by a environment-specific model-fitting regression. For this regression, extreme cases have been evaluated to be used as benchmarks for each version of ALT measure, as if $x \in [2,40]$ agents were perfectly alternating among each other to their top positions, while the rest *n-x* did not move from their initial positions, as analytically explained in [2]. Thus *A(X)* is indicative in terms of the equivalent of how many agents would perfectly alternate if all the rest were not moving at all, as shown in **Figure 1**. Thus, the outcome of those values does not necessarily mean that 3,338 agents out of 5 were indeed perfectly alternating, yet that the collective behavior is the equivalent of such a turn-taking throughout the whole experiment. Indicatively, some results out of a series of 41 experiments are:



	Final Total ALT	Ratio of PA agents Equivelant
FALT	0,657	3,289
EALT	0,647	3,236
eFALT	0,452	3,364
eEALT	0,442	3,326
CALT	0,446	3,342
AALT	0,388	3,47
	Average Ratio:	3,338

Figure 1. The final estimation of ALT version have an average error of *~0.1* for *5* agents and *10000* episodes. Because the *avg. AltRatio* is *>65%,* it is expected that the exponential versions eFALT & eEALT will be higher than FALT & EALT. Also, AALT shows the highest value as its evaluation has the most abrupt curve.

References

[1] J. Perolat, J.Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation, *Adv. Neural Inf. Process. Syst.* 2017-Decem (2017) 3644–3653.

[2] N. Papadopoulos. Study of turn-taking coordination for nagents in game-theoretic scenarios, with reinforcement learning: Proposal of an evaluation framework of Perfect Alternation Equilibria for multi-agent environments, Universitat de Pompeu Fabra, 2020. https://repositori.upf.edu/handle/10230/46270 .

[3] M.J. Gasparrini, and M. Sánchez-Fibla. Loss Aversion Fosters Coordination in Independent Reinforcement Learners, *Front. Artif. Intell. Appl.* 308 (2018) 307–311. doi:10.3233/978-1-61499-918-8-307.

[4] R.X.D. Hawkins, and R.L. Goldstone. The formation of social conventions in real- Time environments, *PLoS One*. 11 (2016) 1–14. doi:10.1371/journal.pone.0151670.