COllective INtelligence with Sequences of Actions extended abstract[1]

P.J. 't Hoen S.M. Bohte

CWI, Centre for Mathematics and Computer Science P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
{hoen,sbohte@cwi.nl}

Agents in a multi-agent system (MAS) contribute to some part of the collective through their private actions. The joint actions of all agents derive reward from the outside world that is divided among the individual agents where each agent aims to increase its received reward by some form of learning. Unless special care is taken as to how this reward is shared, there is a risk that agents in the collective work at cross-purposes. The COllective INtelligence (COIN) framework [2], as introduced by Wolpert et al., suggests how to engineer (or *modify*) the rewards an individual agents receives for its actions (and to which it adapts to optimize) in *private utility functions* so as to optimize the reward received by the collective. As a case study, we investigate the performance of COIN extended for sequences of actions for representative token retrieval problems found to be difficult for agents using classical Reinforcement Learning (RL). We further investigate several techniques from RL (model-based learning, Q(λ)) to scale application of the COIN framework. The interested reader is referred to [1] for details.

Wolpert et al. propose the **Wonderful Life Utility** (WLU) as a private utility function that is both *learnable* and *aligned* with the global utility G for the joint action ζ of the MAS, and that can also be easily calculated. The WLU for agent η is defined as:

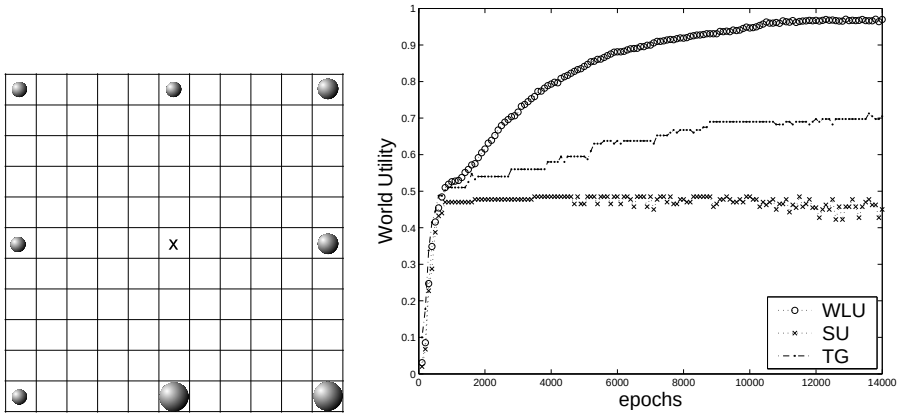
$$WLU_\eta(\zeta) = G(\zeta) - G(CL_{S_\eta^{eff}}(\zeta)) \quad (1)$$

The function $CL_{S_\eta^{eff}}(\zeta)$ "clamps" agent η and returns the utility of the system without the effect of agent η on the remaining agents $\hat{\eta}$ with which it possibly interacts. The WLU hence has a built in incentive for agents to find actions that add to the global utility and also to avoid actions that deduct from the utility of other agents as otherwise the second term in Equation 1 will increase.

In Figure 1(a), as taken from [1], differently valued tokens are placed on the edge of a 11×11 grid and eight agents take five steps in one epoch of learning; they are able to pick up all tokens only if they cooperate perfectly by each focusing on a distinct token. Reward for retrieval of each token is proportional to its size. This problem is representative of a complex set of tasks which must all be completed by one of the agents, but the different tasks have varying priorities.

In Figure 1(b), we show the fitness, the fraction of total value of tokens retrieved, for agents as Q-learners using three types of utility measures. The first

[1] The full version of this paper appeared in [1].



(a) Tokens on the edge

(b) Performance learners

Figure 1: A difficult coordination problem

two utility measures are commonly applied in RL, but may fail to result in a MAS with a high fitness. Agents using a Selfish Utility (SU), i.e. agents are only concerned with their own reward, are attracted to the high token values and ignore the low value tokens. With increasing competition for the high value tokens, the positive reinforcement signal for these targets decreases and the agents become indecisive. For the Team Game (TG), i.e. the total reward is evenly divided over all the agents, a maximum fitness (even after $50,000$ epochs) of ≈ 0.7 is slowly reached as the agents are unable to effectively target a token due to the low signal-to-noise ratio. However, when using the WLU with the extension of a penalty for revisiting states, the agents are able to learn to pick up all the tokens. A fitness of 0.5 is quickly reached and after an agent has chosen "its" token, the extended WLU drives the competing agents to look elsewhere on the grid.

References

[1] P.J. 't Hoen and S.M. Bohte. COllective INtelligence with sequences of actions. In *14th European Conference on Machine Learning*, volume 2837 of *Lecture Notes in Artificial Intelligence*. Springer, 2003.

[2] David Wolpert and Kagan Tumer. Optimal payoff functions for members of collectives. *Advances in Complex Systems*, 4(2/3):265–279, 2001.