Maximally Permissive Reward Machines

Giovanni Varricchione[1], Natasha Alechina[2,1],
Mehdi Dastani[1], and Brian Logan[1,3]

[1] Utrecht University, Utrecht, The Netherlands
{g.varricchione, m.m.dastani, n.a.alechina, b.s.logan}@uu.nl
[2] Open University of the Netherlands, Heerlen, The Netherlands
[3] University of Aberdeen, Aberdeen, United Kingdom

Reward machines were introduced in [4] as a way of defining temporally extended (or "*non-Markovian*") tasks and behaviors, and have been shown to outperform state-of-the-art algorithms in such tasks. However, while learning with a reward machine is guaranteed to converge to an optimal policy *with respect to the reward machine* [5], in general reward machines provide no guarantees that the resulting policy is optimal *with respect to the task* encoded by the reward machine. Moreover, defining reward machines declaratively can be difficult and prone to errors in non-trivial tasks. Illanes et al. in [2] used planning techniques to address the latter issue: given an abstraction of the environment in the form of a planning domain, reward machines are synthesised from a single (sequential or partial-order) plan. Nevertheless, they did not tackle the issue of optimality.

In [6], which has been accepted for publication at ECAI 2024, we have improved on the Illanes et al. approach by synthesising reward machines from the set of all plans to solve the task. We showed that, under certain assumptions, agents trained with our reward machines can learn an optimal policy for the task, rather than just the reward machine as in the approach of Illanes et al. Finally, we provided empirical proof of this by comparing agents trained with our reward machines against agents trained with the reward machines of the approach by Illanes et al. in three tasks from the CraftWorld environment [1]. This is an extended abstract of such paper; we refer the reader to the full paper for formal definitions and a thorough presentation of the results.

As we mentioned above, reward machines (RMs) are used to model tasks that require agents to perform temporally extended behaviours. A reward machine is a finite state automaton that receives in input events that occur in the environment in which the agent acts, updates its internal state, and produces in output a reward signal. To synthesise our "*maximally permissive reward machines*" (MPRMs) we use *planning domains*. A planning domain comprises of a set of planning states and actions, and can be seen as an abstraction of the MDP in which the agent acts: planning actions correspond to sequences of MDP actions which the agent learns to perform in the MDP. Tasks in a planning domain consist in reaching a *goal* planning state from an initial planning state via planning actions. We can synthesise so-called "*plans*" that describe what actions the agent should perform depending on the planning state. While Illanes et al. propose an approach where they use a single plan to synthesise reward machines, our maximally permissive reward machines are synthesised from the entire set of plans to achieve the task, allowing for more flexibility. We formally showed that agents trained with our MPRMs are able to learn policies that obtain higher rewards than agents trained with the RMs of the Illanes et al. approach. Moreover, under certain assumptions, we showed that the reward of the optimal policy of an MPRM-trained agent is equal to the one of the optimal policy for the task.

We evaluated our approach by comparing it against the RMs obtained by the approach of Illanes et al. from [2] in CraftWorld [1], a simplified version of Minecraft. In the full paper we present the results from all the experiments we have performed. Figure 1 shows the MPRM for the task to build a bridge: the agent can decide to build either a grass or iron bridge. In both cases, the agent has to collect wood, but can decide to do so either before or after it has gathered the necessary iron or wood. On the other hand, a single plan for this task would allow the agent to only build either an iron or wood bridge. Figure 2 shows the results of the experiments: as can be seen, agents trained with the MPRM (orange line) achieve higher rewards compared to the other agents from [2] (yellow and blue lines). This shows how, in this task, higher flexibility in achieving the task leads to higher reward. Similarly, in the other tasks we have observed the same; however, in the most complex task, we noticed that the agent trained with the MPRM was slower in converging. We hypothesise that this is due to the fact that, as the agent has more ways to achieve the task, it also has to explore more to find a better policy.

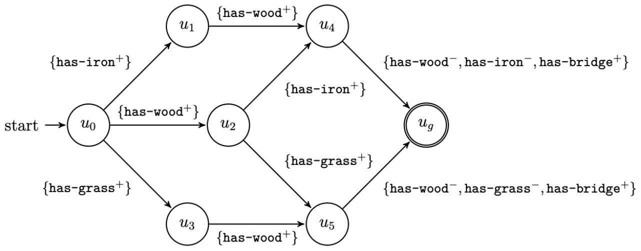

Fig. 1: MPRM for the bridge task.

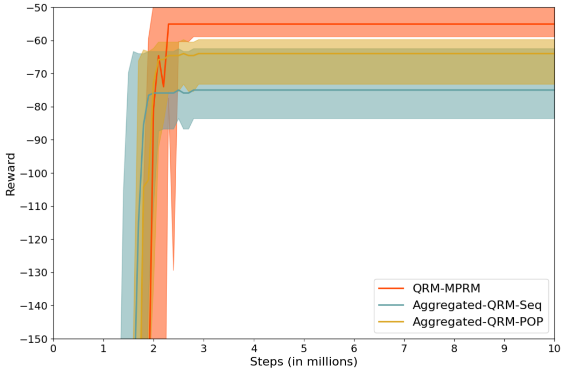

Fig. 2: Results of the bridge task.

To conclude, we have presented an approach where reward machines are synthesised from a set of plans instead of a single one. Our approach allows agents to achieve higher rewards at the cost of a slower convergence. As future works, we would like to investigate the use of *top-k* planning techniques, e.g., [3], to sample a diverse subset of the set of all plans which should hopefully speed up training. Moreover, we would like to test our approach in more complex continuous environments.

References

1. Andreas, J., Klein, D., Levine, S.: Modular multitask reinforcement learning with policy sketches. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2017). pp. 166–175
2. Illanes, L., Yan, X., Toro Icarte, R., McIlraith, S.: Symbolic planning and model-free reinforcement learning: Training taskable agents. In: Proceedings of 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2019). pp. 191–195
3. Katz, M., Lee, J.: K_* search over orbit space for top-k planning. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023). pp. 5368–5376
4. Toro Icarte, R., Klassen, T., Valenzano, R., McIlraith, S.: Using reward machines for high-level task specification and decomposition in reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018). pp. 2107–2116
5. Toro Icarte, R., Klassen, T., Valenzano, R., McIlraith, S.: Reward machines: Exploiting reward function structure in reinforcement learning. Journal of Artificial Intelligence Research **73**, 173–208 (2022)
6. Varricchione, G., Alechina, N., Dastani, M., Logan, B.: Maximally permissive reward machines. In: Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024). To appear.