R5/Sanit Gupta/160100010

August 6, 2020

In this paper, the authors talk about thresholded reward MDPs (TRMDPs). These are MDPs in which one doesn't care about the total reward one gets; instead, one cares only if the total reward is above a certain threshold or not. This setting is clearly applicable to sports like football and cricket where the margin of victory/loss is relevant and the only thing that matters is if one won or lost. For settings like these, it is clear that the optimal policy would be non-stationary (dependent on time remaining). For example, a team, that is behind in the last few minutes of a football match, will have to play aggressively to give themselves a chance of victory even if that might make them extremely susceptible to conceding a goal.

They start by converting the original MDP into a TRMDP. The new description of state must include the time remaining and the total reward received till till the current instance. Unlike previous work, their method gets you a TRMDP that is only polynomially larger than the original MDP. Also, the structure of this TRMDP is such that it allows Value Iteration to be done very efficiently (worst case runtime of $O(|A||S|^2h^2m)$).

The policies learnt by their formulation outperform an agent that learns using the cumulative reward for the football MDP they have made. They also make many similar MDPs with random transition probabilities. Even in cases where the agent was clearly inferior to the opponent across all strategies, using the thresholded reward formulation allows the agent to win more often than not. Also, the non-stationary policies the agent learns are easy to interpret and agree with our intuition. For example, playing defensively when in the lead with little time remaining.

They also talk about various heuristic strategies to reduce the state size while maintaining their performance: uniform-k, lazy-k and logarithmic-k-m.

The lazy-k strategy which uses the TRMDP formulation only for the last k time steps, because more time sensitivity is required when less time is left, works the best empirically.

A few questions I have regarding the paper:

- How do we extend this to a setting where the transition dynamics of a system are unknown?

- In this case, they don't really use intermediate rewards to learn anything. In a case where one doesn't know the transition dynamics, could the intermediate rewards could be leveraged somehow to accelerate learning, instead of giving the agent just the thresholded reward at the end? .
- Why is the difference between the TRMDP and the regular formulation the highest when the capabilities of the teams are similar?