R5/Ishank Juneja/16D070012

February 4, 2020

In the usual Markov Decision Process (MDP) setting the goal of an agent is to maximize the sum of discounted rewards over its lifetime. However, when trying to design an agent that can outperform an opponent in a timed zero-sum game, we must take a different view of the problem. An optimal agent for a zero-sum game maximizes its probability of winning. To win, the only requirement is that the agent scores more points than its opponent. To capture this objective, the paper by McMillen and Veloso sets up the **threshold-rewards (TR) problem**. Under this setup, rather than maximizing cumulative discounted reward over some fixed number of time steps (h), the agent's goal is to maximize $r_{true} = f(r_{intermediate})$. Here r_{true} is the objective that is maximized by the optimal policy π^*_{TR}, f is a function called the *threshold rewards objective* and $r_{intermediate}$ is the usual cumulative reward over agent lifetime.

An interesting contrast between π^*_{TR} in the TR setup and the optimal policy under the usual **maximizing expected reward (MER)** setup is that, π^*_{TR} is in general non-stationary. To illustrate this, the paper considers a simplified robot soccer setup with each time step being associated with one of three states - *FOR* (Agent scores: $\Delta r_{intermediate} = 1$), *AGAINST* (Opponent scores: $\Delta r_{intermediate} = -1$) and *NONE* (Neither score: $\Delta r_{intermediate} = 0$). Further at each time step, the agent's soccer team as a whole can decide to play in one of three formations - *balanced, offensive* or *defensive*. These three formations are interpreted as the possible actions that the agent can take. Each of these actions has a fixed pre-specified probability associated with transitioning to one of the three game states described earlier. For certain instances of transition probabilities, the MER scheme would lead to a policy that always plays the balanced action. However, this scheme will come with a high loss probability. Under the TR framework the optimal policy increases its chances of winning by alternating between the *offensive, defensive* and *balanced* actions.

To obtain this non-stationary optimal policy using well known MDP planning algorithms, the paper synthesizes an associated MDP in such a way that the optimal policy on it would be π^*_{TR}. In simplified terms they do this by incorporating the time index $t \in \{1, \ldots, h\}$ and the intermediate reward $r_{intermediate}$ into the state of this new MDP ($M^{'}$). It is fascinating to note here that, even on $M^{'}$, planning for π^*_{TR} using value-iteration has a run-time that is polynomial in the MDP parameters. On deploying this approach onto the simplified robot-

soccer problem, the agent learns a policy that is more aggressive in its decisions as the game reaches its end. On comparing the performances of the agents, the paper finds that the TR agent is able to win with high probability even against a non-adaptive opponent that is a better scorer.

Even though value-iteration takes at most polynomial time on the problem, a large sized state-space in the original MDP could prove to a significant bottle-neck to planning. To overcome this, the paper proposes some heuristics that decrease the number of times a non-stationary policy is updated, thereby decreasing the state-space of $M^{'}$. The first is *uniform-k* where the policy is adapted only every k time steps. The next is *lazy-k* where the policy is not adapted until the last k time steps. Lastly, there is the *logarithmic-k-m* heuristic that is a hybrid approach between the other two. Empirically, it is observed that the *lazy-k* heuristic performed the best on the robot soccer problem.

A question I have is -

- As the paper points out, a drawback of the present TR model is that it does not consider that the agent may be adaptive. I was wondering if a convergent non-stationary policy would give us any real advantage over an adversary/opponent that was adaptive, since the opponent could change their policy to counter ours.