R4/Ishank Juneja/16D070012

February 1, 2020

Unlike machine-learning models learnt using supervised learning, reinforcement learning methods do not have the limitation that complete training happen in a learning phase prior to their deployment. However, even in most reinforcement learning solutions, the focus is often on converging to a single stationary solution to the problem. Most often the reason learning is not continued after deployment of the agent is its slow adaptation. This makes significant progress unfeasible during online interaction.

The paradigm of tracking is a well accepted method for performing well on non-stationary problems. This paper by Sutton and others, makes the case for using tracking even for stationary problems. Particularly when an agent-environment interaction has a large state space with temporal coherence. The authors describe two such stationary environments where tracking based solutions perform better than the best convergent solution.

The first illustrative example is that of the Black and White world. The problem setup consists of an agent undertaking a random walk. At every position on its path, the agent could choose (through an action) to observe the colour (either Black or White) of the square it is on. The agent's task is to predict the colour of the square it is about to observe. Depending on the accuracy of the agent's prediction, it incurs a loss and subsequently its probability estimate is updated. The environment has temporal coherence since both colours occur in contiguous blocks of squares. A logistic sigmoid model is used to assign a probability to choosing a colour. The model relies on a single parameter w_t. When the problem is viewed over a long horizon, the probability of predicting either black or white is 0.5, since both coloured squares are equal in number. However, due to the temporal coherence present in the problem, this is not true over shorter time scales. To exploit this, the paper uses a large learning rate in the gradient descent updates of w_t to effectively create a *tracking* based solution.

Next, the paper considers a 5×5 game of Go. In this example the convergent solution approach would be to seek the best possible global policy that can be learnt. But considering the complexity of the state-space in Go, a solution that computes the best policy starting from the games current state would likely perform better. The authors point to a few situation where despite using the same representation, a tracking agent is able to identify the correct move while the converging agent fails to do so. This feat becomes possible since on starting

from a particular state, it becomes computationally feasible to roll out entire games within a reasonable inference time. From recent advances in Go playing by *alphaGo*, it has been seen that even if complete roll-outs can not be performed due to high complexity, partial roll-outs padded with estimates also work well when applying this approach.

Lastly the paper discusses how tracking can be used to assess the viability of meta-learning methods on certain problems. To illustrate this feature of tracking the authors compare the performance of the IDBD step size adaptation algorithm between temporally coherent and incoherent Black and White worlds. Simulations show that the step size adaptation meta-learning method provides no significant advantage in the temporally incoherent setting.

A specific question I have is,

- At the end of the conclusion, the authors suggest that tracking may help in providing a justification to the choice of sequence of tasks in meta-learning. What do they mean by this?