Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper: When Waiting Is Not an Option: Learning Options with a Deliberation Cost
Date: 23 August 2024

Make sure your critique addresses the following points:

1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem

Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 Problem Statement

The usual RL algorithms assume action decision-making is always instantaneous and cost-free. In actual real-life environments, this may not be true. For example, in an autonomous driving scenario, delayed decisions for actions can sometimes lead to sub-optimal and maybe even unfavorable outcomes. This paper aims to address the above problem so that the agent can make timely decisions and a trade-off between deliberation and performance.

2 Key Contributions

Bounded rationality is crucial for understanding both physical and artificial systems. This paper proposes bounded rationality for constructing temporal abstractions with the goal of reducing computing time. "Good" options are those that help to learn faster and are close to the optimal policy. The paper implements an optimization problem in an "option critic" framework:

- Introduction of "deliberation costs"
- Using deliberation costs in the options framework to determine a proper trade-off between switching to a different option and performing the same option
- Implementation of a gradient-based learning algorithm

3 Proposed Algorithm/Framework

3.1 Deliberation Cost Model

The key idea is to penalize switching from one option to another. As the option gets longer, the cost rate decreases. For a very long option, with cost = η, the cost per step turns out to be $\eta/d = (1-\gamma\kappa)\eta$

$$D_\theta(z) = \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\gamma^t\tilde{c}(Z_t, A_t, Z_{t+1}) \mid Z_0 = z\right] \tag{1}$$

Goal/objective function formulation :

$$\max_\theta J_\alpha(\theta)\,, \text{where } J_\alpha(\theta)\sum_{s,o}\alpha(s,o)\left(Q_\theta(s,o) - \eta D_\theta(s,o)\right) \tag{2}$$

$$\tilde{V}^c_\theta(z)\sum_{a,z'}\pi_\theta(a \mid s,z)\tilde{P}_\theta(z' \mid z,a)\left(\tilde{r}(s,a,z') - \eta\tilde{c}(z,a,z') + \gamma\tilde{V}^c_\theta(z')\right) \tag{3}$$

3.2 Switching cost and its interpretation as a margin

$$Q^c_\theta(s,o) = \sum_a \pi_\theta(a \mid s,o) \left(r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \left[Q^c_\theta(s',o) - \beta_\theta(s',o) \left(A^c_\theta(s,o) + \eta \right) \right] \right). \tag{4}$$

$$\frac{\partial J_\alpha(\theta)}{\partial \theta_\beta} = \mathbb{E}_{\alpha,\theta} \left[-\frac{\partial \beta_\theta(S_{t+1}, O_t)}{\partial \theta_\beta} \left(A^c_\theta(S_{t+1}, O_t) + \eta \right) \right] \tag{5}$$

Here η is for the correction margin. It is a baseline for how good an option is supposed to be. Increasing its value reduces the gap in the advantage function $\beta_\theta(s',)$ is the mean of a Bernoulli random variable over the two possible outcomes, switching or continuing (1 or 0),

3.3 Computation horizon

$$J^{\gamma,\tau}_\alpha(\theta) = \sum_{s,o} \alpha(s,o) \left(Q^\gamma_\theta(s,o) - \eta D^\tau_\theta(s,o) \right) \tag{6}$$

$$\begin{aligned} \frac{\partial D^\tau_\theta(s,o)}{\partial \theta_\beta} = & \frac{\partial}{\partial \theta_\beta} \sum_a \pi_\theta(a \mid s,o) \sum_{s'} P(s' \mid s,a) \left(c_\theta(s',o) \right. \\ & \left. + \tau \left[(1 - \beta_\theta(s',o)) D^\tau_\theta(s',o) + \beta_\theta(s',o) \sum_{o'} \mu_\theta(o' \mid s') D^\tau_\theta(s',o') \right] \right) \end{aligned} \tag{7}$$

$$\frac{\partial J^{\gamma,\tau=0}_\alpha(\theta)}{\partial \theta_\beta} = \gamma \mathbb{E}_{\alpha,\theta} \left[-\frac{\partial \beta_\theta(s',o)}{\partial \theta_\beta} \left(A_\theta(s',o) + \eta \right) \right] \tag{8}$$

$$\frac{\partial J^{\gamma=\tau}_\alpha(\theta)}{\partial \theta_\beta} = \gamma \mathbb{E}_{\alpha,\theta} \left[-\frac{\partial \beta_\theta(s',o)}{\partial \theta_\beta} \left(A^c_\theta(s',o) + \eta \right) \right] = \frac{\partial J_\alpha(\theta)}{\partial \theta_\beta} \tag{9}$$

The discount factor λ is used for truncating the sum of costs $\lambda = 1$ is a Temporal regularisation - predictions become more complex - as our agent needs to account for longer horizons

3.4 A2OC

The asynchronous Advantage Option Critic framework helped in stable online learning. Parallel agents are used to sample from different states.

4 Advantages and Conclusions of the Algorithm

It was observed that the effect of training on deliberation cost- the agent continues to take the same option for a long time. Certain options terminate - **Key decision points**. By using deliberation costs, the paper addresses a way to reduce computation time and unwanted decision while preserving performance and taking close to the optimal set of actions.
The concept of deliberation cost could be taken beyond computation efficiency. It can also include factors such as "missed opportunity." If agent spends too much time deliberating, it might miss out on taking better actions. We also need to address the errors and capacity in representing the states through its values, etc., and how deliberation cost could be used to address these.