The option-critic architecture

Pierre-Luc Bacon and Doina Precup
Reasoning and Learning Lab (RLLAB), McGill University

Summary

- We extend the policy gradient theorem (Sutton, McAllester, et al., 2000) to the options framework.
- We provide two new results for computing the intra-option policy gradients as well as the termination gradients.
- Their algorithmic implementation gives rise to the option-critic architecture.

Options framework

Options (Sutton, Precup, and Singh, 1999) formalize the idea of temporally extended actions (also sometimes called skills or macro-actions).

A Markovian option $\omega \in \Omega$ is a tuple $(\mathcal{I}_\omega, \pi_\omega, \beta_\omega)$:
- Initiation set $\mathcal{I}_\omega \subseteq \mathcal{S}$
- Policy $\pi_\omega$ (stochastic or deterministic)
- Termination function $\beta_\omega : \mathcal{S} \rightarrow [0, 1]$.

Augmented state space

Even with an MDP structure and Markov options, the induced flat process over primitive actions is not Markovian. We then need to consider the gradient of $V_\omega(s) \equiv Q_\omega(s, w)$, were $\hat{S} \equiv \mathcal{S} \times \Omega$:

$$\frac{\partial}{\partial \theta} Q_\omega(s, \omega) = \sum_a \pi_\omega(a | s) Q_\omega(s, \omega, a)$$

$$Q_\omega(s, \omega, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) U(s', \omega)$$

$$U(s, \omega) = (1 - \beta_\omega(s)) Q_\omega(s, \omega) + \beta_\omega(s) V_\omega(s)$$

Intra-option policy gradient theorem

Given a set of fixed Markov options and a fixed policy over them, in the start-state formulation,

$$\frac{\partial}{\partial \theta} Q_\omega(s, \omega) = \mathbb{E}_{(s, w) \sim \pi_\omega \beta_\omega} \left\{ \sum_a \pi_\omega(a | s) \frac{\partial}{\partial \theta} \pi_\omega(a | s) Q_\omega(s, \omega, a) | \omega_0, \omega_0 \right\}$$

Termination gradient theorem

Given a set of fixed Markov options and a fixed policy over them, in the start-state formulations,

$$\frac{\partial}{\partial \theta} Q_\omega(s, \omega) = \mathbb{E}_{(s, w) \sim \pi_\omega \beta_\omega} \left\{ \frac{\partial \beta_\omega(s)}{\partial \theta} (V_\omega(s) - Q_\omega(s, \omega)) | \omega_0, \omega_0 \right\}$$

Learning intra-option policies with fixed terminations

We first studied the behavior of the intra-option policy gradient algorithm when the initiation sets and subgoals are fixed by hand. In this case, options terminate with probability 0.9 in a hallway state and four of its incoming neighboring states. We chose to parameterize the intra-option policies using the softmax distribution with a one-hot encoding of state-action pairs as basis functions.

Learning both intra-option policies and terminations

We used the same softmax parametrization as in the previous experiment but chose to represent the termination functions using the hyperbolic tangent function. We found that option-critic converges faster than an MDP-based actor-critic approach with a single softmax policy over primitive actions. When overlaid to the grid layout, a plot of the termination probabilities for option 0 (fig. 3b) shows that option-critic learned to terminate around a hallway state, agreeing with our intuition.

Opportunities and future work

- Option-critic opens the way to end-to-end learning of RL agents.
- It enables joint study of temporal and state representation learning.

Ongoing work:

- Function approximation: provide an analogue to the feature compatibility condition (Sutton, McAllester, et al., 2000)
- Two-timescale convergence analysis (Konda and Tsitsiklis, 2004)
- Regularization: we are developing a bounded rationality approach that favors learning fast and robust options. Come see us at the NIPS 2015 Bounded Rationality workshop.

References