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 triple $\langle \mathcal{I}_\omega, \pi_\omega, \beta_\omega \rangle$:

- Initiation set $\mathcal{I}_\omega \subseteq \mathcal{S}$
- Policy π_ω (stochastic or deterministic)
- Termination function $\beta_\omega : \mathcal{S} \to [0, 1]$.

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 consisder the gradient of $V_\Omega(\tilde{s}) \equiv Q_\Omega(s, w)$, were $\tilde{\mathcal{S}} \equiv \mathcal{S} \times \Omega$.

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

$$Q_U(s, \omega, a) = r(s, a) + \gamma \sum_{s'} \mathrm{P}(s' \mid s, a)\, U(s', \omega)$$

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

Intra-option policy gradient theorem

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

$$\frac{\partial Q_\Omega(s_0, \omega_0)}{\partial \theta} = \mathop{\mathbb{E}}_{(s,\omega)\sim\mu} \left\{ \sum_a \frac{\partial}{\partial \theta} \pi_{\omega,\theta}(a \mid s)\, Q_U(s, \omega, a) \;\middle|\; s_0, \omega_0 \right\}$$

Termination gradient theorem

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

$$\frac{\partial Q_\Omega(s_0, \omega_0)}{\partial \theta} = \mathop{\mathbb{E}}_{(s,\omega)\sim\mu} \left\{ \frac{\partial \beta_{\omega,\theta}(s)}{\partial \theta} \left(V_\Omega(s) - Q_\Omega(s, \omega) \right) \;\middle|\; s_0, \omega_0 \right\}$$

Option-critic architecture

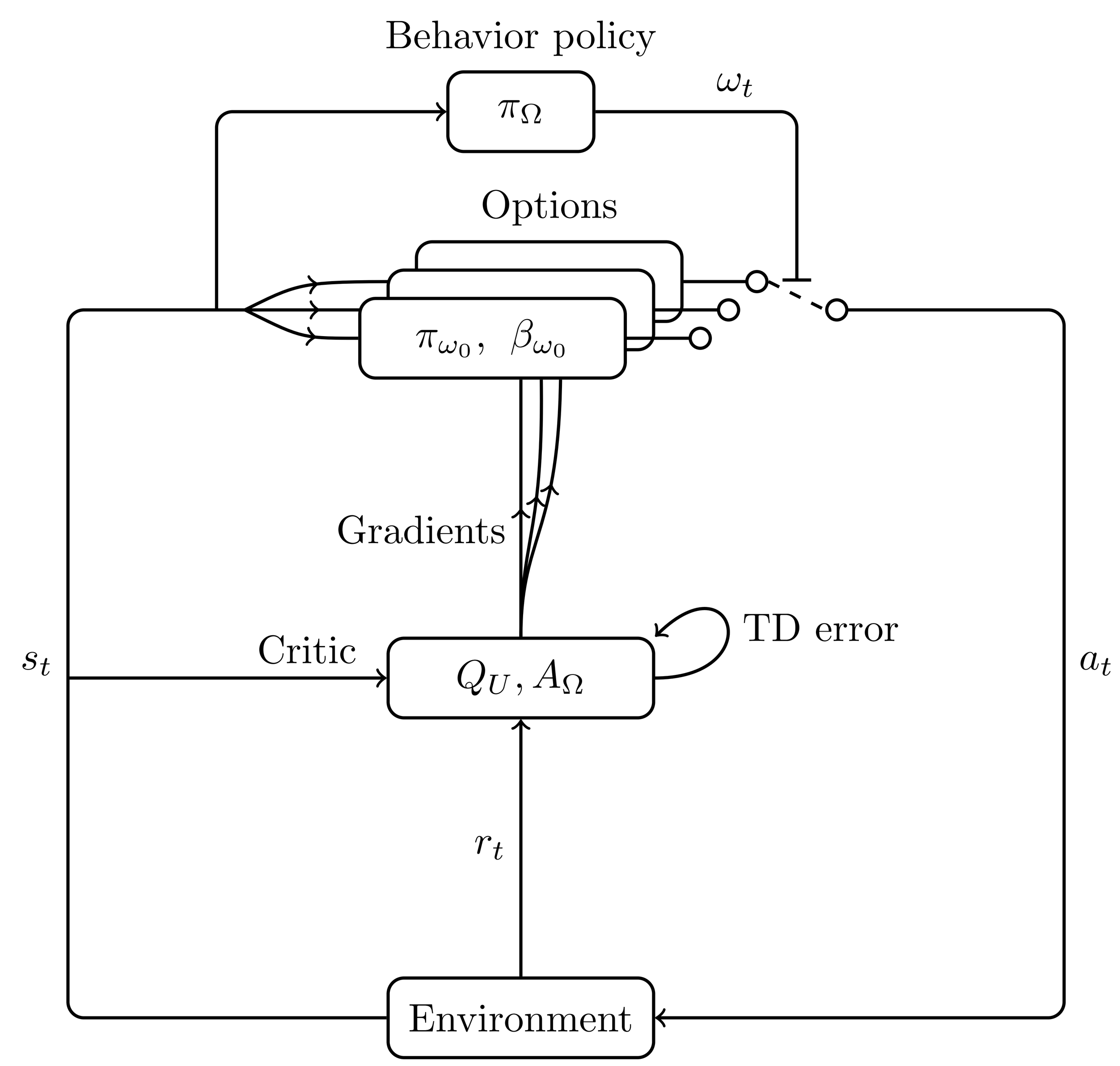
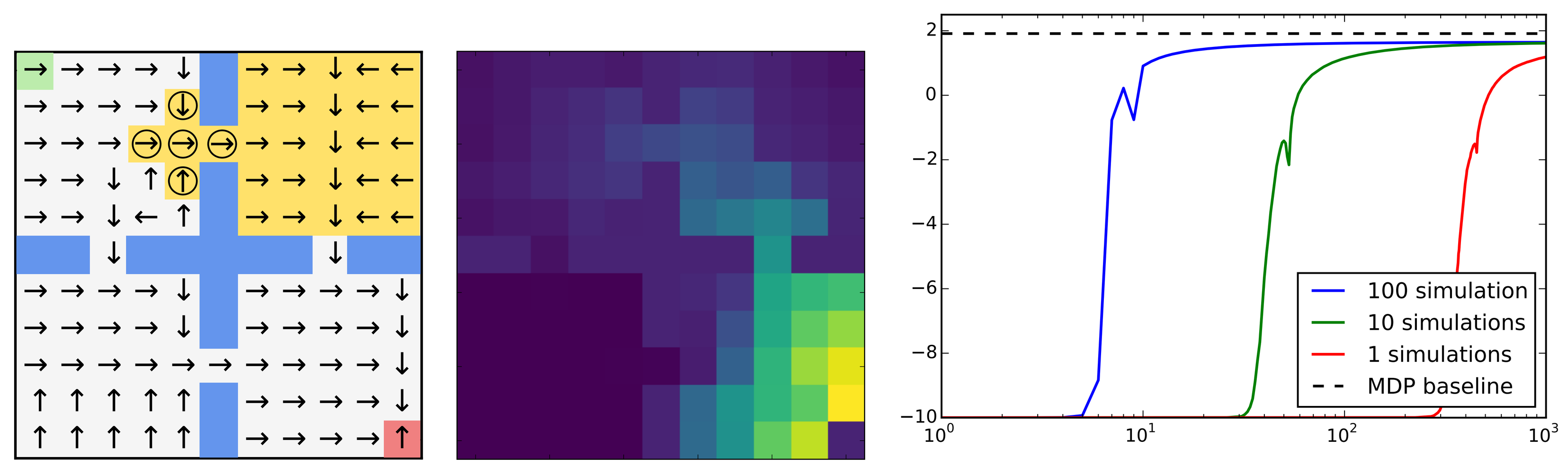


Figure : The option-critic architecture consists of a set of options, a policy over them and a critic. Gradients can be derived from the critic for both the intra-option policies and termination functions. The execution model is suggested pictorially by a *switch* $\perp$ over the *contacts* $\multimap$. Switching can only take place when a termination event is encountered.

Learning intra-option policies with fixed terminations

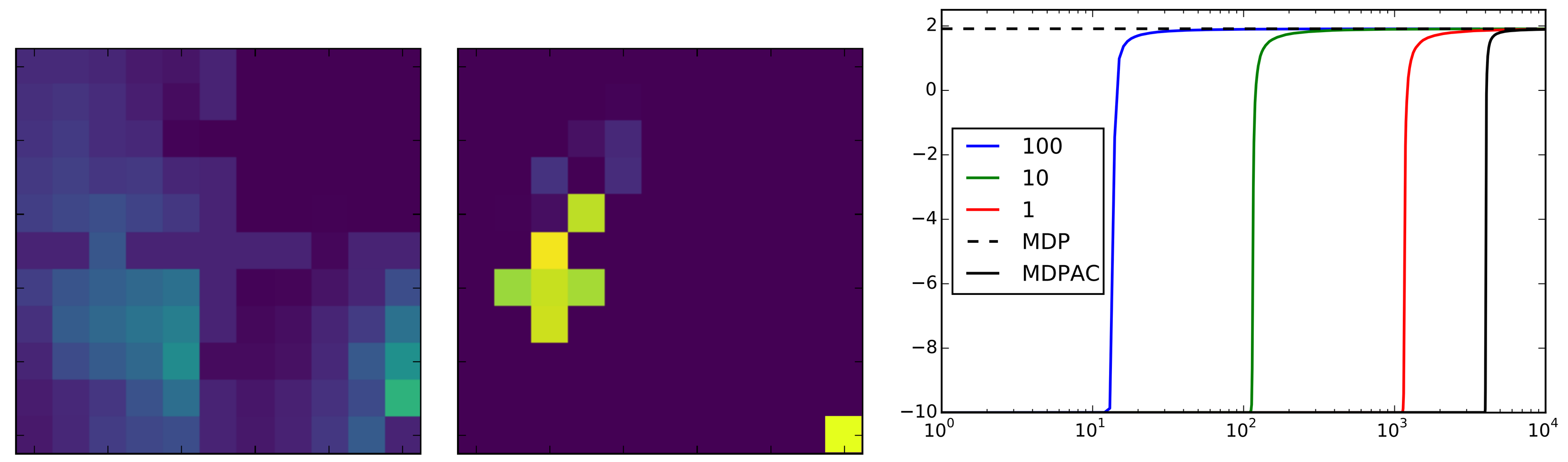


(a) Four-rooms domain (b) Value function over options (c) $V_\Omega(0)$ vs optimization steps

Figure : Four-rooms domain (Sutton, Precup, and Singh, 1999) overlaid with a deterministic optimal MDP policy. The blue color is for *wall* cells, the green cell corresponds to the initial state and the red one is a goal state. The initation set of option 1 is highlighted with the color yellow. The circles represent the subgoal states for option 0

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 parametrize 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



(a) Value function over options (b) Learned termination probabilities (c) $V_\Omega(0)$ vs optimization steps

Figure : Simultaneous learning of option policies and termination functions

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 an 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 developping a bounded rationality approach that favors learning *fast* and *robust* options. Come see us at the NIPS 2015 Bounded Rationality workshop.

References

- Richard S. Sutton, David A. McAllester, et al. (2000). "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In: *Advances in Neural Information Processing Systems 12*, pp. 1057–1063
- Richard S. Sutton, Doina Precup, and Satinder P. Singh (1999). "Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning". In: *Artif. Intell.* 112.1-2, pp. 181–211
- Vijay R. Konda and John N. Tsitsiklis (2004). "Convergence Rate of Linear Two-Time-Scale Stochastic Approximation". In: *The Annals of Applied Probability* 14.2, pp. 796–819