Safe Option-Critic: Learning Safety in the Option-Critic Architecture

Arushi Jain, Khimya Khetarpal, Doina Precup

Reasoning and Learning Lab (McGill University), Mila Lab
Montreal, Canada

Reinforcement Learning

In RL, we have state s, action a, reward r, policy $\pi(a|s)$, transition probability $P(s'|s,a)$ and discount factor γ.

- Return: $\sum_{t=0}^{\infty} \gamma^t r_{t+1}$
- Value of state: $V(s) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s]$
- State-action value: $Q(s,a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s, a_0 = a]$
- One-step temporal difference (TD) error:
  $\delta(s,a) = r(s,a) + \gamma P(s'|s,a)\,\pi(a'|s')\,Q(s',a') - Q(s,a)$

RL Lab McGill University

Mila

Background I: Options framework

An option $\omega \in \Omega$ is a triple of:

- Initiation set: I_ω
- Internal policy: π_ω
- Termination condition: β_ω

Let $\Theta = \{\theta, \nu\}$, where following represents parameter for:

- θ: Internal policy $\pi_{\omega,\theta}$
- ν: Termination condition $\beta_{\omega,\nu}$

The update for Q value [Bacon et al., 2017]:

$$Q(s,\omega,a) = r(s,a) + \gamma P(s'|s,a)\{(1-\beta_{\omega,\nu}(s))Q_\Theta(s,\omega) + \beta_{\omega,\nu}(s)V_\Omega(s)\}$$

RLLab McGill University

Mila

Background II: Safety

Unintended *or* ***harmful behavior*** *that may emerge from machine learning systems when we specify the wrong objective function, are not careful about the learning process, or commit other machine learning-related implementation errors.* [Amodei et al., 2016]

Controllability: Negation of variance in the TD error, controlling uncertainty in the value of a state-option pair [Gehring and Precup, 2013].

Contribution of this work

Safe Option-Critic (SOC) framework provides a novel mechanism to learn end-to-end safe options.

- Derived a policy-gradient style update for a new safe objective function

$$\max_{\Theta} J(\Theta|d),$$
$$\text{where } J(\Theta|d) = \mathbb{E}_{(s_0,\omega_0)\sim d}[Q_\Theta(s_0,\omega_0) + \psi\, C_\Theta(s_0,\omega_0)]$$

Here $C_\Theta(s_0,\omega_0) = -\,\mathbb{E}_{a\sim\pi_{\omega,\theta}(a|s)}\left[\delta^2(s,\omega,a)\right]$ is the controllability, ψ is the regularizer on controllability, d is initial state-option distribution.

McGill University

Results: Update for gradient

θ update for internal policy of option

$$\mathbb{E}[\frac{\partial \log(\pi_{\omega,\theta}(a|s))}{\partial\theta} Q_{U,\Theta}(s,\omega,a) - \frac{\partial \log(\pi_{\omega_0,\theta}(a_0|s_0))}{\partial\theta} \psi\delta^2(s_0,\omega_0,a_0)]$$

Interpretation: Take better primitive action with a regularization on minimizing TD error inside an option.

ν update for termination function of option

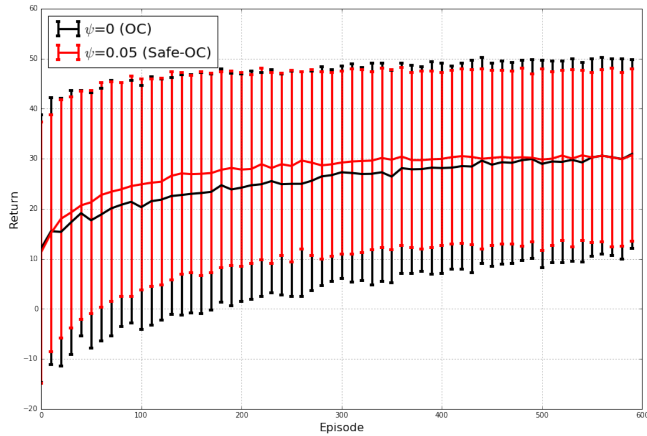
$$\mathbb{E}[\frac{\partial \beta_{\omega,\nu}(s')}{\partial\nu}(Q_\Theta(s',\omega) - V_\Omega(s'))]$$

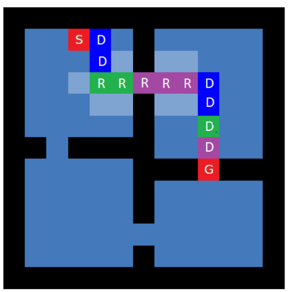
Interpretation: Termination unaffected by the controllability.

RLLab McGill University

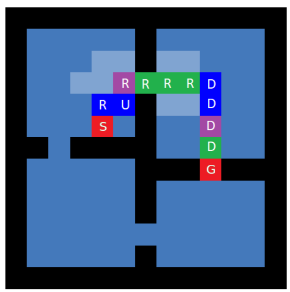
Mila

Results: Four room environment

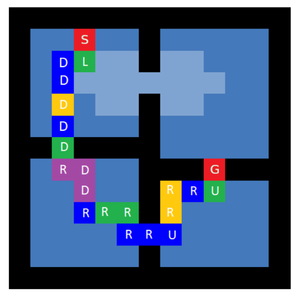




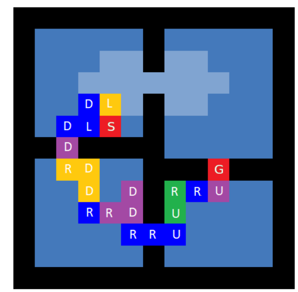
(a) OC



(b) OC

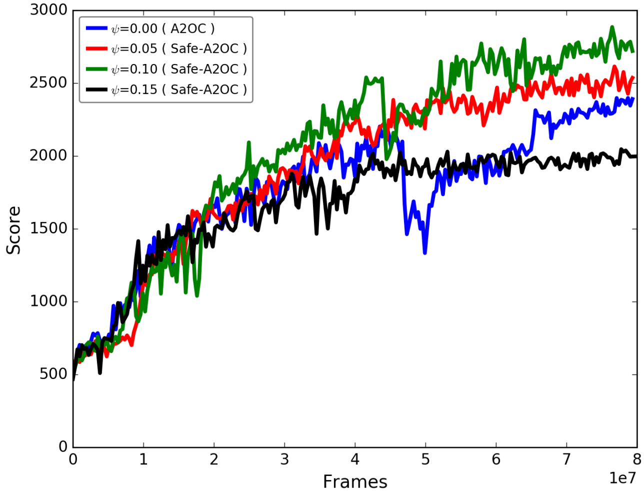


(c) Safe-OC

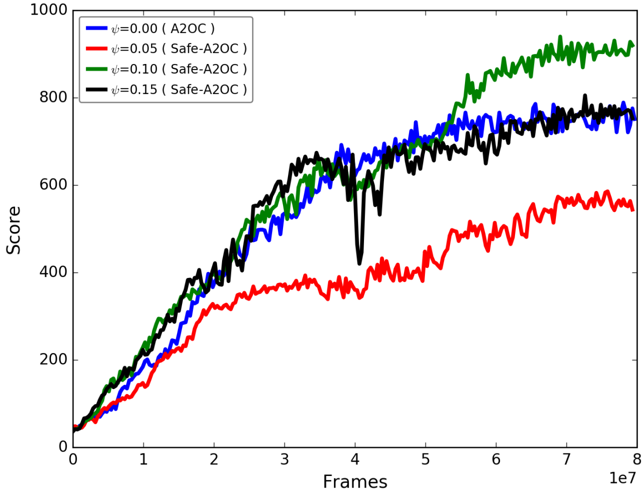


(d) Safe-OC

Results: Arcade Learning Environment



(a) MsPacman



(b) Amidar

RLLab McGill University

Mila

Conclusion

- Novel work to incorporate **safety** in **end-to-end options** learning.
- Safe-OC framework is **scalable** to include non-linear function approximation.

Future Work:

- Using **n-step return** calculation at option switching (current work return calculation limited until option terminates).
- To learn **initiation set** while learning safe options.
- Notion of safety to **different levels of hierarchy**.

RLLab McGill University

Mila

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. (2016).
Concrete problems in AI safety.
CoRR.

Bacon, P.-L., Harb, J., and Precup, D. (2017).
The option-critic architecture.
In *AAAI*, pages 1726–1734.

Gehring, C. and Precup, D. (2013).
Smart exploration in reinforcement learning using absolute temporal difference errors.
In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 1037–1044.

RLLab
McGill University

Mila