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

RLLab McGill University

Mila

Abstract

- Novel work on introducing **safety in hierarchical reinforcement learning** (Option-Critic architecture).
- Safety introduced by **regularizing variance in the TD error**.
- Demonstrate effectiveness of framework in **tabular** and **Arcade Learning Environments** (ALE).

Reinforcement Learning

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

- 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:
  $\delta(s,a) = r(s,a) + \gamma P(s'|s,a)\,\pi(a'|s')\,Q(s',a') - Q(s,a)$

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 intra-option Bellman update for Q value:

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

Safety Definition

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.* [1]

Our notion of safety -
Controllability: Negation of variance in the TD error, controlling uncertainty in the value of a state-option pair [2].

Contribution

Safe Option-Critic (SOC) framework provides a novel mechanism to learn end-to-end safe options in Option-Critic Architecture [3].

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

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

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.

Results: Updates for Gradient

- Gradient update for θ **parameter** of **internal policy of option**

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

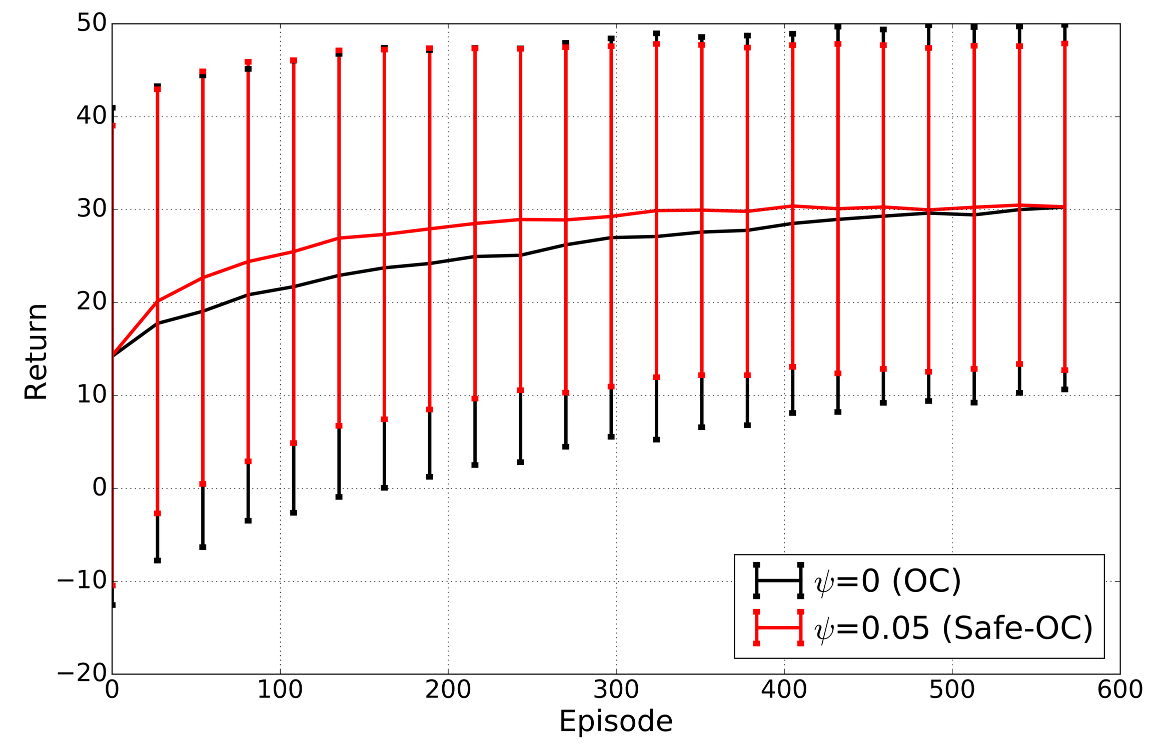
Take better primitive action with regularization on minimizing variance in TD error.

- Gradient update for ν **parameter** of **termination function of option**

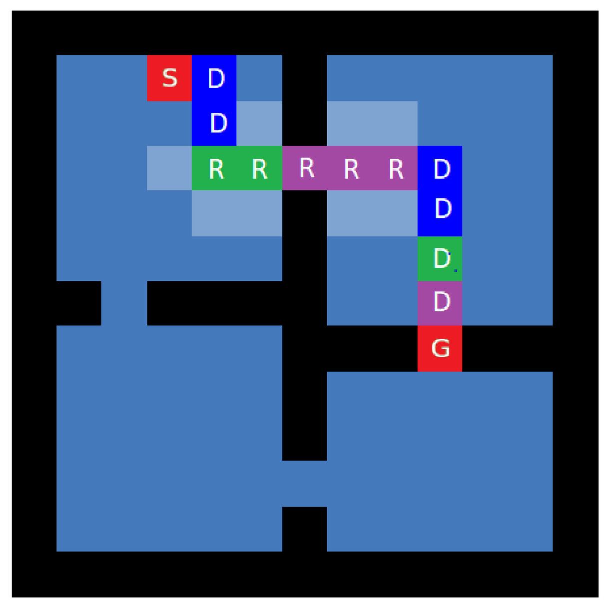
$$\mathbb{E}[\frac{\partial\beta_{\omega,\nu}(s')}{\partial\nu}(Q_\Theta(s',\omega) - V_\Omega(s'))]$$

Termination condition is unaffected by addition of the controllability factor.

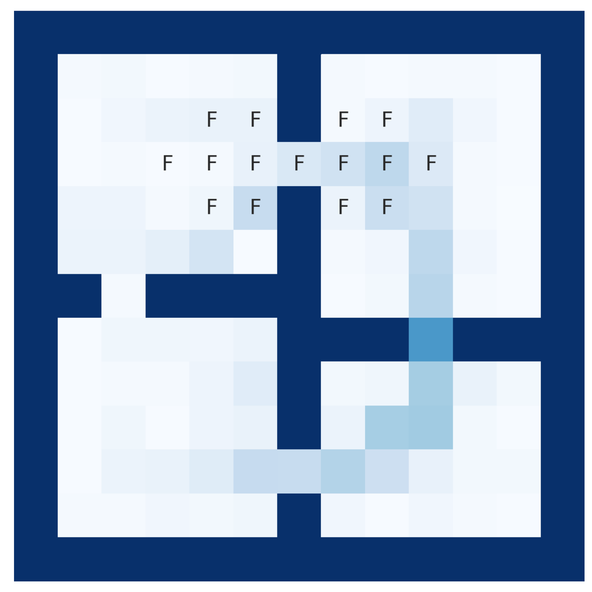
Experiments: Four Rooms Grid World



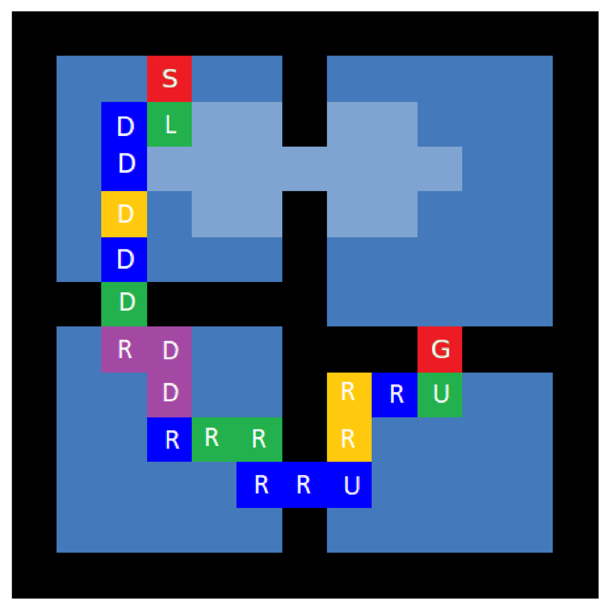
Learning curve with 4 options averaged over 200 trials



Option Critic

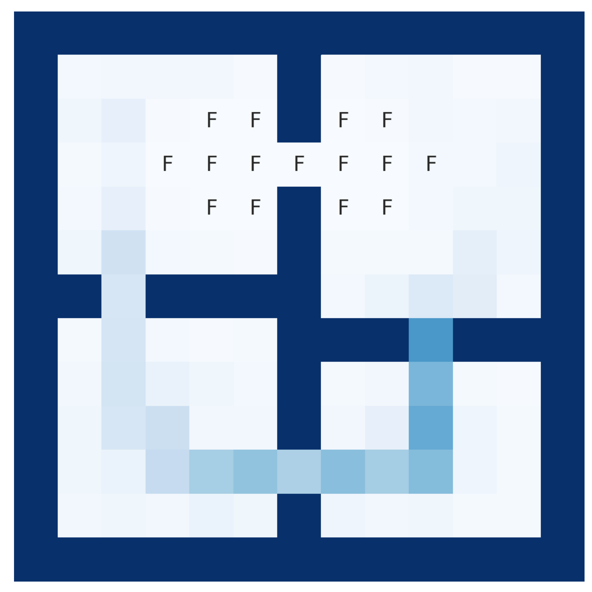


Option Critic



Safe Option Critic

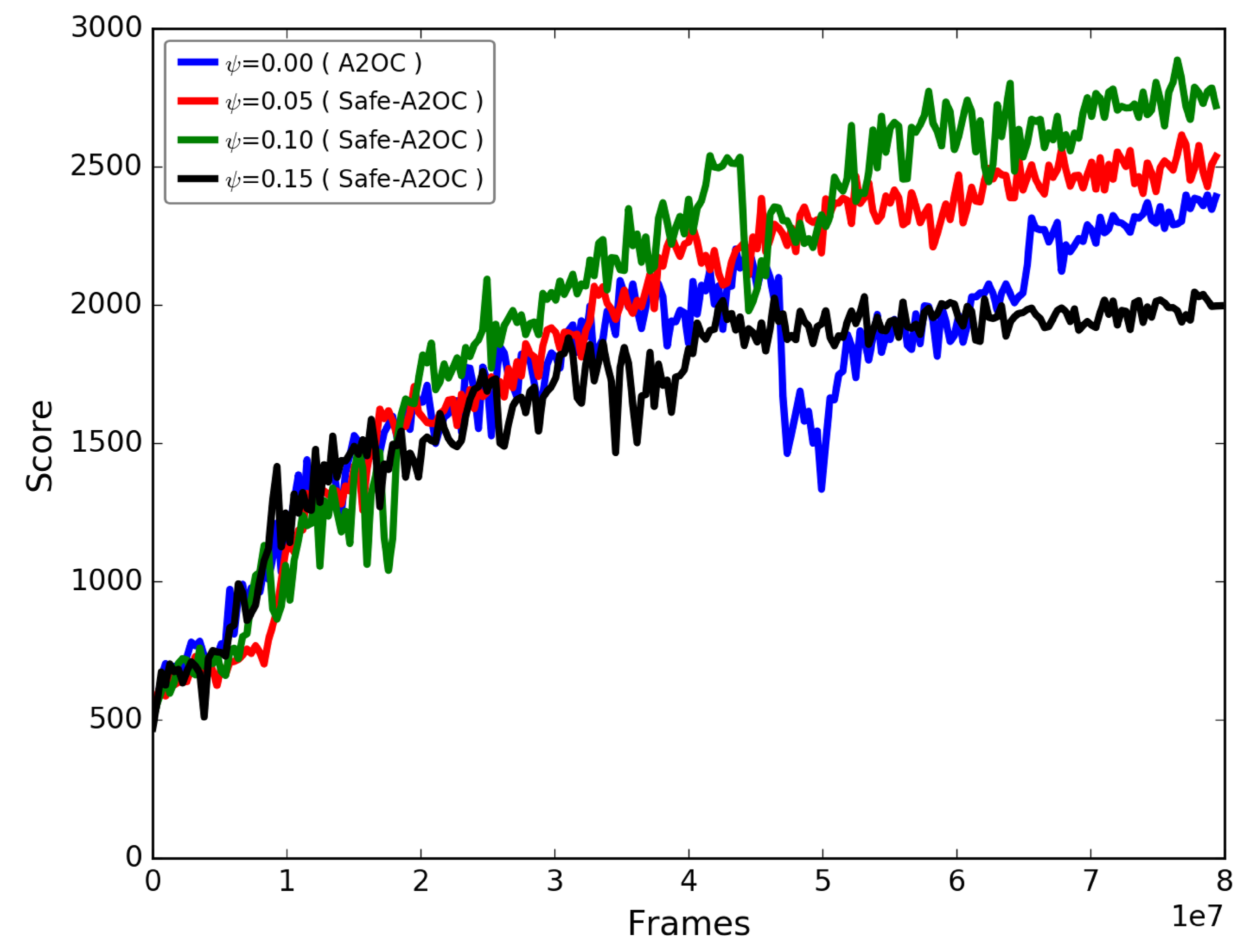
Sampled Policies



Safe Option Critic

State Frequency

ALE - MsPacman



Learning curve with 4 options in MsPacman

Conclusion & Future Work

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

Future Work

- Using **n-step return** calculation (current work is one-step return).
- Notion of safety to **different levels of hierarchy**.

References

[1] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *CoRR*, 2016.

[2] C. Gehring and D. Precup, "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, pp. 1037–1044, 2013.

[3] P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture.," in *AAAI*, pp. 1726–1734, 2017.