Safe Policy Learning with Constrained Return Variance

Arushi Jain, Doina Precup

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

RLLab
McGill University

Mila

What is Reinforcement Learning?

Learning by interacting with an environment to achieve a goal.

$$s_0, a_0, r_1, s_1, a_1, r_2, \ldots$$

$$V_\pi(s) = \mathbb{E}_\pi[\sum_t^\infty \gamma^t r_t | S_t = s]$$

$$= \mathbb{E}_\pi[r_t + \gamma V(S_{t+1}) | S_t = s]$$

RLLab McGill University

Mila

What is Hierarchical Reinforcement Learning?

Learning to perform a task in a hierarchical fashion using combinations of skills/options [Sutton et al. 1999]

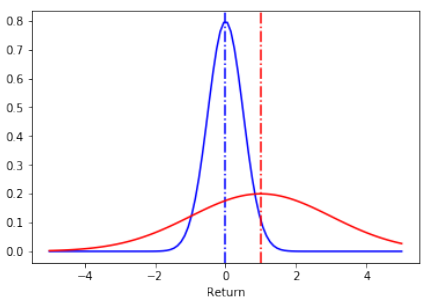
Option o: 1. intra-option policy, 2. termination condition, 3. policy over options

RLLab McGill University

Mila

Why Safety?

Safe RL *is a way of learning policies/behavior that not only maximize the expected return for reasonable performance but also respect certain safety constraints while learning or/and testing phase.*



Which is a better policy? Red? Blue?

RLLab McGill University

Mila

Safe Option Critic

Novel safe hierarchical policy learning in SOTA **Option-Critic Architecture** [Bacon et al. 2017]

Safe Objective Function

$$J_{\text{Safe}}(\theta) = \mathbb{E}_{(s,o)}[V_\theta(s,o) - \psi \underbrace{\sigma_\theta(s,o)}_{\text{Constraint}}]$$

θ: [intra-option policy, termination condition, policy over options]
$\sigma(s,o) = \mathbb{E}[\delta_t^2 + \gamma^2 \sigma(S_{t+1}, O_{t+1}) | S_t = s, O_t = o]$: Variance in Return
δ_t: TD error
ψ: regularizer controlling *risk-seeking* or *risk-sensitive* behavior.

RLLab McGill University

Mila

Contribution of this work

Automatic approach for learning a **safe-hierarchical-policy** in option-critic (OC) where **regularization** is placed on the **variance in return**.

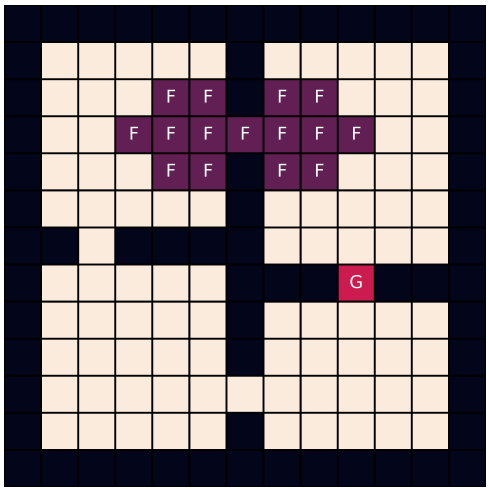
The **Safe OC** is a scalable solution

- It is an online, model-free and continual learning approach.
- No prior knowledge required about the environment - no need for knowing what safe or unsafe.
- Can be applied to general continuous state-action space and scales well to tasks in Mujoco environments.

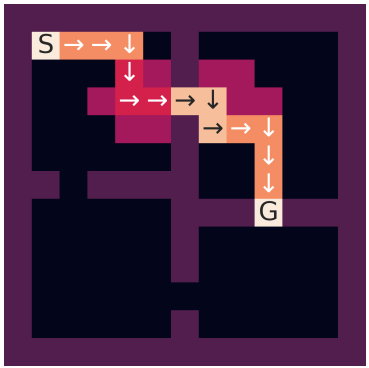
RLLab McGill University

Mila

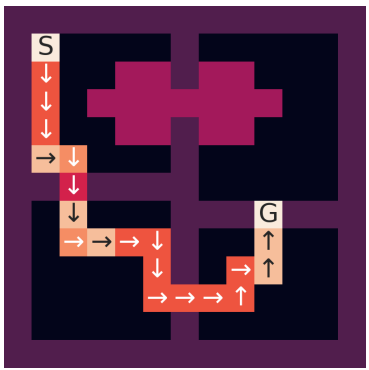
Discrete Grid World



(a) Four Rooms Environment



(b) OC

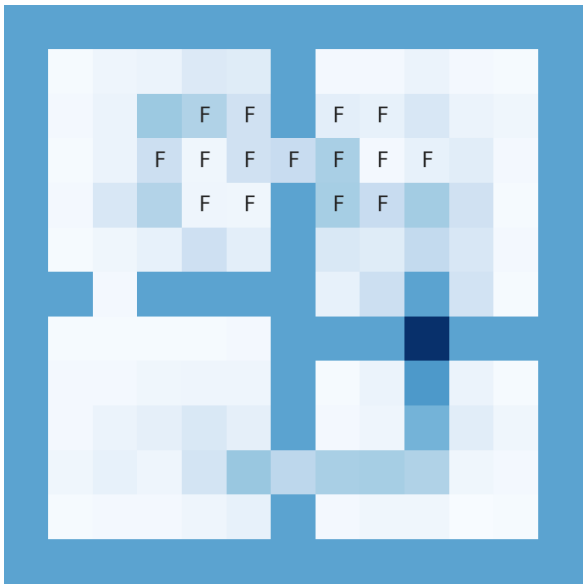


(c) Safe-OC

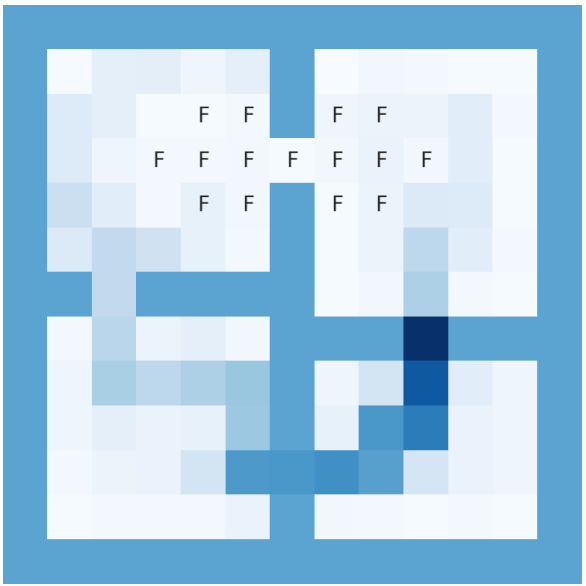
RLLab McGill University

Mila

Discrete Grid World: Frequency Plots



(a) OC

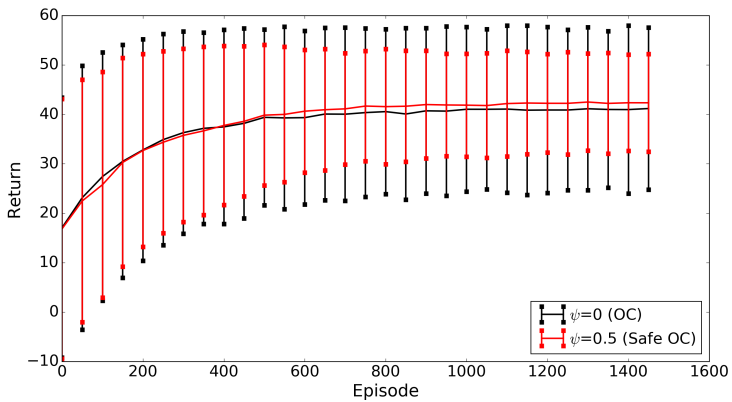


(b) Safe-OC

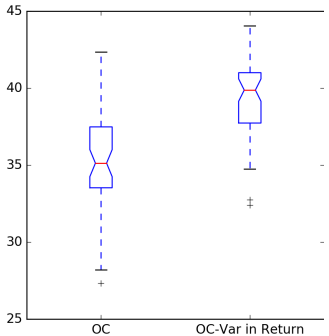
RLLab McGill University

Mila

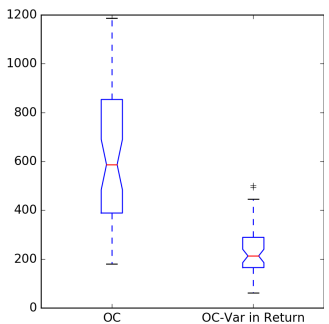
Discrete Grid World



(a) Learning Curve
red curve → Safe Policy
black curve → Unsafe Policy



(b) Return Mean



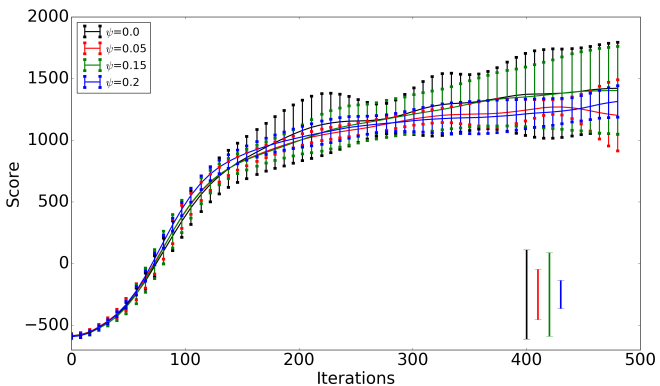
(c) Return Variance

RLLab McGill University

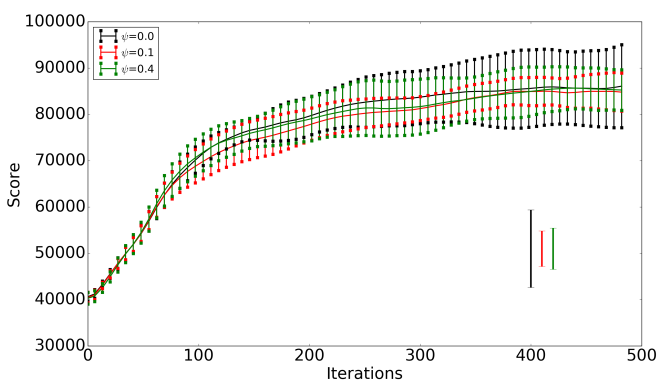
Mila

Mujoco environment: Continuous state-action

Added safety in proximal policy option critic (PPOC) [Klissarov et al. 2017] using constraint variance in return.



(a) Half Cheetah



(b) Humanoid Standup

Video `https://sites.google.com/view/safeoc/home`.

RLLab McGill University

Mila

Conclusion

- Novel **Safe** approach of learning policy in **Option-Critic** style methods.
- Constrained unsafe regions by **regularizing** the direct estimate of **variance in the return**.
- Scalable framework, comparable results as PPOC in Mujoco environments.

Future Work:

- Variable ψ value for different options/skills introducing diversity in behavior.
- More results!

RLLab McGill University

Mila

Conclusion

- Novel **Safe** approach of learning policy in **Option-Critic** style methods.
- Constrained unsafe regions by **regularizing** the direct estimate of **variance in the return**.
- Scalable framework, comparable results as PPOC in Mujoco environments.

Future Work:

- Variable ψ value for different options/skills introducing diversity in behavior.
- More results!

RLLab McGill University

Mila