Learning Options using Constrained Return Variance

Arushi Jain[1], Doina Precup[1,2]

[1] Mila and McGill University, [2]Google Deepmind, Montreal

Overview

Problem: Modelling **safe temporal abstractions** that can avoid regions of state-space with high uncertainty.
Solution: We model a hierarchical reinforcement learning algorithm that minimizes the effects of *model uncertainty* on the expected return in addition to maximizing it.
Main Contributions:

- Propose a new objective in **Option-Critic** framework which uses **variance in the return** as a regularizer.
- Use the above objective to derive the policy-gradient for automatically learning "risk averse" options.
- Experimentally demonstrate the effective of algorithm in tabular and Mujoco environments.

Background

Options:
MDP $= \{S, A, r, \gamma, \mathbb{P}, \}$. We use discounted optimality criteria here.
An option $w \in W$ is a triple of - {initiation set I_w; internal policy π_w; termination condition β_w}. The intra-option Bellman update for Q value:

$$Q(s, w, a) = r(s, a) + \gamma \mathbb{P}(s'|s, a)\{(1 - \beta_w(s))Q(s', w) + \beta_w(s)V_W(s')\}$$

We define **safe** behaviour as the ability of the agent to *avoid regions with high model uncertainty.*

Our Contribution

Safe OC

Taking inspiration from [1], we derived the variance in return for a given augmented state (state-option) space $z \in Z$ as:

$$\boldsymbol{\sigma(z, a)} = \mathbb{E}_{\pi,\mu}[\boldsymbol{\delta_t^2 + \gamma^2 \sigma(Z_{t+1}, A_{t+1}) | Z_t = z, A_t = a}]$$

where $\mu(w|s)$ is policy over options and δ_t is the single-step TD error.

We define the new objective function which desires to maximize the mean performance, but also, minimize the variance in the return of the policy for **Safe OC** architecture as:

$$\boldsymbol{J}_{\text{Safe}}(\boldsymbol{\Theta}) = \mathbb{E}_{d\sim(s_0,w_0)}[Q_\Theta(s_0, w_0) - \underbrace{\boldsymbol{\psi}\, \sigma_\Theta(s_0, w_0)}_{\text{Regularization Term}}\ ]$$

where-
ψ: is regularization constant controlling *risk-sensitive* behavior,
Θ: is the vector of parameterized internal policy, policy over options and termination condition.

Experiments - Tabular Environment

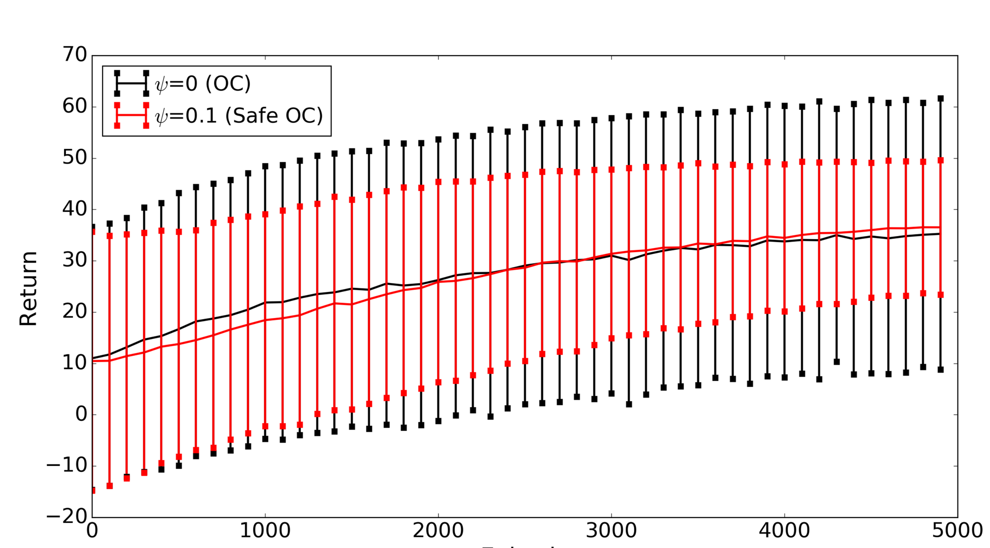
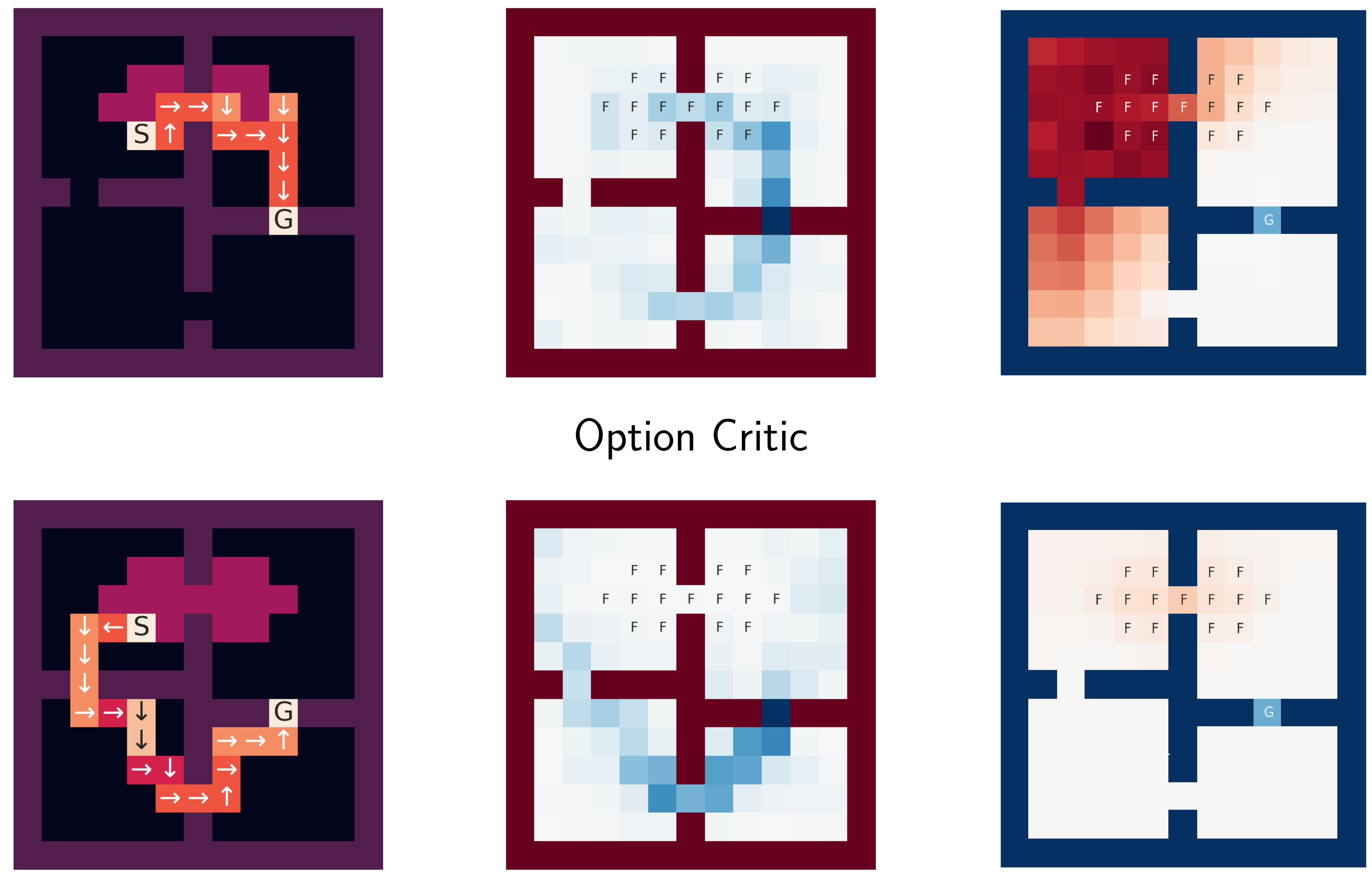


Figure: Return in FourRooms Environment



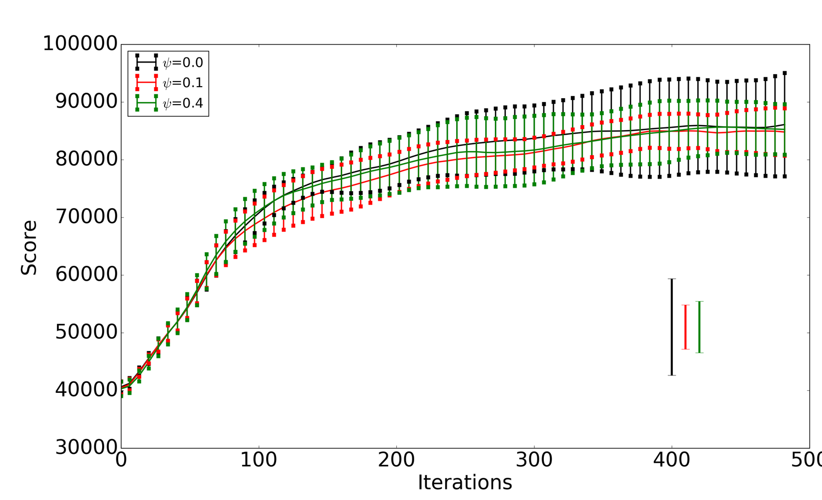
Option Critic

Safe Option Critic

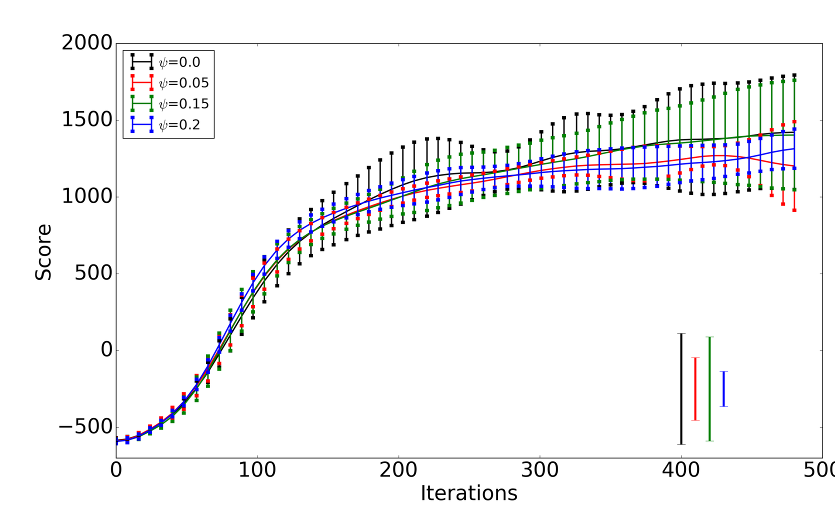
Policy **State Frequency** **Variance**

The purple patch above denotes the unsafe frozen region F in one of the hallways which safe policy learns to avoid. The switch in colors in sampled policy denotes the 4 options.

Experiments - Mujoco Environment



(a) Humanoid Standup



(b) Half Cheetah

Added safe architecture on PPOC [2] framework which is inspired from PPO in primitive actions.

Conclusion & Future Work

- Proposed a **novel** safe hierarchical policy learning framework for **Options** where the **regularization** is placed on the **variance in return**.
- Its on online, generic and **scalable** approach which also includes **non-linear function approximations**.

Future Work

- Learn **diverse skills/options** using mix of risk sensitive/averse policies.

References

[1] C. Sherstan, D. R. Ashley, B. Bennett, K. Young, A. White, M. White, and R. S. Sutton, "Comparing direct and indirect temporal-difference methods for estimating the variance of the return.," in *UAI*, pp. 63–72, 2018.

[2] M. Klissarov, P.-L. Bacon, J. Harb, and D. Precup, "Learnings options end-to-end for continuous action tasks," *arXiv preprint arXiv:1712.00004*, 2017.

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

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