Safe Actor-Critic

Arushi JAIN[1], Ayush JAIN[2], Doina PRECUP[1]

[1] Reasoning and Learning Lab (McGill University), Mila Lab, Montreal, and [2] SPORTLOGiQ, Montreal

Abstract

In this paper we propose a safe policy learning framework in the actor-critic style. We base the safety criteria on regularizing the variance of return in a learned policy. We estimate the variance of λ-return directly using temporal difference (TD) approach [1]. We first demonstrate the effectiveness of our approach in the four rooms grid world environment, and then present the results on four environments with continuous action tasks in Mujoco domain using distributed proximal policy optimization (DPPO) framework. The proposed algorithm outperforms the baselines in all the environments with a significant reduction in the standard deviation of the scores.

- Novel work on introducing **safety in policy gradient style** algorithms - Safe Actor-Critic.
- Safety introduced by **regularizing variance in the return**.
- Demonstrate effectiveness of framework in **tabular** and four **Mujoco Environments**.

Safety Definition

Unintended *or* **harmful behavior** *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.* [2]

Our notion of safety -
Constraining the **variance of return** $\sigma(s)$ using direct method of estimating the variance of λ-return using new Bellman operator. [1]

$$\sigma_\pi(s) = \mathbb{E}_\pi[\delta_t^2 + \bar{\gamma}\sigma(s_{t+1})|s_t = s]$$

where, $\bar{\gamma} = \gamma^2\lambda^2$ and $\delta_t = r_{t+1} + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)$ is the TD error.

Contribution

Safe Actor-Critic: Integrated the notion of safety by constraining the variance of return [1] with actor-critic style methods to automatically learn a safe policy.

- Derived a **policy-gradient theorem for safe-AC framework** which would help to avoid the states with inconsistent behavior.
- The new **safe objective function** is:

$$J(\theta) = \mathbb{E}_{s_0\sim d}[V(s_0) - \psi\sigma(s_0)]$$

Here d is the initial state distribution and $\psi \in \mathbb{R}$ is the **regularizer** for controlling the amount of the variance in return and θ is parameter for the policy.

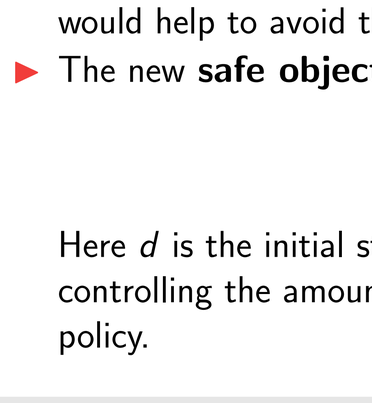
Results: Updates for Gradient

- Update for θ **parameter** of **policy**

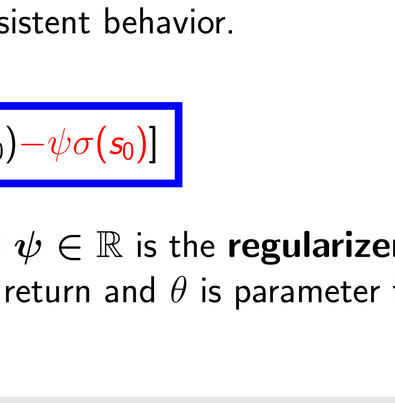
$$\theta \leftarrow \theta + \alpha_\theta \frac{\partial \log(\pi(a|s,\theta))}{\partial\theta}\big[Q(s,a) - \underbrace{\psi\,\sigma(s,a)}_{\text{Regularization Term}}\big]$$

- The policy is learned such that is maximizes the sum of discounted return (as usual) but also constrains the variance of return given by $\sigma(s,a)$.

Experiments: Sampled Policies from Tabular FourRoom Env



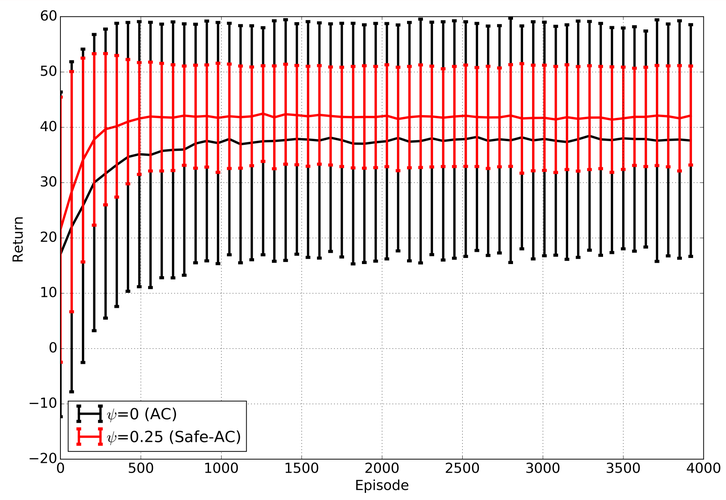
Actor Critic (AC)



Safe Actor Critic (Safe AC)

Here purple colored region is unsafe region with variable reward. Rest of states are normal states. Safe agent, learns to avoid unsafe region from states space.

Return plot for FourRoom Env

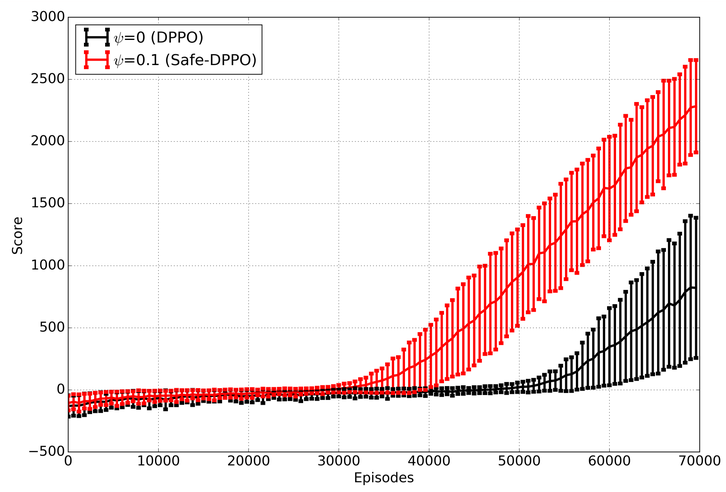


Learning curve for both safe and vanilla AC averaged over 50 trials. With safety (red curve), a significant reduction in standard deviation along with faster learning is observed.

Experiments: Non-Linear Function Approximation

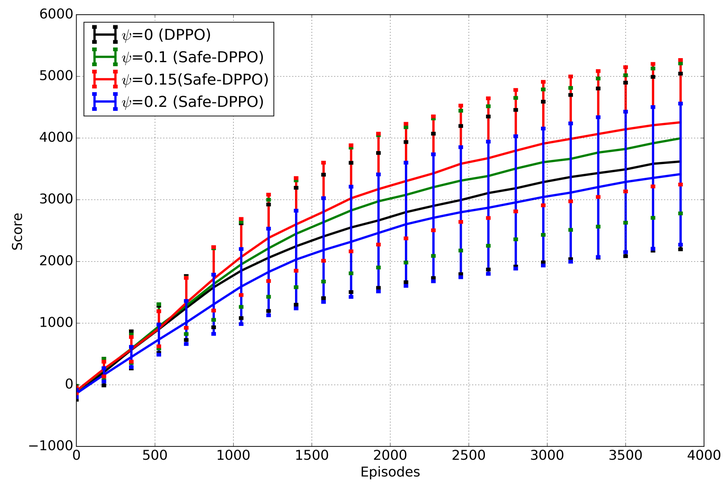
- Experiments on Mujoco to prove **scalibility** of safety model with **non-linear function approximation**.
- Used **state-of-the-art DPPO framework** to extend safety on top of it.
- With safety constraint in DPPO framework, the **boost in performance** was achieved in terms of **faster learning** and significant **reduction in standard deviation** of score across multiple trials.

Ant Env (Mujoco)



Learning curve for both safe (red) and vanilla (black) DPPO averaged over 5 trials. Safe architecture leads to better mean performance with reduction in the standard deviation of the score.

HalfCheetah Env (Mujoco)



Learning curve for both various value of ψ for safe and vanilla DPPO averaged over 20 trials. Best performance of safety is reached with $\psi = 0.15$ (red). Safety leads to better mean performance with reduction in the standard deviation of the score compared to vanilla DPPO (black).

Conclusion & Future Work

- Novel work to incorporate **safety in actor-critic** style methods by constraining the direct estimation of variance of the return.
- Safety framework is **scalable** to include non-linear function approximation.

Future Work

- Experiment with variable ψ value along the time scale rather than constant ψ value.

References

[1] C. Sherstan, B. Bennett, K. Young, D. R. Ashley, A. White, M. White, and R. S. Sutton, "Directly Estimating the Variance of the λ-Return Using Temporal-Difference Methods," *ArXiv e-prints*, Jan. 2018.

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