Safe Actor-Critic

Arushi Jain, Ayush Jain, Doina Precup

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

RLLab
McGill University

Mila

What is safety?

Prevention from ***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]

RLLab McGill University

Mila

Safe in Safe Actor-Critic (SAC)

- Many different approaches of incorporating safety.
- **Safe** : Using a constrained based optimization strategy where regularization is placed on the variance of the return.
- Higher the variance in return $\longrightarrow$ Higher would be uncertainty in the value of state.

RL Lab McGill University

Mila

Variance in return

Approaches to estimate the variance in return:

- Indirect methods: second order moment methods
  $Var(R) = \mathbb{E}[R^2] - (\mathbb{E}[R])^2$
- **Direct methods** : **Bellman operator** for estimating the variance ($\sigma(s)$) [Sherstan et al., 2018]

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

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

RLLab McGill University

Mila

Contribution of this work

Automatic approach for learning a **safe-policy** using actor-critic style methods where **constrained** is placed on the **variance of the return** using direct method.

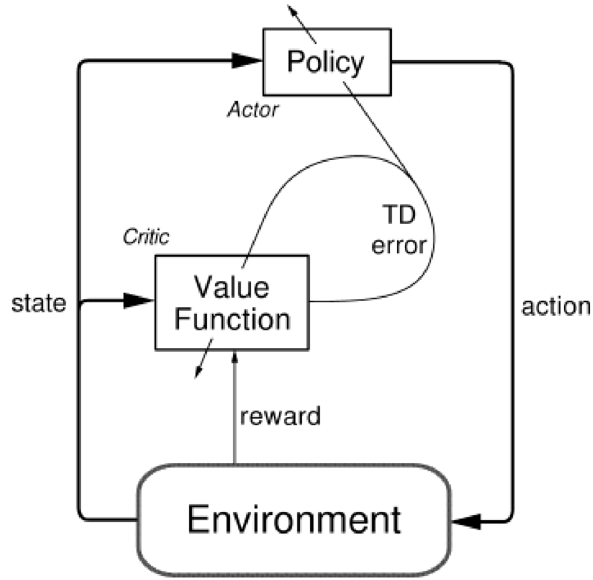
The **Safe Actor-Critic** is a scalable solution

- 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.
- SAC approach leads to stable solution and in many tasks leads to faster learning.

RLLab McGill University

Mila

Actor-Critic Architecture (Sutton 1984)



actor improvement : improves current policy
critic evaluation : evaluate current policy by bootstrapping the value

(image source: `cs.wmich.edu`)

RLLab McGill University

Mila

Objective function

Constrained based optimization

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

$\sigma(s)$: variance in a state s
$V(s)$: value of a state s
ψ: regularizer for maintaining trade-off between expectation and variance
d: initial state distribution
θ: parameter for policy $\boldsymbol{\pi_\theta(a|s)}$

RLLab McGill University

Mila

Results: Update for gradient

θ update for policy

$$\mathbb{E}\left[\frac{\partial \log(\pi_\theta(a|s))}{\partial \theta}\{Q(s,a) - \underbrace{\psi\sigma(s,a)}_{\text{Regularization Term}}\}\right]$$

Interpretation: Take better action that improve Q value but also minimize the variance σ in the return caused by that action.

RLLab McGill University

Mila

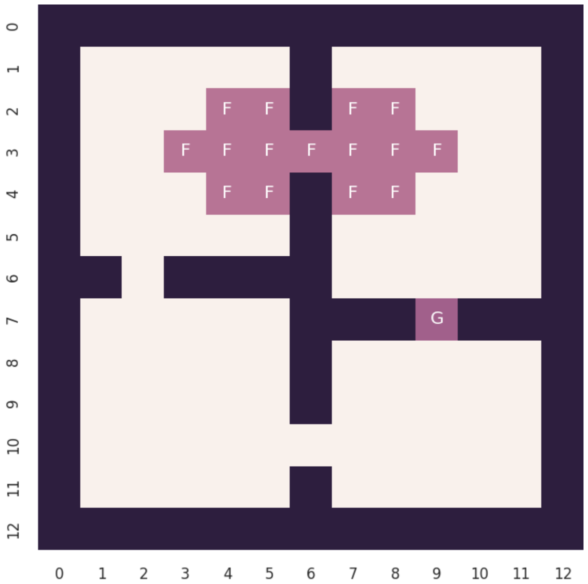
Results: Update for gradient

θ update for policy

$$\mathbb{E}\left[\frac{\partial \log(\pi_\theta(a|s))}{\partial \theta}\{Q(s,a) - \underbrace{\psi\sigma(s,a)}_{\text{Regularization Term}}\}\right]$$

Interpretation: Take better action that improve Q value but also minimize the variance σ in the return caused by that action.

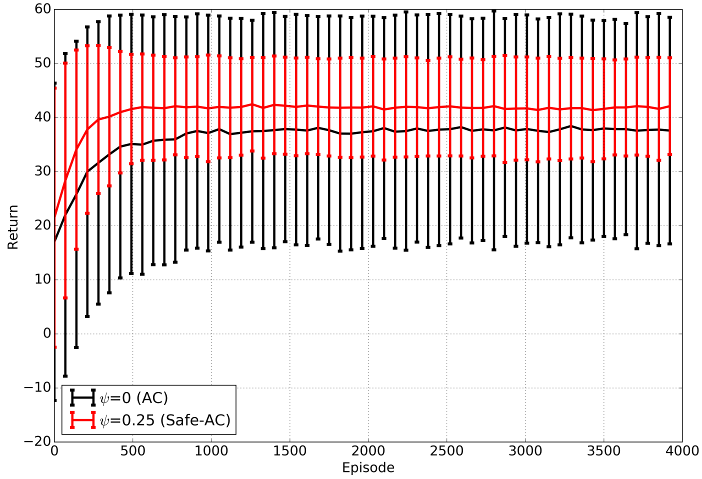
Results: Tabular



F : Frozen states $\longrightarrow$ Unsafe states (with variable reward)

G : Goal state

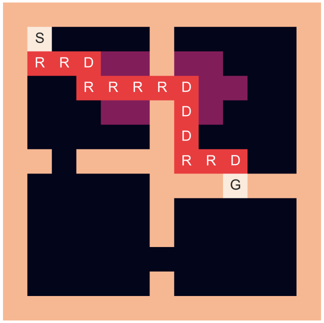
Results: Tabular



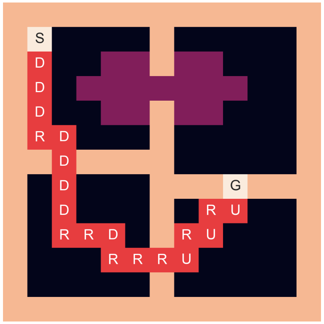
(a) Learning Curve

red curve → Safe Policy

black curve → Unsafe Policy



(b) AC



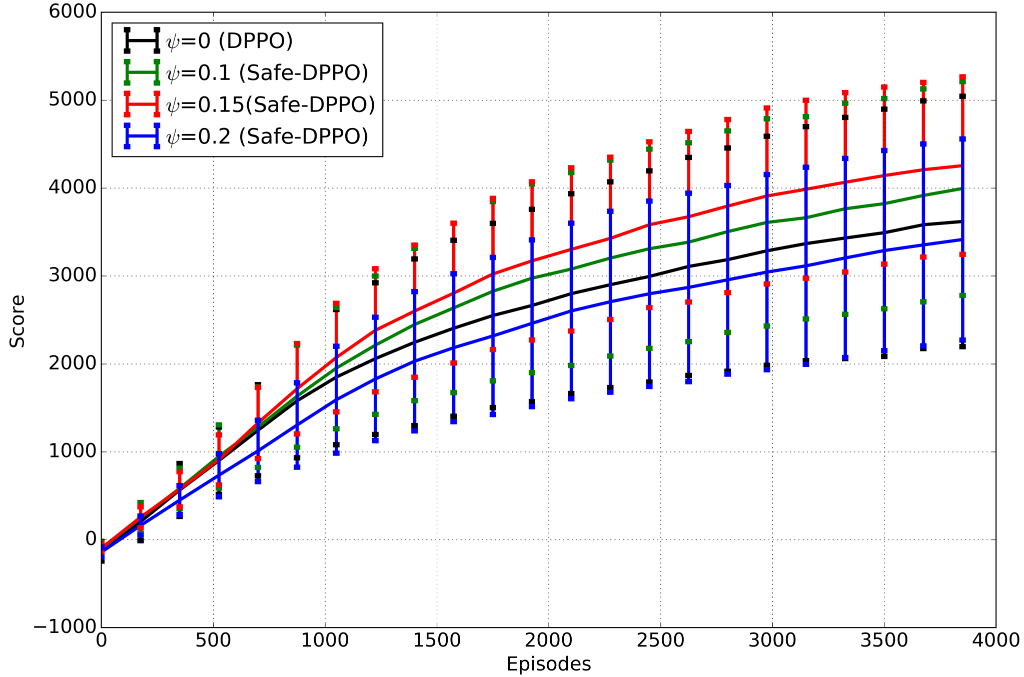
(c) Safe-AC

RLLab McGill University

Mila

Results: Mujoco environment

Added safety in distributed proximal policy optimization (DPPO) using constrained on the variance of return.

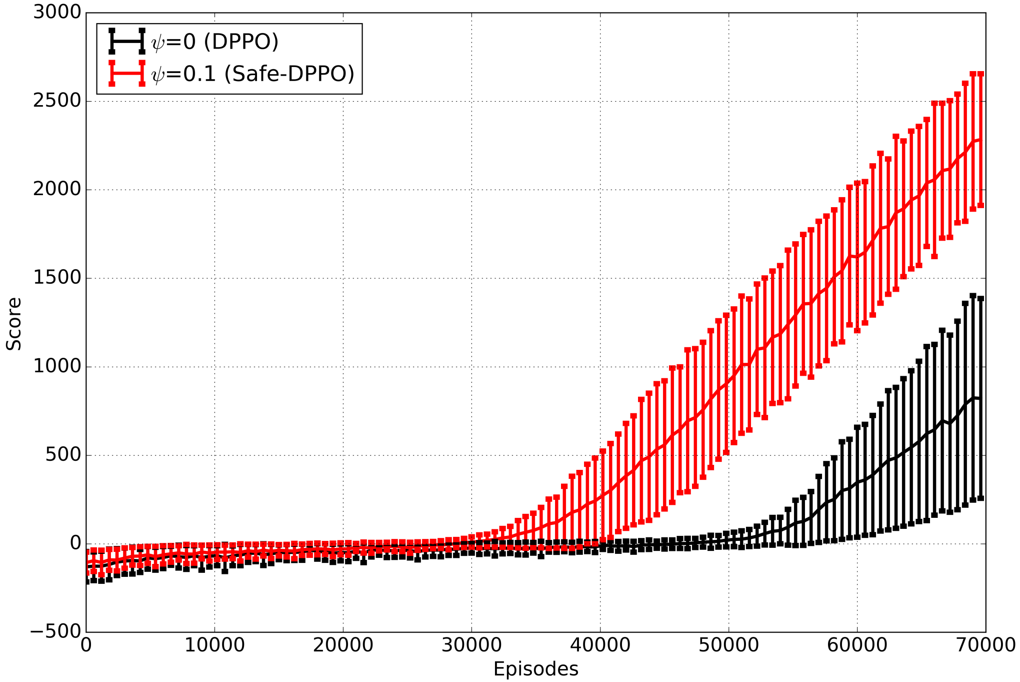


(a) HalfCheetah

RLLab McGill University

Mila

Results: Mujoco environment



(a) Ant

Conclusion

- **Safe** approach of learning policy in **Actor-Critic** style methods.
- Constrained unsafe regions by **regularizing** the **variance in the return**.
- Scalable framework, comparable or better results than DPPO in Mujoco environments.

Future Work:

- Variable value of ψ ranging from $0 \rightarrow$ high where in beginning it promotes exploration and later curb visitation to unsafe or highly varied behavior.
- More results !

RLLab McGill University

Mila

Conclusion

- **Safe** approach of learning policy in **Actor-Critic** style methods.
- Constrained unsafe regions by **regularizing** the **variance in the return**.
- Scalable framework, comparable or better results than DPPO in Mujoco environments.

Future Work:

- Variable value of ψ ranging from $0 \rightarrow$ high where in beginning it promotes exploration and later curb visitation to unsafe or highly varied behavior.
- More results !

RLLab
McGill University

Mila

Bibliography

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

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

RLLab McGill University

Mila