Certificated Actor-Critic: Hierarchical Reinforcement Learning with Control Barrier Functions for Safe Navigation

Junjun Xie[1†], Shuhao Zhao[1†], Liang Hu[1*], Huijun Gao[2*]

1 Harbin Institute of Technology, Shenzhen

2 Harbin Institute of Technology, Harbin

arXiv youtube

Motivations

- Two limitations of RL with CBFs
  - Lack of quantitative results about policy performance

    Through the definition of the reward function, the value function can be used to evaluate policy performance. However, these results are often qualitative, only suitable for comparing the relative merits of different policies, and struggle to characterize the true performance of a policy (such as whether collisions occur).
  - Performance degradation due to multi-objective framework

    To simultaneously account for both safety and task performance, reward functions are often designed in a trade-off manner. However, improper weight settings may either compromise safety or result in overly conservative behavior, ultimately degrading task performance.

Contributions

- A hierarchical RL algorithm: Certificated Actor-Critic

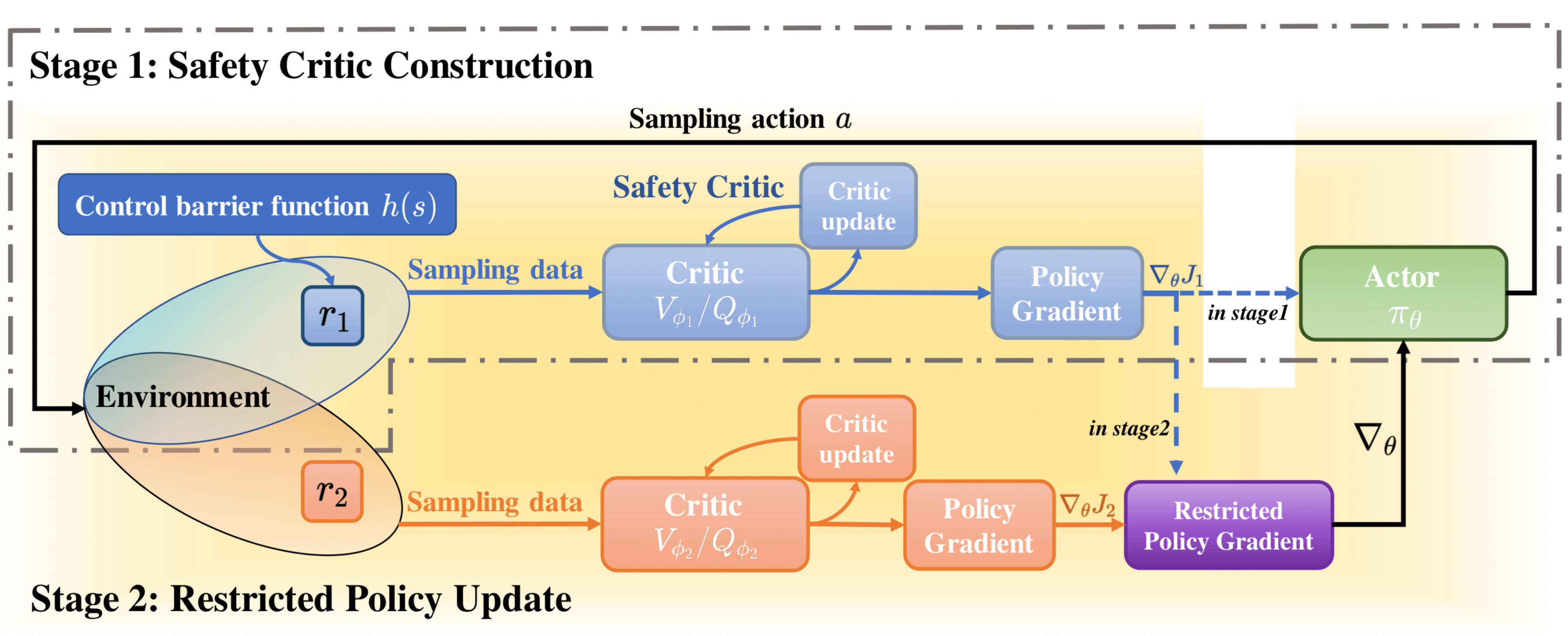
  We design a hierarchical framework that accommodates safety and goal-reaching objectives in robot navigation, and improve its goal-reaching capability yet maintaining safety via novel restricted policy update.
- Quantitative estimation about safety via CBF-based reward

  We propose a CBF-based reward function, which can quantitatively estimate the safety of the policies and states for learned policies.
- Experiments with detailed comparative analysis

  We conduct two experiments with detailed comparative analysis, showing the effectiveness of our proposed algorithm.

Methodology

Certificated Actor-Critic



Stage 1: Safety Critic Construction

- **Safety reward based on control barrier functions**

  $r_1 = \min(h(s_{t+1}) + (\alpha_0 - 1)h(s_t), 0)$
- **Value function as safety critics** — **Only consider safety**

  $V_1^\pi(s_0) = 0 \Rightarrow$ *The system is safe from initial state* s_0

  $Q_1^\pi(s_0, a_0) = 0 \Rightarrow$ *The system is safe from initial state-action pair* (s_0, a_0)

Stage 2: Restricted Policy Update

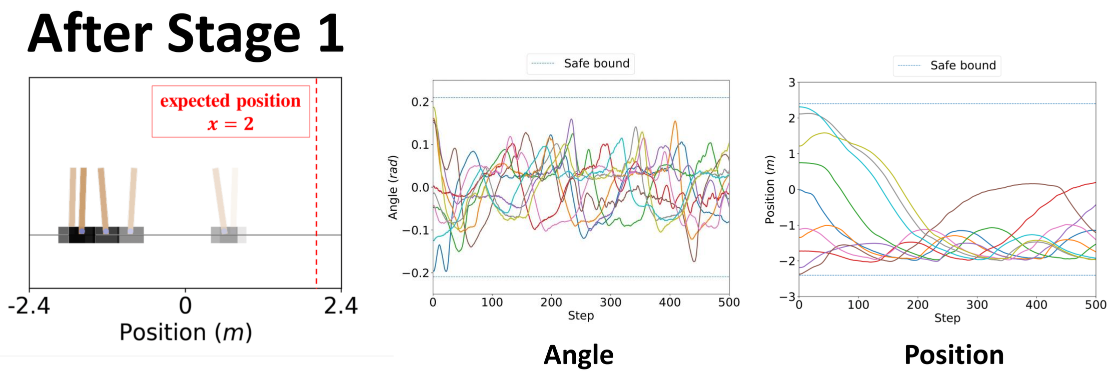
- **Update the policy for the other task while maintaining safety**

$$\nabla_\theta = \arg\max_e e \cdot \nabla_\theta J_2(\theta)$$
$$\text{s.t.} \quad e \cdot \nabla_\theta J_1(\theta) \geq 0$$
$$\|e\| \leq \|\nabla_\theta J_2(\theta)\|$$

Experiments

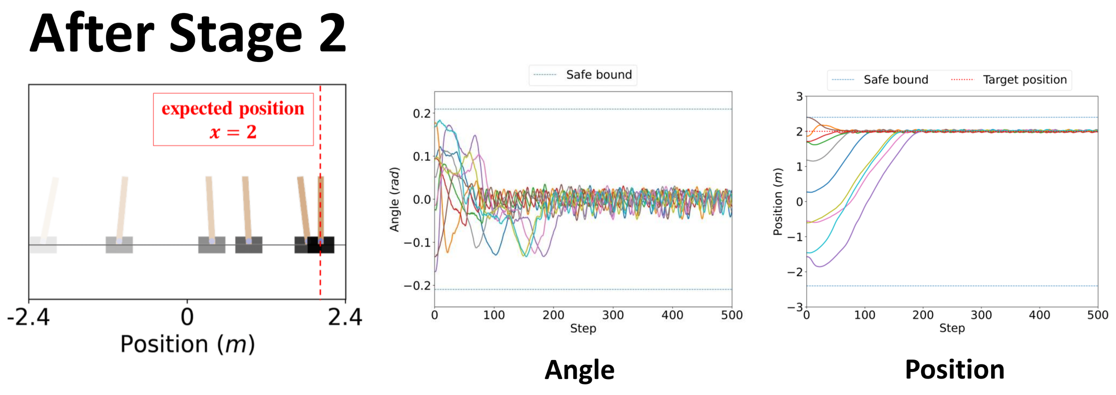
Cartpole

After Stage 1



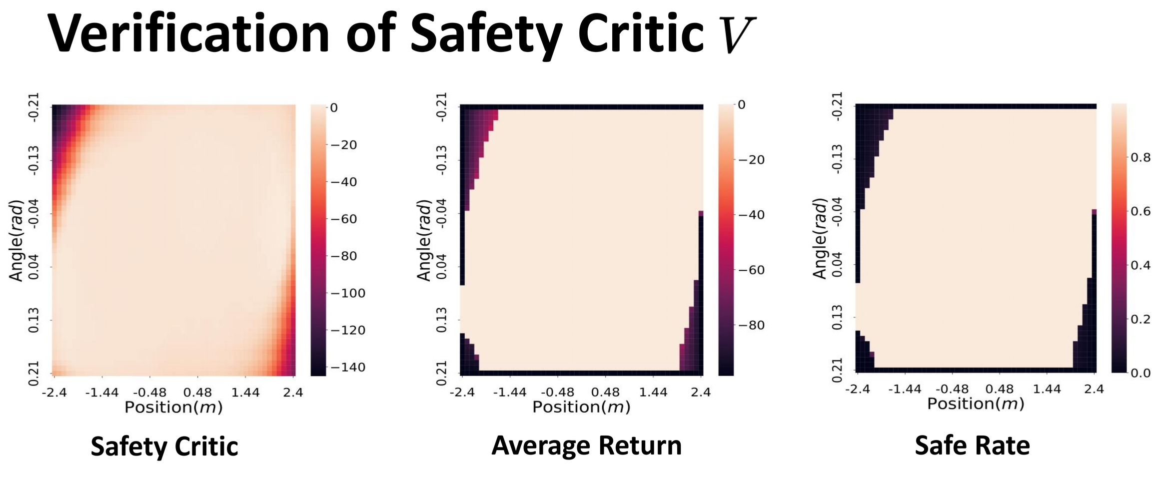
Safe but not convergent

After Stage 2

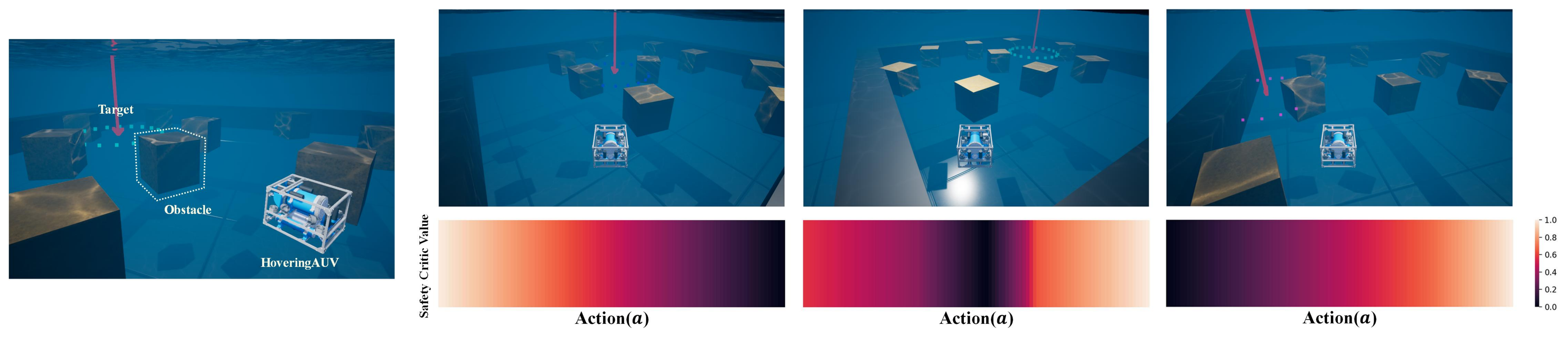


Safe and convergent

Verification of Safety Critic V



Autonomous Underwater Vehicle Navigation

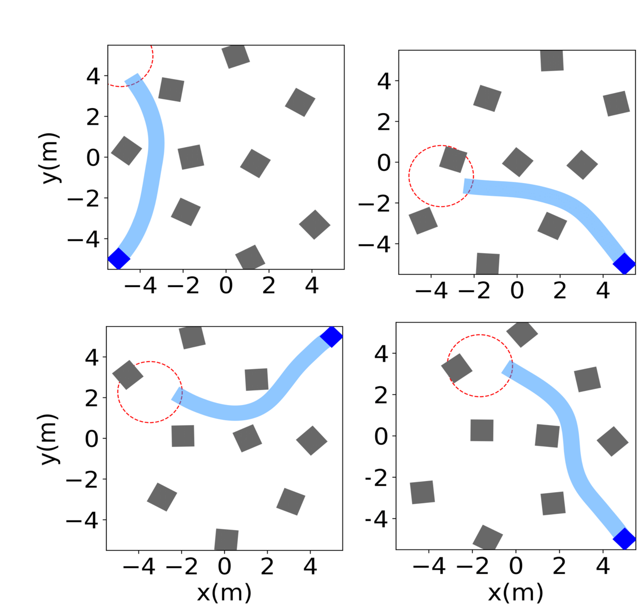


Verification of Safety Critic Q

ABLATIONS RESULTS IN 100 RANDOM EPISODES

Algorithm	Success Rate[a] ↑	Collision Rate ↓	Average length[b] ↓
T-O(0.5)[c]	48%	31%	852
T-O(0.25)	67%	33%	679
Stage 1	N/A	18%	N/A
W/o Re.[d]	49%	51%	**616**
Our CAC	**86%**	**11%**	625

a): reaching target position without collision is regarded as a successful episode;
b): the average length of successful episodes;
c): T-O(x) represents the reward function is set as a trade-off with x as the coefficient of safety reward, and $1 - x$ as that of navigation reward;
d): policy update only using $\nabla_\theta J_2(\theta)$ without restriction (10).



Trajectories