Project Update:
Safe Robotic Learning via Reachability Theory

Claire Tomlin
tomlin@berkeley.edu

Vicenç Rubies-Royo
vrubies@berkeley.edu

Roberto Calandra
rcalandra@fb.com

Our efforts of merging reachability analysis and reinforcement learning started with the work in ICRA'19 [1], where we blended reachability-based safety analysis and reinforcement learning. The next goal was to extend this safety framework to the more challenging setting of reach-avoid reinforcement learning and, additionally with FAIR, develop an algorithm able to incorporate high-dimensional observations such as LiDAR or camera feedback. Our recently published paper at RSS [2] demonstrated the applicability of reinforcement learning tools for reach-avoid problems. Fig. 1 shows our approach computing the reach-avoid set and control policy for a 2-dimensional environment with three purple obstacles (left, middle and right) and a yellow target set at the top. The green lines represent the analytic boundary of the reach-avoid set.

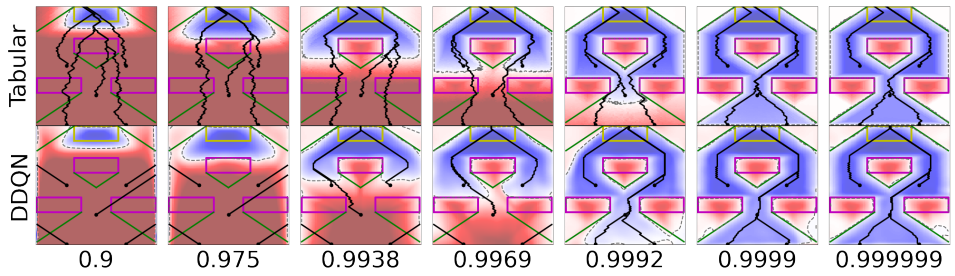

Figure 1: A convergent family of under-approximations that asymptotically approaches the undiscounted reach-avoid set as $\gamma \to 1$. The red region indicates positive state value, while blue region indicates negative state value. The dashed gray line specifies the zero level set or the discounted reach-avoid set boundary, while green lines specify analytic reach-avoid set boundary. The solid black lines show trajectory rollouts from five initial states.

This latest work also revealed some limitations; in particular, it suggests that certain reinforcement learning algorithms are better suited than others for reach-avoid problems. This limitation comes as a direct consequence of a phenomenon which we termed "the value function flattening problem".

1 The Value Function Flattening Problem

In reinforcement learning the cost functional is a sum of discounted rewards. In contrast, the (infinite time) reach-avoid cost functional looks as follows,

$$\mathcal{V}^{\mathbf{u}}(s) = \min_{\tau \in \{0,1,\ldots\}} \max \Big\{ l(\xi_s^{\mathbf{u}}(\tau)), \max_{\kappa \in \{0,\ldots,\tau\}} g(\xi_s^{\mathbf{u}}(\kappa)) \Big\}. \tag{1}$$

Both l and g are implicit surface functions which are bounded below. Their zero level set defines the target set and the constraint set respectively. Given these definitions, the outer minimum-over-time will produce a value which is also bounded below. This means that this minimum cost can be achieved by many initial states, which results in a value function that will be mostly "flat". This loss of gradient information poses an important challenge for computing the optimal policy. Through the discounted formulation in [2] this problem can be circumvented, albeit it requires proper tuning of the discount factor γ, which is also more challenging as the state dimension grows.

2 Lessons Learned and Current Work

The flattening problem becomes apparent when using actor-critic methods such as DDPG, TD3 or SAC. For these algorithms the critic provides a good representation of the reach-avoid set, but the associated policy does not reach the target unless the discount factor is properly chosen. Again, this is due to the fact that these algorithms first update the critic, and then they use the critic to update the actor. Since the critic loses gradient information in the interior of the reach-avoid set for $\gamma \sim 1$, the actor can't be properly updated.

To circumvent this problem it is necessary to change the order in which the different components are learned. Policy gradient algorithms are therefore more suitable for reach-avoid problems, since the policy is being learned directly, and it is this policy which is then used to approximate the value function.

In our current work we have implemented a "vanilla" policy gradient algorithm which is showing promising results in resolving the flattening problem. Beyond this, our goals are to implement more sophisticated policy gradient algorithms (TRPO, PPO etc.) for reach-avoid problems and then incorporate high-dimensional observations.

References

[1] J. F. Fisac, N. F. Lugovoy, V. Rubies-Royo, S. Ghosh, and C. J. Tomlin. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8550–8556, 2019.

[2] Kai-Chieh Hsu*, Vicenç Rubies-Royo*, Claire Tomlin, and Jaime Fisac.

Safety and liveness guarantees through reach-avoid reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2021.