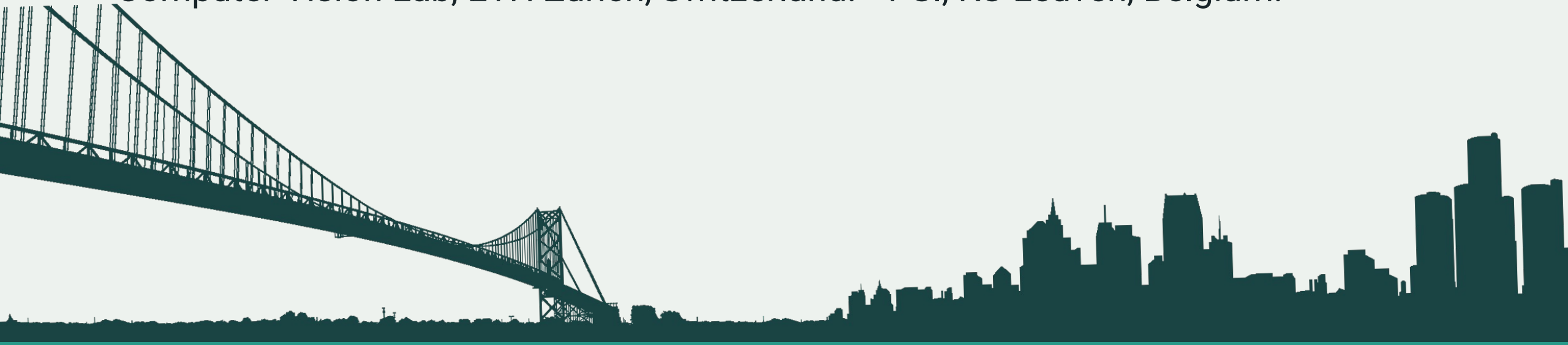
A Multiplicative Value Function for Safe and Efficient Reinforcement Learning

Nick Bührer [1], Zhejun Zhang [1], Alexander Liniger [1], Fisher Yu [1], Luc Van Gool [1,2].

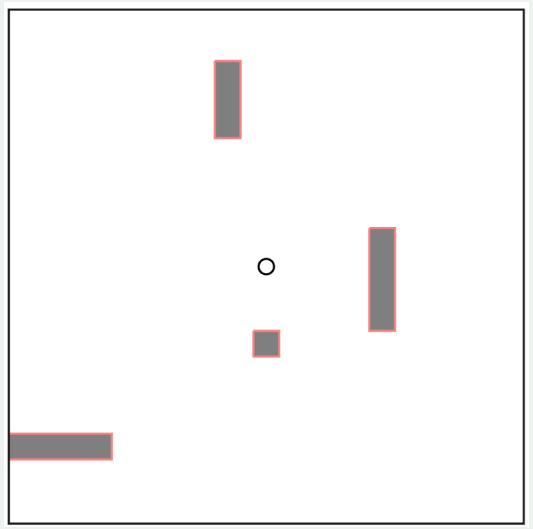
[1] Computer Vision Lab, ETH Zurich, Switzerland. [2] PSI, KU Leuven, Belgium.



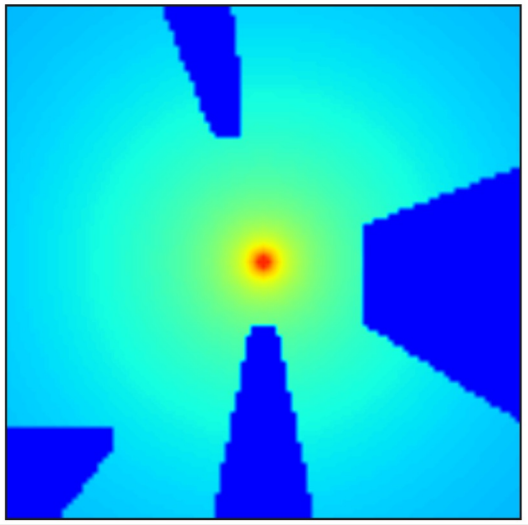
ETH zürich CVL Computer Vision Lab

https://zhejz.github.io/saferl

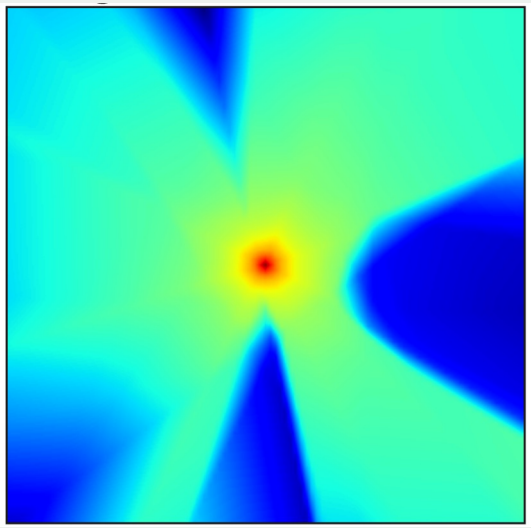
Motivation



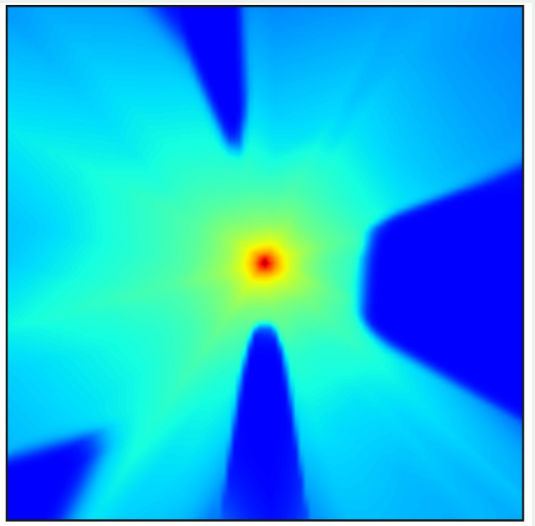
(A) Environment



(B) Ground-Truth Value



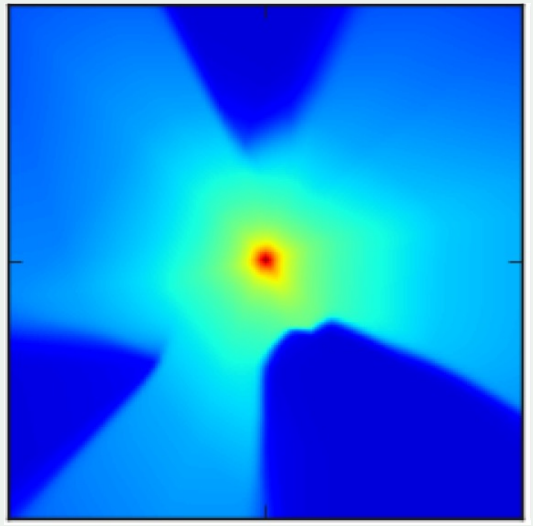
(C) Regular Value Function



(D) Multiplicative Value Function

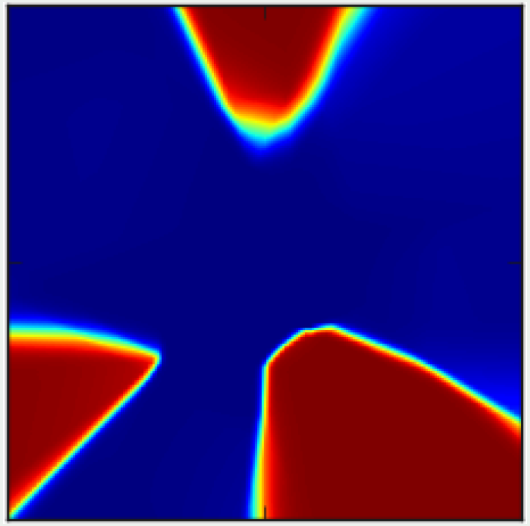
Multiplicative Value Function

Multiplicative Value Function V^{π}_{mult}



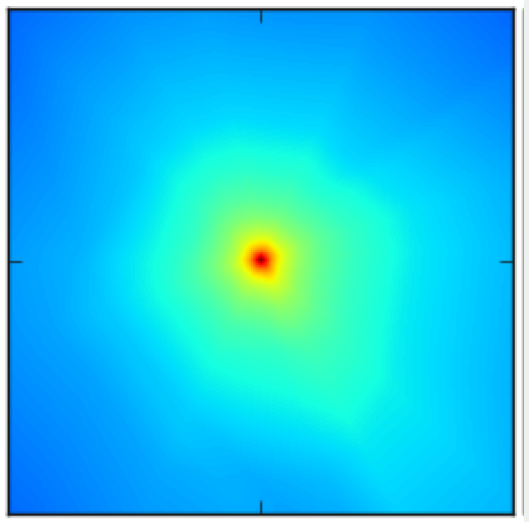
=

Probabilistic Safety Critic Ψ^{π}



⊗

Constraint-Free Reward Critic $\bar{V}^{\pi}$

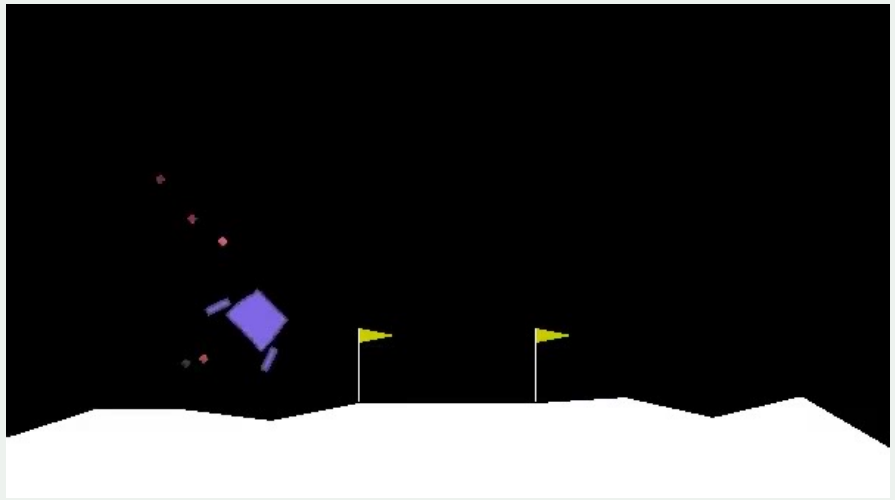


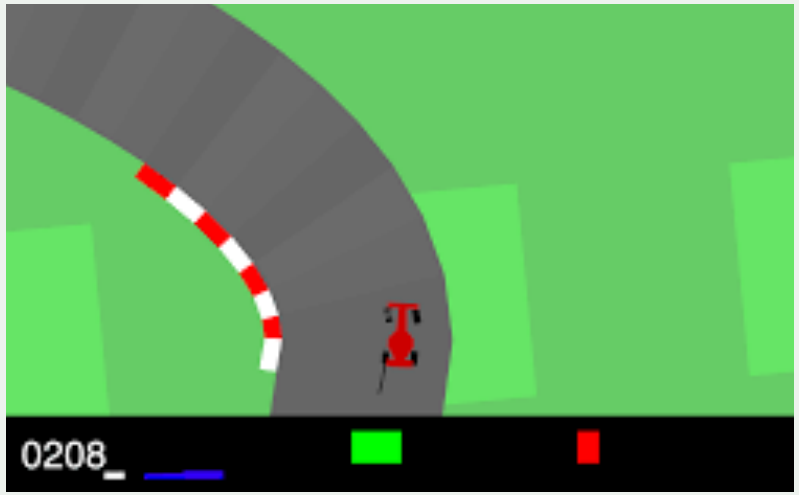
$$V^{\pi}_{\text{mult}}(s) = (\bar{V}^{\pi}(s) - \bar{v}_{min}) \cdot (1 - \Psi^{\pi}(s)) + \bar{v}_{min},\ \bar{v}_{min} \coloneqq \min_{s} \bar{V}^{\pi}(s)$$

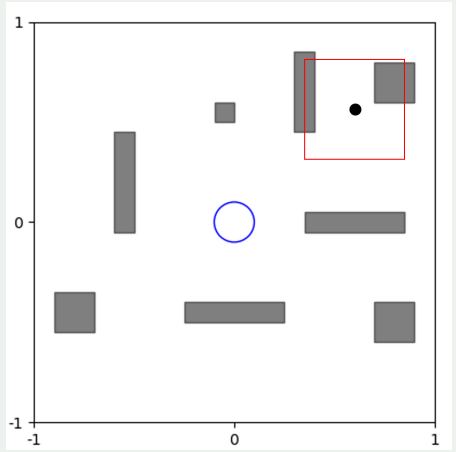
Apply to SAC and PPO

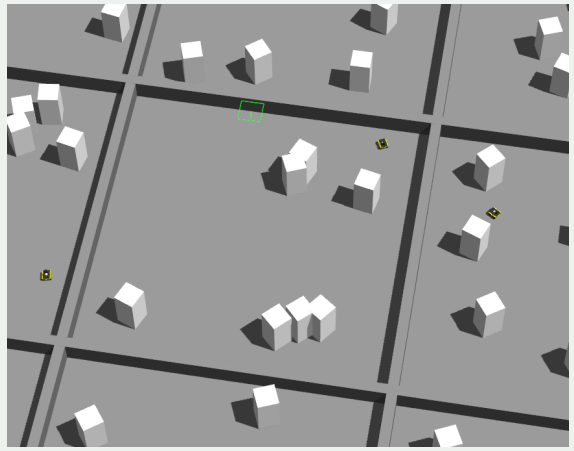
- **Q Function**: $Q^{\pi}_{\text{mult}}(s,a) = (\bar{Q}^{\pi}(s,a) - \bar{q}_{min}) \cdot (1 - \Psi^{\pi}(s,a)) + \bar{q}_{min}$
- **Advantage**
  - V1 bootstrap Q: $A^{\pi}_{\text{mult}}(s_t,a_t) = [\overline{r_t} + \gamma V^{\pi}_{\text{mult}}(s_{t+1})] - V^{\pi}_{\text{mult}}(s_t)$
  - V2 without bootstrap: $A^{\pi}_{\text{mult}}(s_t,a_t) = Q^{\pi}_{\text{mult}}(s_t,a_t) - V^{\pi}_{\text{mult}}(s_t)$
  - V3 bootstrap the safety critic inside $Q^{\pi}_{\text{mult}}(s_t,a_t)$ of V2
- **SAC:** $\max_{\theta} \mathbb{E}_{a_\theta \sim \pi_\theta}[Q^{\pi_\theta}_{\text{mult}}(s,a_\theta) - \alpha \log \pi_\theta(s_\theta|x)]$
- **PPO**: $\max_{\theta} \mathbb{E}_{a\sim\pi_\theta}\left[\min\left\{\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{\pi_{\theta_k}}_{\text{mult}}(s,a), g\left(\epsilon, A^{\pi_{\theta_k}}_{\text{mult}}(s,a)\right)\right\}\right]$

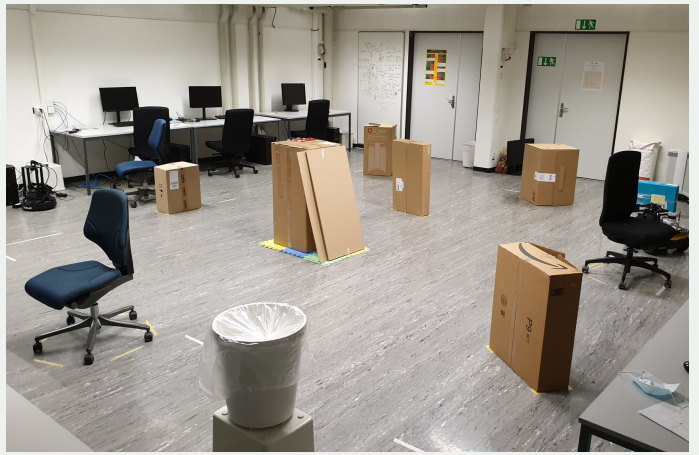
Safety-Focused Environments



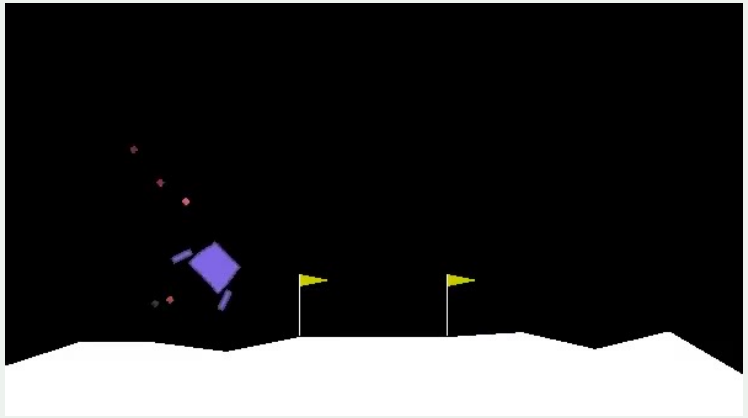




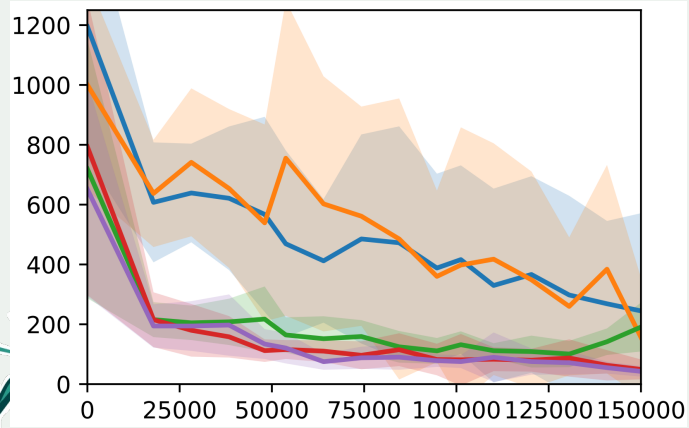




Quantitative Results



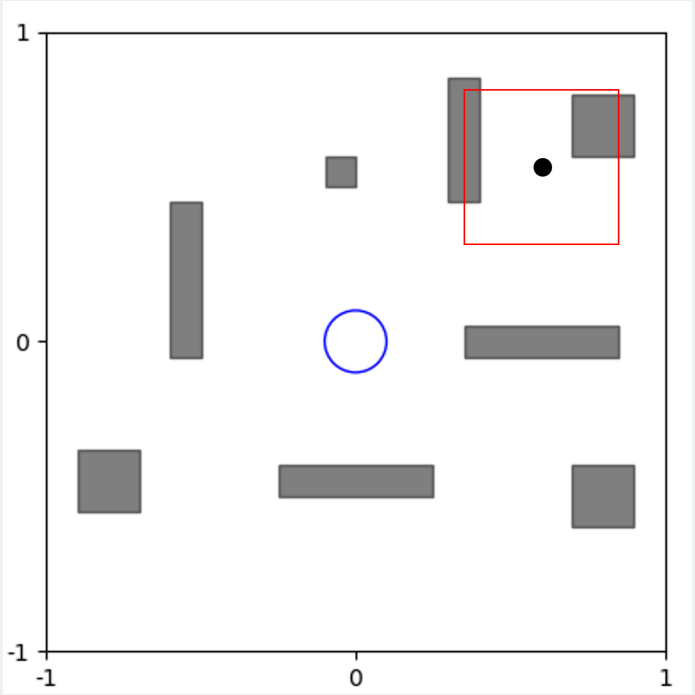
Value Loss of Lunar Lander Safe



Blue: Orange: Green: Violet: Red:

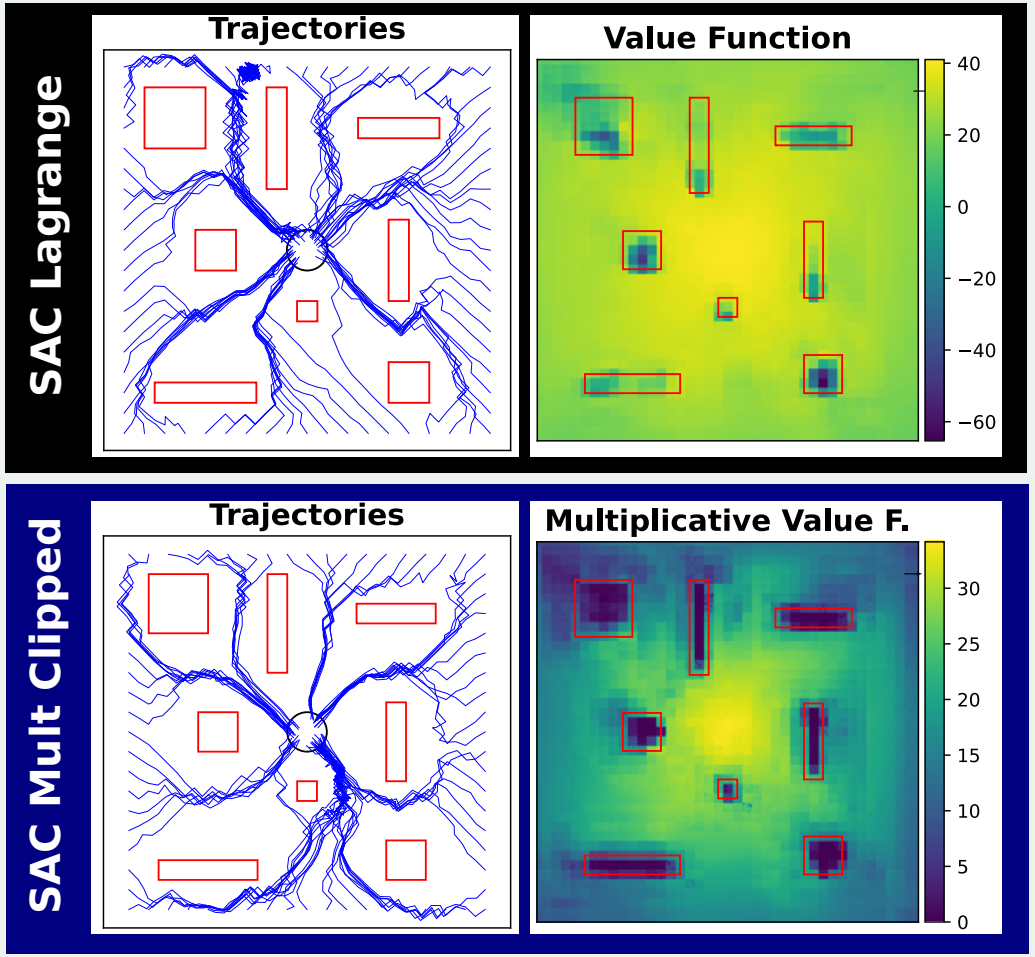
	Reward ↑	% Constraint violations ↓	Reward ↑	% Constraint violations ↓
	Lunar Lander Safe			
SAC	50k		150k	
SAC base	90 ± 108	10 ± 16	181 ± 117	3 ± 6
Lagrange	111 ± 105	17 ± 13	184 ± 128	2 ± 3
Mult	-35 ± 27	$\mathbf{3 \pm 5}$	-34 ± 22	3 ± 4
Mult Clipped	$\mathbf{134 \pm 94}$	14 ± 13	243 ± 49	8 ± 15
Mult Lagrange	125 ± 59	29 ± 15	$\mathbf{251 \pm 20}$	$\mathbf{2 \pm 2}$
PPO	50k		150k	
PPO base	-126 ± 158	77 ± 29	225 ± 100	10 ± 30
Lagrange	-24 ± 146	54 ± 39	204 ± 116	12 ± 24
V1	101 ± 84	41 ± 19	205 ± 78	7 ± 16
V2	89 ± 122	44 ± 34	251 ± 28	5 ± 9
V3	$\mathbf{144 \pm 4}$	$\mathbf{26 \pm 22}$	$\mathbf{264 \pm 5}$	$\mathbf{1 \pm 2}$
FOCOPS	-129 ± 21	64 ± 24	117 ± 80	30 ± 19

Qualitative Results



Point Robot Navigation

Via local occupancy grid and a vector pointing to the goal.



Simulation and Real-world

- Differential drive robot with 1D-Lidar
- Gazebo simulation
- Zero-shot sim-to-real

Summary

- Multiplicative value function
  - Constraint-free reward critic ⊗ Probabilistic safety critic
- Integration into SAC and PPO
  - Increased sample efficiency and learning stability.
- Experiments
  - Safety-focused RL environments and real-world robot navigation.
- Future works
  - Theoretical justification.