Genetic Soft Updates for Policy Evolution in Deep Reinforcement Learning

Enrico Marchesini, Davide Corsi, Alessandro Farinelli
University of Verona, Italy

enrico.marchesini@univr.it

05/2021

ICLR

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Deep Reinforcement Learning Issues

1 A huge number of trials is required to achieve good performance.

↓

Devising robust learning approaches improving sample efficiency

2 Convergence to local optima, mainly caused by the lack of diverse exploration in high-dimensional spaces.

↓

Recent approaches for the exploration[1, 2] problem relies on task-specific hyperparameters

[1] Pathak et al., "Curiosity-Driven Exploration by Self-Supervised Prediction", CVPRW 2017.
[2] Ostrovski et al., "Count-Based Exploration with Neural Density Models", ICML 2017.

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Evolutionary Algorithms[1]

Gradient-free **population-based** approaches are a natural way to complement DRL:

- The population search enable diverse exploration
- More diversified samples improve training robustness
- Low computational cost

They struggle to solve high-dimensional problems and have poor sample efficiency

[1] Pathak et al., "Curiosity-Driven Exploration by Self-Supervised Prediction", CVPRW 2017.
[2] Ostrovski et al., "Count-Based Exploration with Neural Density Models", ICML 2017.

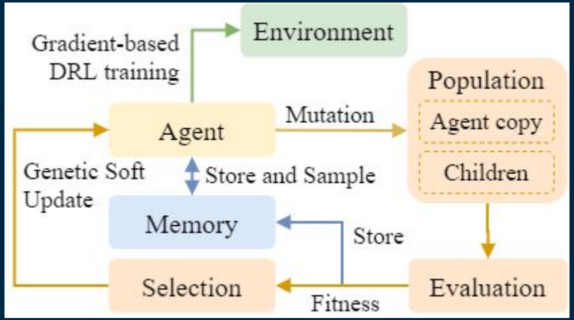
Supe-RL

Inspired by nature, we **combine** Evolutionary Algorithms with DRL algorithms to address the limitation of both approaches.

Polyak averaging is used to shift the agent policy towards an improved version of itself.

↓

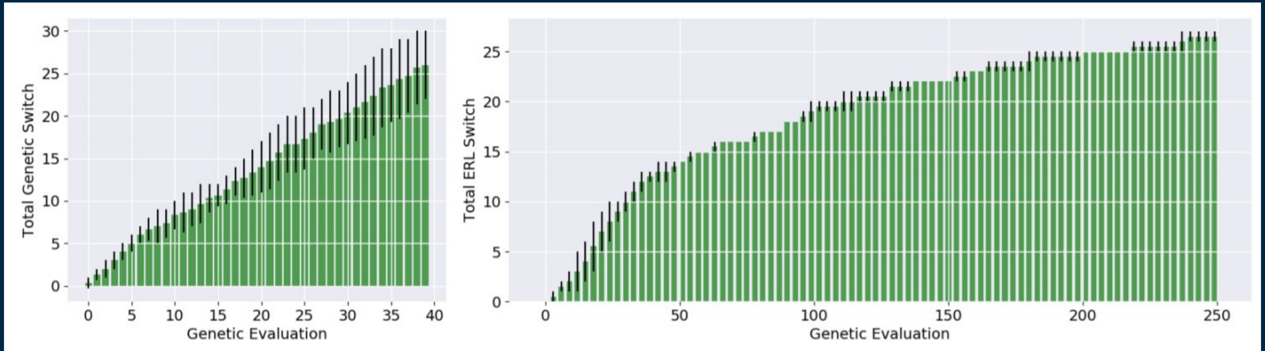
This simulates a **gradient step** towards a better policy.

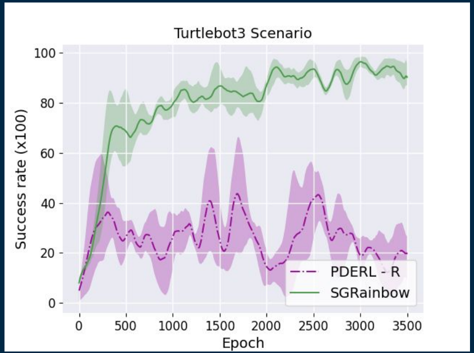


Previous Approaches

Supe-RL addresses the limitations of previous approaches[1, 2, 3]:

- The concurrent training of EA and DRL causes significant **overhead**.
- The **actor-critic** formalization hinders the combination with value-based DRL.
- Previous combination strategy do not ensure better performance and can result in **detrimental behaviours**.



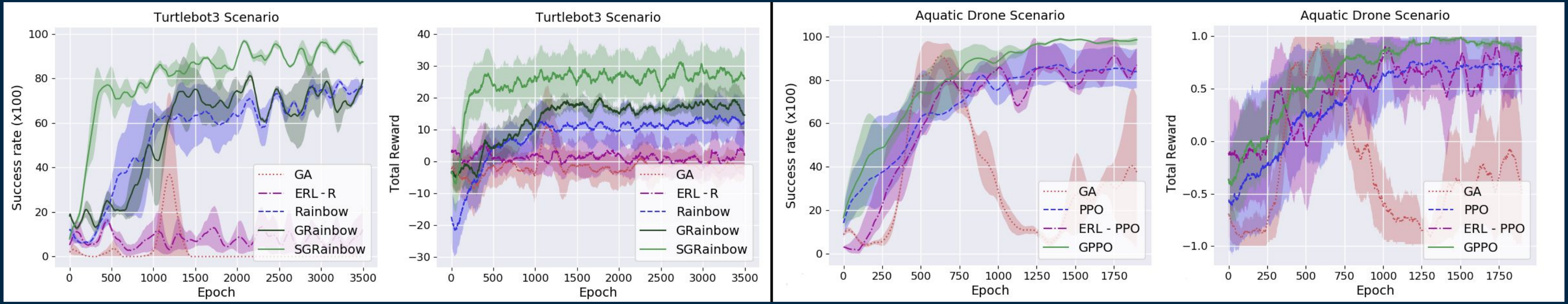


[1] Khadka and Tumer, "Evolutionary Reinforcement Learning", NIPS 2018.
[2] Pourchot and Sigaud, "CEM-RL: Combining evolutionary and gradient-based methods for policy search", ICLR 2019.
[2] Bodnar et al., "Proximal Distilled Evolutionary Reinforcement Learning", AAAI 2020.

Results in Real Robotic Applications

We evaluated both a value-based and a policy-gradient implementation of Supe-RL in two real robotic navigation environment; a well-known task in recent literature[1, 2]

- An indoor Turtlebot3 mapless navigation task with a discrete action-space.
- A novel outdoor highly dynamic aquatic navigation task with a continuous action-space.



Standard Evaluation Issues

1 — Standard metrics related to performance (e.g., return) are not informative in real applications.

2 — Common evaluation strategies rarely consider low visited situations, causing safety problems.

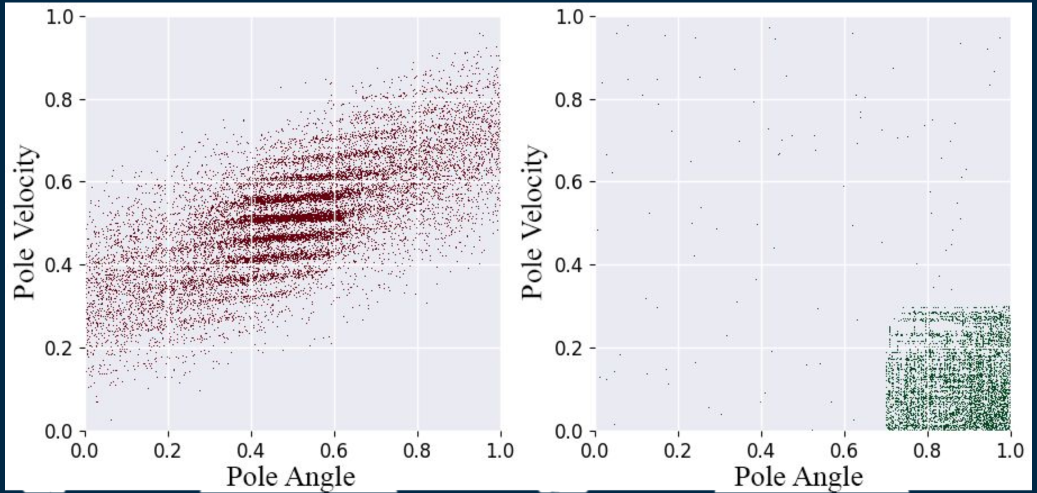
We rely on formal verification[1] to measure the violations of desired safety properties.

We also use the verification to confirm that Supe-RL bias the exploration in direction of more robust policy regions.

[1] Marchesini et al., "Formal Verification for Safe Deep Reinforcement Learning in Trajectory Generation", IRC 2020.

Why Formal Verification Is Crucial

Standard evaluations and metrics are **not informative** in low-visited states and can not formally guarantee the behaviors of the model.



UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Formal Verification Results in Navigation

Verification performance of simple safety properties (e.g., do not turn in the direction of a close obstacle) in the indoor (top) and outdoor (down) robotic navigation.

	Violation (%)			Time (s)			Memory (MB)		
Model	$\Theta_{I,0}$	$\Theta_{I,1}$	$\Theta_{I,2}$	$\Theta_{I,0}$	$\Theta_{I,1}$	$\Theta_{I,2}$	$\Theta_{I,0}$	$\Theta_{I,1}$	$\Theta_{I,2}$
Rainbow	2.21	9.11	0	79.7	75.5	92.6	3.74	3.96	6.92
SGRainbow	0	4.75	0	66.7	74.1	80.5	2.18	2.91	4.1

	Violation (%)		Time (s)		Memory (MB)	
Model	$\Theta_{A,0}$	$\Theta_{A,1}$	$\Theta_{A,0}$	$\Theta_{A,1}$	$\Theta_{A,0}$	$\Theta_{A,1}$
PPO	0.9	1.2	3.4	124	0.1	5.8
ERL	0.5	0.7	3.4	3.4	0.1	0.15
GPPO	0	0.1	3.1	3.2	0.1	0.1

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Questions?

enrico.marchesini@univr.it

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA