Evolution + deep RL

Using evolution to optimize policies

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Introduction

- RL methods follow a gradient, but do not evaluate the obtained agent
- Evolutionary methods perform random variations, but evaluate the agents
- Evo+RL: looking for the best of both worlds
- Four perspectives:
  - Evolving policies to improve performance
  - Evolving actions to improve performance
  - Evolving policies to improve diversity
  - Evolving something else (hyper-params, environment, reward function...)
- Older perspective: learning classifier systems

Sigaud, O. (2022) Combining evolution and deep reinforcement learning for policy search: a survey. *arXiv preprint arXiv:2203.14009*

Sigaud, O. and Wilson, S. W. (2007) Learning classifier systems: A survey. *Journal of Soft Computing*, 11(11):1065–1078

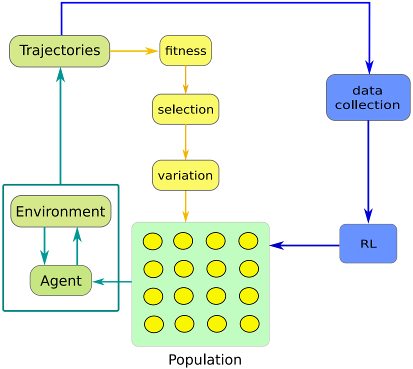
INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

Evolving policies to improve performance: overview

Algo. \ Prop.	RL algo.	Evo. algo.	Actor Injec.	+ Comb. Mech.	Surr. Fitness	Soft Update	Buffer Filt.
ERL Khadka and Turner (2018)	DDPG	GA	•	x	x	x	x
CERL Khadka et al. (2019)	TD3	GA	•	x	x	x	x
PDERL Bodnar et al. (2020)	TD3	GA	•	x	x	x	x
ESAC Suri et al. (2020)	SAC	ES	•	x	x	x	x
FIDI-RL Shi et al. (2019)	DDPG	ARS	•	x	x	x	x
X-DDPG Espositi and Bonarini (2020)	DDPG	GA	•	x	x	x	x
CEM-RL Pourchot and Sigaud (2019)	TD3	CEM	x	•	x	x	x
CEM-ACER Tang (2021)	ACER	CEM	x	•	x	x	x
SERL Wang et al. (2022)	DDPG	GA	•	x	•	x	x
SPDERL Wang et al. (2022)	TD3	GA	•	x	•	x	x
PGPS Kim et al. (2020)	TD3	CEM	•	x	•	•	x
BNET Stork et al. (2021)	BBNE	CPG	•	x	•	x	x
CSPC Zheng et al. (2020)	SAC + PPO	CEM	•	x	x	x	•
SUPE-RL Marchesini et al. (2021)	RAINBOW or PPO	GA	•	♠	x	•	x
G2AC Chang et al. (2018)	A2C	GA	x	♠	x	x	x
G2PPO Chang et al. (2018)	PPO	GA	x	♠	x	x	x

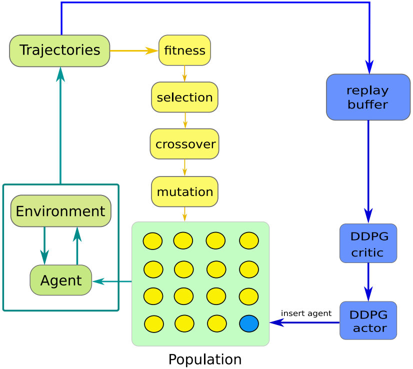
- The most active line of combinations

General architecture



- A template for many algorithms

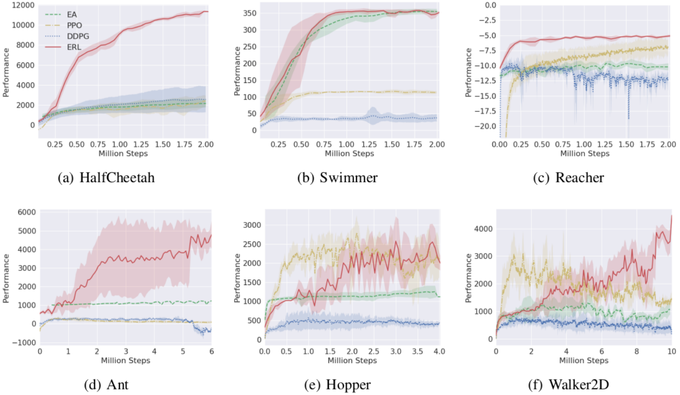
The renewal of Evo+RL: ERL



- One of the main Evo+RL approaches

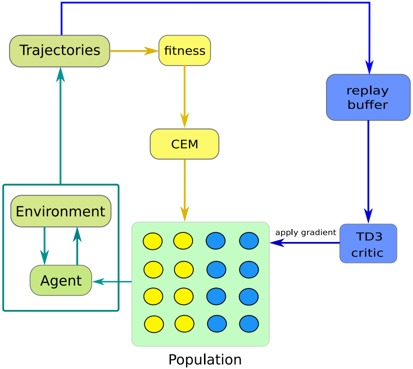
Khadka, S. & Tumer, K. (2018) Evolution-guided policy gradient in reinforcement learning. In *Neural Information Processing Systems*

ERL results



- Consistent improvement in performance and sample efficiency

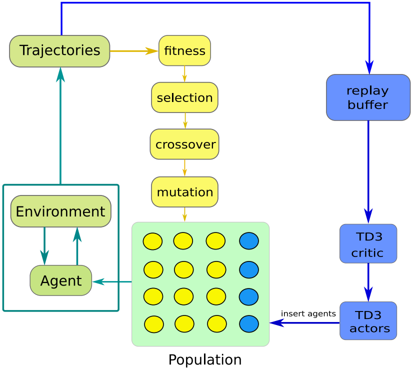
Competitor: CEM-RL



- Applies the TD3 critic gradient to half the population at all steps

Pourchot, A. & Sigaud, O. (2018) CEM-RL: Combining evolutionary and gradient-based methods for policy search. *arXiv preprint arXiv:1810.01222* (ICLR 2019)

ERL follow-up: CERL



- CERL: ERL with a population of TD3 actors
- Similar to ERL, but:
  - Uses TD3 instead of DDPG
  - Uses a population of actors sharing the same critic
  - Does not evolve their hyper-parameters, as in PBT

Khadka, S., Majumdar, S., Miret, S., Tumer, E., Nassar, T., Dwiel, Z., Liu, Y., & Tumer, K. (2019) Collaborative evolutionary reinforcement learning. *arXiv preprint arXiv:1905.00976* (ICML 2019)

Other follow-up

- FIDI-RL: In ERL, replaces the GA with Augmented Random Search
- ESAC: In ERL, replaces DDPG with SAC
- CEM-ACER: In CEMRL, replaces TD3 with ACER

Mania, H., Guy, A., and Recht, B. Simple random search of static linear policies is competitive for reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018

Shi, L., Li, S., Cao, L., Yang, L., Zheng, G., and Pan, G. Fidi-RL: Incorporating deep reinforcement learning with finite-difference policy search for efficient learning of continuous control. *arXiv preprint arXiv:1907.00526*, 2019

Suri, K., Shi, X. Q., Plataniotis, K. N., and Lawryshyn, Y. A. Maximum mutation reinforcement learning for scalable control. *arXiv preprint arXiv:2007.13690*, 2020

Tang, Y. (2021) Guiding evolutionary strategies with off-policy actor-critic. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1317–1325

ERL Limitations

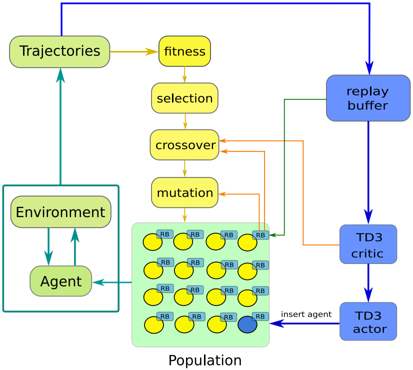
- With direct encoding of NN parameters, the standard n-point crossover and mutations can be disruptive.
- PDERL replaces them with:
  - A combined policy distillation operator for crossover, inspired from Gangwani & Peng (2017)
  - The safe mutation operator of Lehman et al. (2018)

Bodnar, C., Day, B., and Lió, P. Proximal distilled evolutionary reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3283–3290, 2020

Gangwani, T. and Peng, J. Policy optimization by genetic distillation. *arXiv preprint arXiv:1711.01012*, 2017

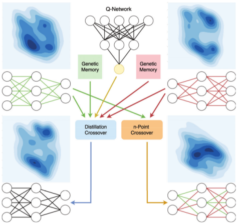
Lehman, J., Chen, J., Clune, J., and Stanley, K. O. Safe mutations for deep and recurrent neural networks through output gradients. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 117–124, 2018

PDERL Architecture



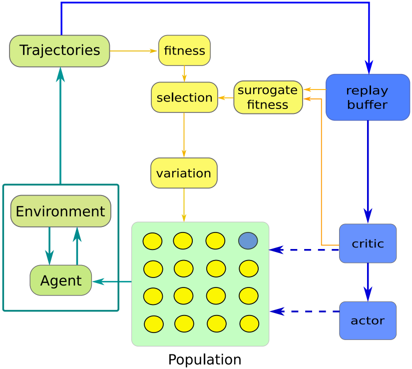
- Genetic memory: small local replay buffer
- Code available at https://github.com/crisbodnar/pderl.

PDERL: New crossover: based on experience



- Each worker has its own replay buffer
- Crossover is between the replay buffers of parents
- The choice of parents is based on:
  - Greedy: the sum of their fitnesses
  - Diversity: the sum of their parameter distances
- Which choice is used is not specified nor studied!
- If policies are stochastic, one could use their KL divergence
- In an ERL-like approach, the most distant policy is often the TD3 agent
- The choice of samples is based on their Q-value, according to the central critic
- Offspring trained with behavioral cloning based on its synthetic replay buffer

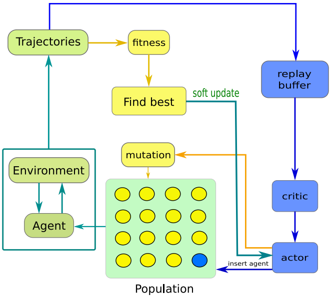
SC-ERL (Surrogate-Assisted Controller)



- The SC component is generic, it can be plugged into most Evo+RL algos
- The critic gives the surrogate fitness over a set of states taken from the replay buffer

Wang, Y., Zhang, T., Chang, Y., Wang, X., Liang, B., and Yuan, B. (2022) A surrogate-assisted controller for expensive evolutionary reinforcement learning. *Information Sciences*, 616:539–557

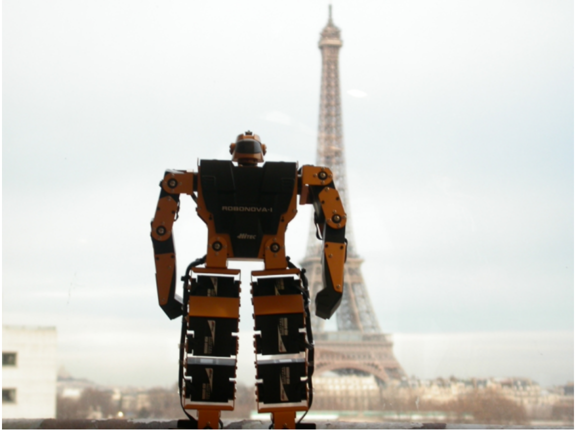
Supe-RL



- In ERL, CERL, CEMRL, the main loop is evolutionary, the RL loop generates faster improvement
- In SUPE-RL, the main loop is the RL loop, from time to time it performs an evolutionary step and improve the RL agent if a better offspring is found
- Performs a soft update of the RL agent towards the best offspring

Marchesini, E., Corsi, D., and Farinelli, A. Genetic soft updates for policy evolution in deep reinforcement learning. In *International Conference on Learning Representations*, 2021

Any question?



Send mail to: `Olivier.Sigaud@isir.upmc.fr`

Bodnar, C., Day, B., and Lió, P. (2020).
Proximal distilled evolutionary reinforcement learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3283–3290.

Gangwani, T. and Peng, J. (2017).
Policy optimization by genetic distillation.
arXiv preprint arXiv:1711.01012.

Khadka, S., Majumdar, S., Miret, S., Tumer, E., Nassar, T., Dwiel, Z., Liu, Y., and Tumer, K. (2019).
Collaborative evolutionary reinforcement learning.
arXiv preprint arXiv:1905.00976.

Khadka, S. and Tumer, K. (2018).
Evolution-guided policy gradient in reinforcement learning.
In *Neural Information Processing Systems*.

Lehman, J., Chen, J., Clune, J., and Stanley, K. O. (2018).
Safe mutations for deep and recurrent neural networks through output gradients.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 117–124.

Mania, H., Guy, A., and Recht, B. (2018).
Simple random search of static linear policies is competitive for reinforcement learning.
Advances in Neural Information Processing Systems, 31.

Marchesini, E., Corsi, D., and Farinelli, A. (2021).
Genetic soft updates for policy evolution in deep reinforcement learning.
In *International Conference on Learning Representations*.

Pourchot, A. and Sigaud, O. (2018).
CEM-RL: Combining evolutionary and gradient-based methods for policy search.
arXiv preprint arXiv:1810.01222 (ICLR 2019).

Shi, L., Li, S., Cao, L., Yang, L., Zheng, G., and Pan, G. (2019).

Fidi-rl: Incorporating deep reinforcement learning with finite-difference policy search for efficient learning of continuous control.
arXiv preprint arXiv:1907.00526.

Sigaud, O. (2022).
Combining evolution and deep reinforcement learning for policy search: a survey.
ACM Transactions in Evolutionary Learning and Optimization.

Sigaud, O. and Wilson, S. W. (2007).
Learning classifier systems: A survey.
Journal of Soft Computing, 11(11):1065–1078.

Suri, K., Shi, X. Q., Plataniotis, K. N., and Lawryshyn, Y. A. (2020).
Maximum mutation reinforcement learning for scalable control.
arXiv preprint arXiv:2007.13690.

Tang, Y. (2021).
Guiding evolutionary strategies with off-policy actor-critic.
In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1317–1325.

Wang, Y., Zhang, T., Chang, Y., Wang, X., Liang, B., and Yuan, B. (2022).
A surrogate-assisted controller for expensive evolutionary reinforcement learning.
Information Sciences, 616:539–557.