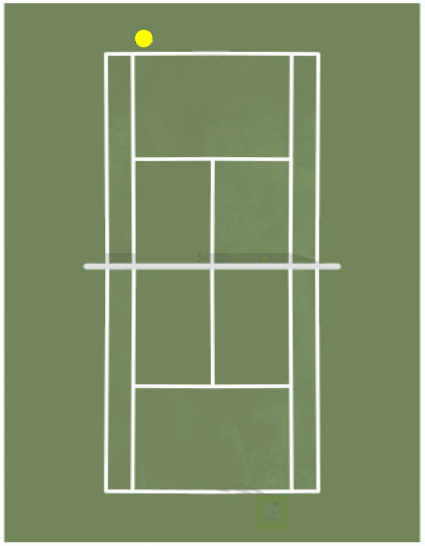
Evolution + deep RL

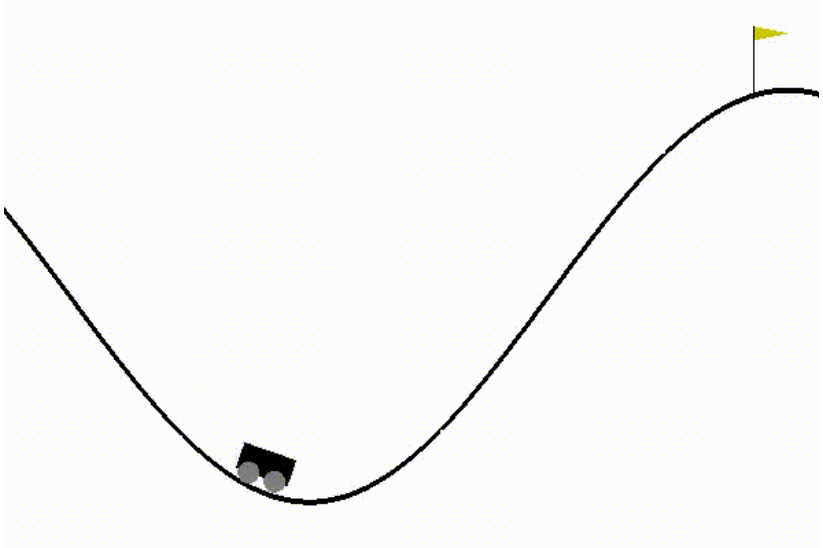
Optimizing diversity

Olivier Sigaud

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

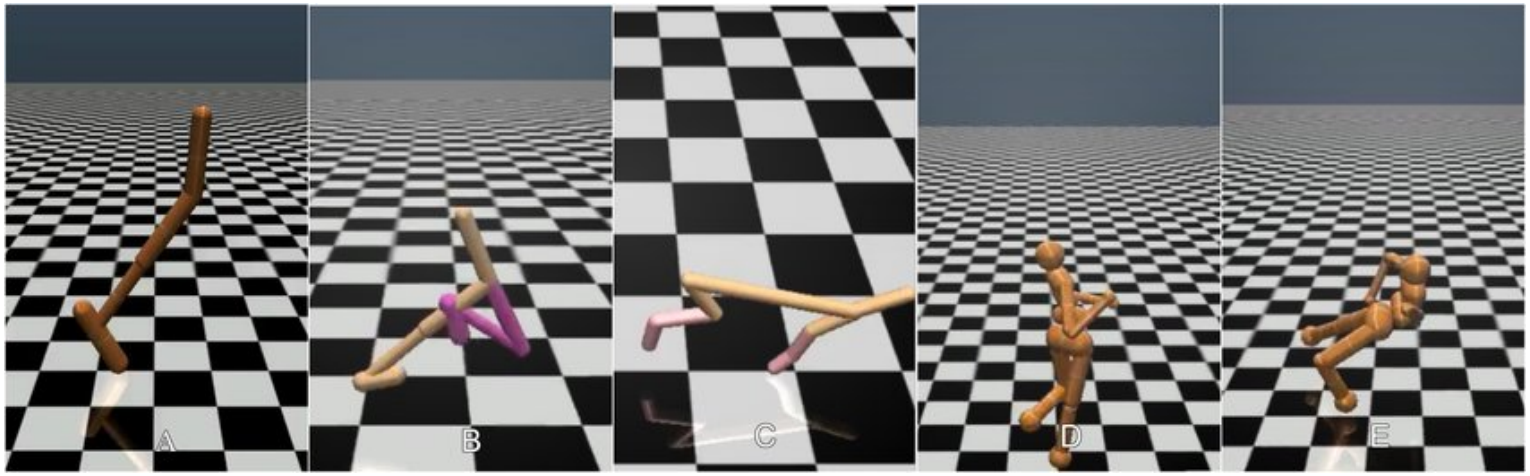
Reinforcement learning issues: sparse rewards and deceptive gradients





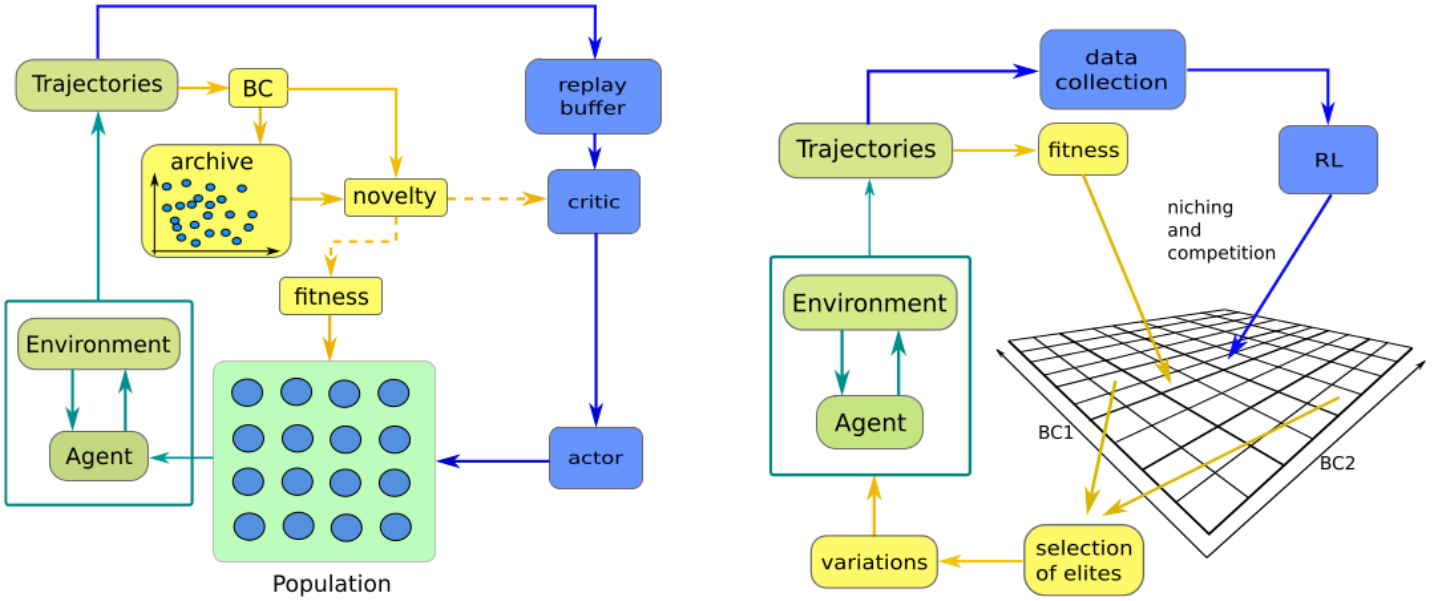
- Sparse reward: very few trajectories are rewarded, the agent learns nothing
- Deceptive gradient: drives the agent away from the target trajectory
- Looking for diversity helps finding sparse rewards
- Looking for diversity only handles deceptive gradients
- But diversity in irrelevant behaviors is not a good solution

Diversity that matters: behavior descriptors



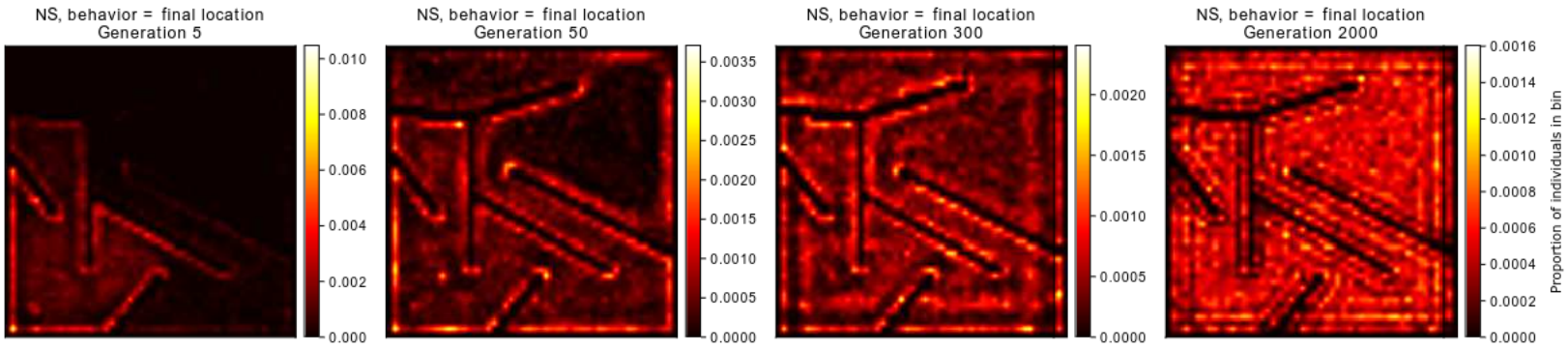
- We would like to get policies that behave differently with respect to domain-relevant criteria
- The easiest approach is to define a set of behavior descriptors and to cover the space of these descriptors
- E.g. in locomotion: running speed, frequency of ground contacts, head height...
- E.g. in maze navigation: final point, distance travelled...
- These behavior descriptors could be learned (not covered)

Evolving policies for diversity: two frameworks



- Behavior characterizations (BC) = B. descriptors (BD), describe trajectories
- The NS approach only looks for diversity. It is better in the absence of reward, or when the reward signal is very sparse or deceptive
- The QD approach is more appropriate when the reward signal is more dense and can contribute to the policy search process

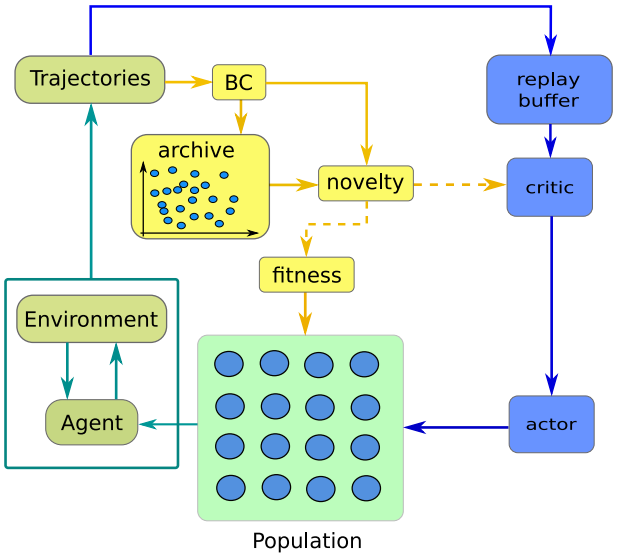
Novelty Search basics



- Evolution of just novelty with respect to an archive of policies
- The fitness is a function of a distance to other policies in the BD space
- In practice, mean distance to K nearest neighbors (more continuous than distance to closest or K-closest)
- Several archive management methods (not covered)

Lehman, J. and Stanley, K. O. (2011) Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223

NS with RL

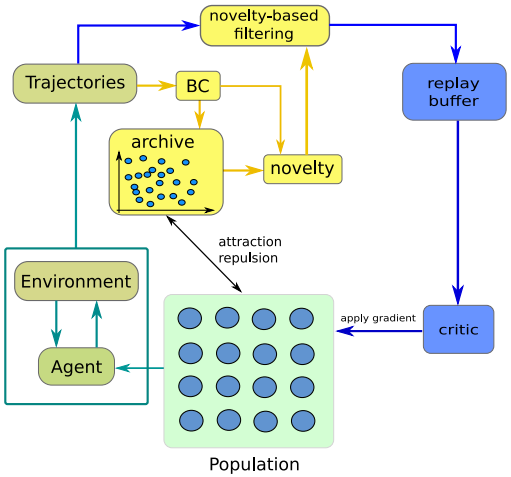


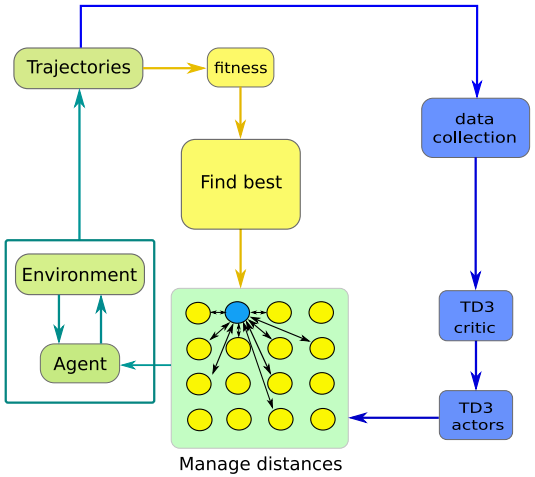
Prop. / Algo.	RL algo.	Diversity algo.	Distance space
P3S-TD3 [32]	TD3	Find best	Policy params.
DEPRL [48]	TD3	CEM	Policy params.
ARAC [14]	SAC	NS-like	Policy params.
NS-RL [78]	GC-DDPG	True NS	Manual BC
PNS-RL [49]	TD3	True NS	Manual BC

- The figure suggests 3 cases: novelty used as fitness, as reward, or both
- Used in both: NS-RL, with goal-conditioned policies
- If novelty is used just as fitness, the combination is close to a QD method
- Future research: investigate the differences to QD

Shi, L., Li, S., Zheng, Q., Yao, M., and Pan, G. Efficient novelty search through deep reinforcement learning. *IEEE Access*, 8:128809–128818, 2020.

ARAC and P3S-TD3



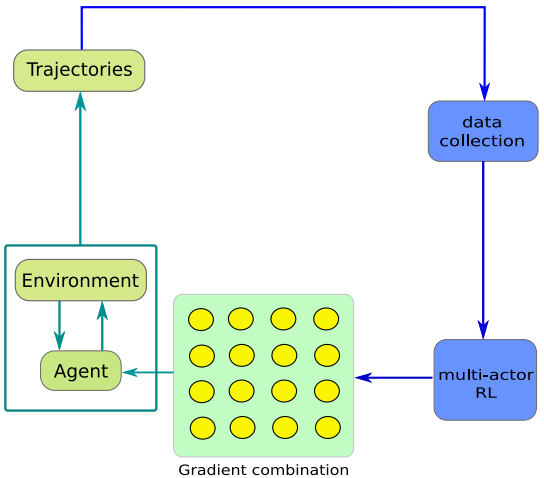


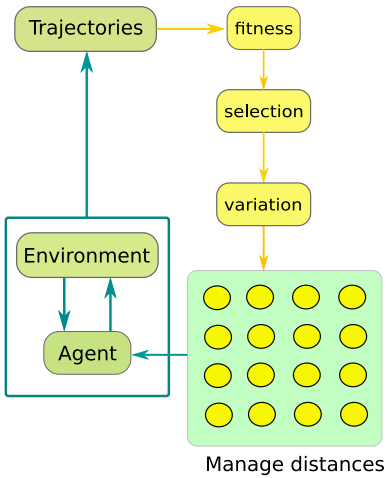
- ARAC: Only data from the most novel agents are sent to the replay buffer
- P3S-TD3: attraction towards the top agent
- P3S-TD3: No BC space, distance in policy param space

Doan, Thang and Mazoure, Bogdan and Durand, Audrey and Pineau, Joelle and Hjelm, R. Devon (2019) Attraction-Repulsion Actor-Critic for Continuous Control Reinforcement Learning, *arXiv preprint arXiv:1909.07543*

Jung, W., Park, G., and Sung, Y. (2020) Population-guided parallel policy search for reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net

Diversity without combination: SVPG and DVD



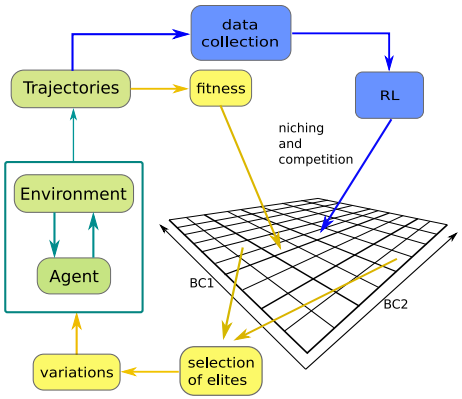


- SVPG: each agent is a particle, attraction and repulsion
- Models a distribution of high-performing policies with SVGD
- DVD: maximize the volume between agents (global vs one-to-one)
- Both: No BC space, distance in policy param space

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017) Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*

Parker-Holder, J., Pacchiano, A., Choromanski, K., and Roberts, S. (2020) Effective diversity in population-based reinforcement learning. *arXiv preprint arXiv:2002.00632*

QD Methods

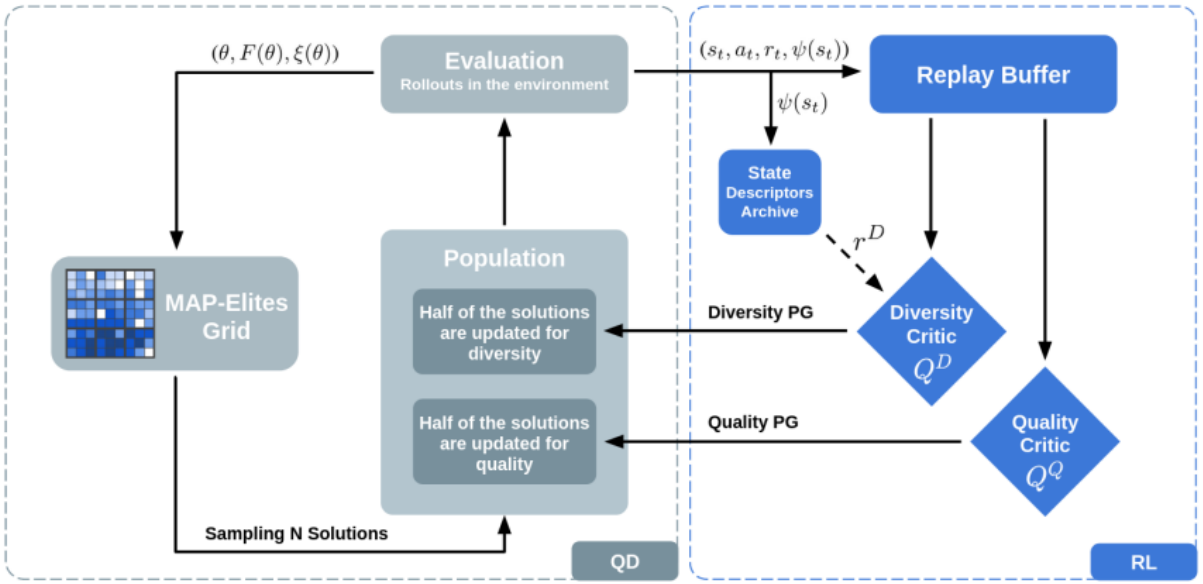


Algo. \ Prop.	Q. improvement	D. improvement
PGA-ME [59]	TD3 or GA	TD3 or GA
QD-PG-PF [8]	TD3	TD3
QD-PG-ME [64]	TD3	TD3
CMA-MEGA-ES [90]	CMA-ES	CMA-ES
CMA-MEGA-(TD3, ES) [90]	TD3 + CMA-ES	CMA-ES

- RL and evo agents are inserted in the archive if they win the competition (Pareto front or Map-Elites)
- One can use evolution and/or RL to improve quality and/or diversity
- All combinations exist

Pierrot, T., Richard, G., Beguir, K., and Cully, A. (2022b) Multi-objective quality diversity optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 139–147

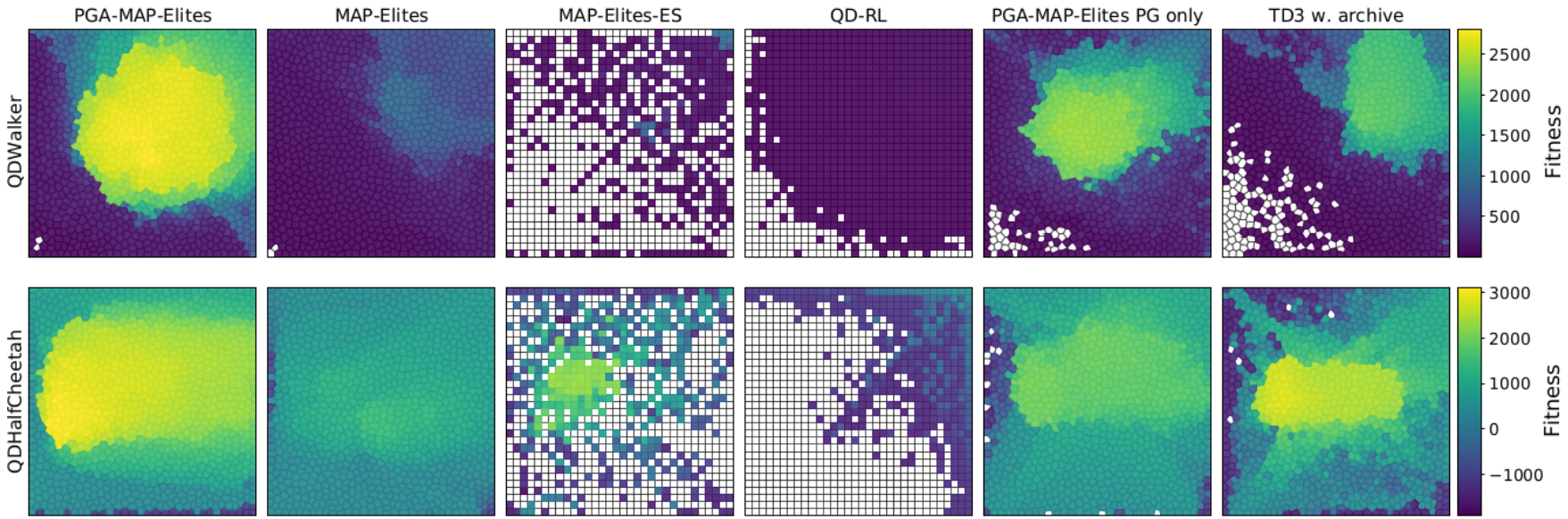
QDPG



- Two separate diversity and quality critics
- Uses a state descriptor in addition to BD to favor step-based diversity

Pierrot, T., Macé, V., Chalumeau, F., Flajolet, A., Cideron, G., Beguir, K., Cully, A., Sigaud, O., and Perrin-Gilbert, N. (2022) Diversity policy gradient for sample efficient quality-diversity optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1075–1083

PGA-ME Archives

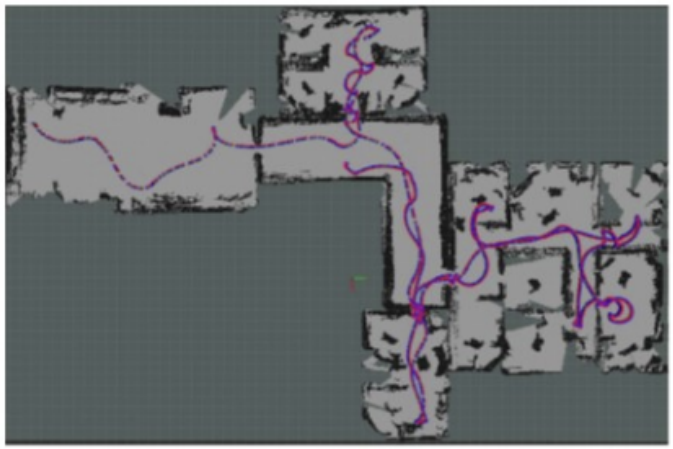


- Note the low covering of ME-ES and QD-PG
- A very active domain

Colas, C., Huizinga, J., Madhavan, V., and Clune, J. (2020) Scaling map-elites to deep neuroevolution. *arXiv preprint arXiv:2003.01825*

Nilsson, O. and Cully, A. (2021) Policy gradient assisted Map-Elites. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 866–875

Final remarks



- Looking for diversity is key to solving hard exploration problems (sparse rewards)
- A growing field, moving from evo conferences (GECCO, ECC, ...) to machine learning conferences (NeurIPS, ICLR, ...)
- A lot of questions remain to be investigated
- Combination algorithms have more potential, but more hyper-parameters
- Population-based training facilitates hyper-parameter tuning

Doncieux, S., Laflaquière, A., and Coninx, A. (2019) Novelty search: a theoretical perspective. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 99–106. ACM

Any question?

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

Colas, C., Huizinga, J., Madhavan, V., and Clune, J. (2020).

Scaling map-elites to deep neuroevolution.
arXiv preprint arXiv:2003.01825.

Doan, T., Mazoure, B., Durand, A., Pineau, J., and Hjelm, R. D. (2019).

Attraction-repulsion actor-critic for continuous control reinforcement learning.
arXiv preprint arXiv:1909.07543.

Doncieux, S., Laflaquière, A., and Coninx, A. (2019).

Novelty search: a theoretical perspective.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 99–106. ACM.

Jung, W., Park, G., and Sung, Y. (2020).

Population-guided parallel policy search for reinforcement learning.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lehman, J. and Stanley, K. O. (2011).

Abandoning objectives: Evolution through the search for novelty alone.
Evolutionary computation, 19(2):189–223.

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017).

Stein variational policy gradient.
arXiv preprint arXiv:1704.02399.

Nilsson, O. and Cully, A. (2021).

Policy gradient assisted map-elites.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 866–875.

Parker-Holder, J., Pacchiano, A., Choromanski, K., and Roberts, S. (2020).

Effective diversity in population-based reinforcement learning.
arXiv preprint arXiv:2002.00632.

Pierrot, T., Macé, V., Chalumeau, F., Flajolet, A., Cideron, G., Beguir, K., Cully, A., Sigaud, O., and Perrin-Gilbert, N. (2022a).

Diversity policy gradient for sample efficient quality-diversity optimization.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1075–1083.

Pierrot, T., Richard, G., Beguir, K., and Cully, A. (2022b).

Multi-objective quality diversity optimization.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 139–147.

Shi, L., Li, S., Zheng, Q., Yao, M., and Pan, G. (2020).

Efficient novelty search through deep reinforcement learning.
IEEE Access, 8:128809–128818.

ISIR
INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE