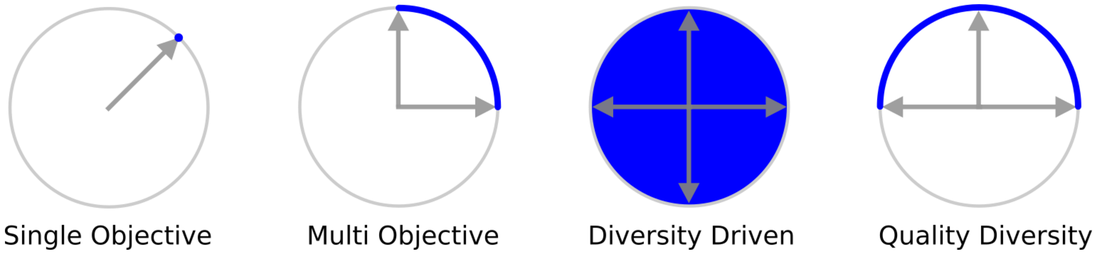
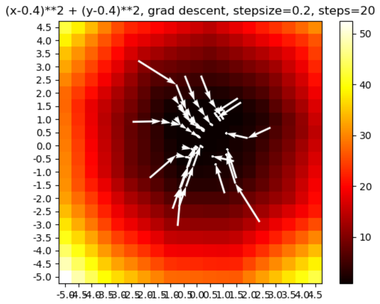
Think Before You Act: Generating a High-Quality Repertoire of Diverse Policies

Ryan Boldi, Matthew Fontaine, Stefanos Nikolaidis
University of Southern California

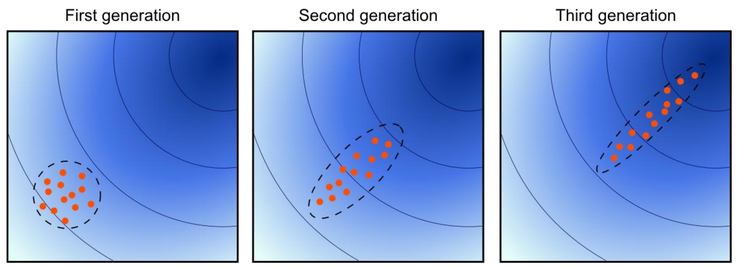
Quality Diversity Optimization



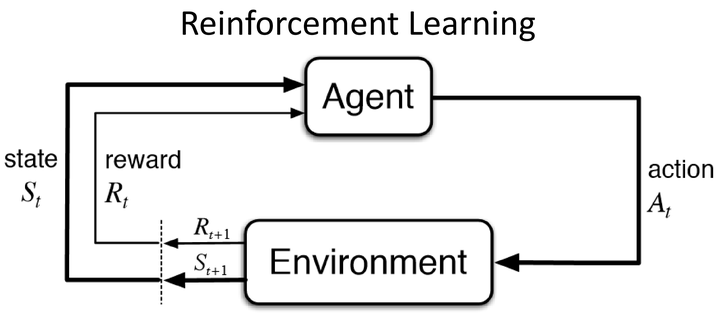
Gradient Descent



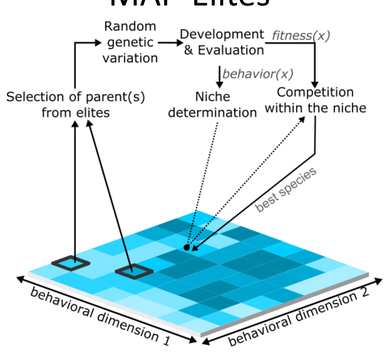
Covariance Matrix Adaptation



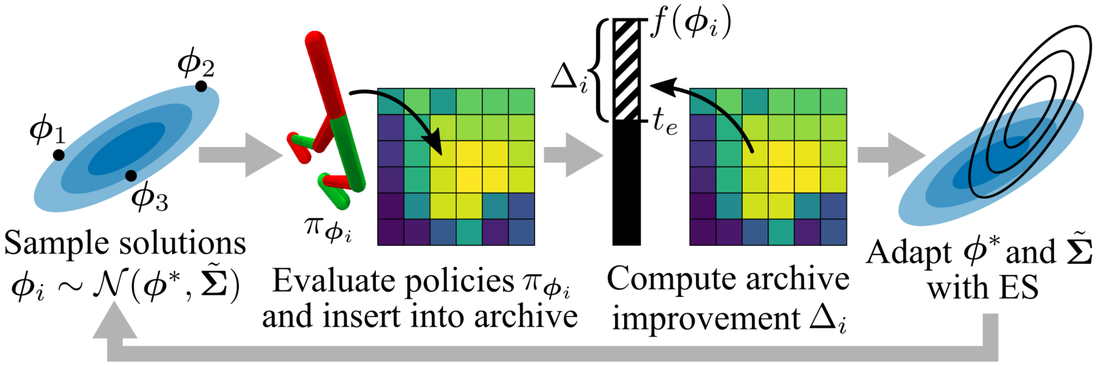
Reinforcement Learning



MAP-Elites



Covariance Matrix Adaptation MAP-Annealing



Using a Q-Function as a Policy

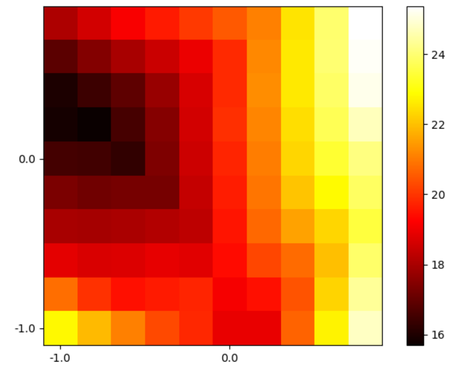
$Q(s,a) \to \mathbb{R}$: Value of taking action a in state s

$\pi(s) \to a$: A policy

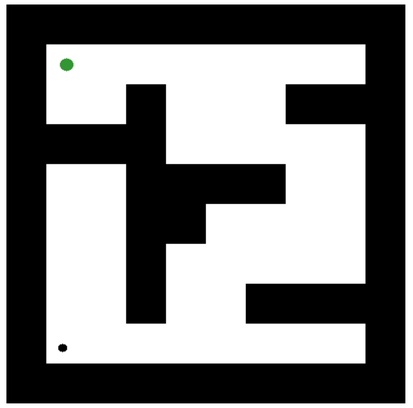
$\pi(s) = \arg\max_a Q(s,a)$

How to compute argmax in a continuous action space?

Gradient Descent!
- Non differentiable process, but this doesn't matter as we are using a gradient free optimizer



Domain: Hard Maze

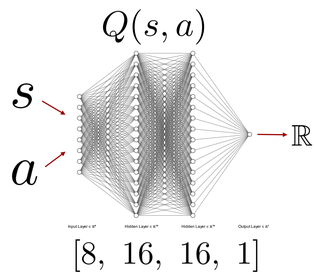


Inputs: $[x, y, x_{\text{vel}}, y_{\text{vel}}, x_{\text{acc}}, y_{\text{acc}}]$

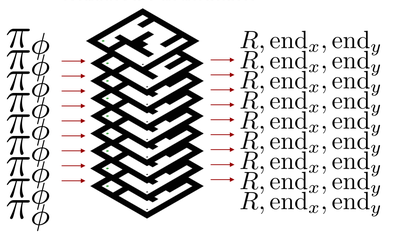
Outputs: $[\Delta x_{\text{acc}}, \Delta y_{\text{acc}}]$

Reward: $707.1 - \sqrt{(g_x - x)^2 + (g_y - y)^2}$
$- 100\mathbb{I}[\text{collided?}]$
$- \|\text{vel}\|_2^2 \, \mathbb{I}[\text{done?}]$

Policy and Implementation Details



Massive Parallelism



Results

Best Objective: 700+!

References

Matthew Fontaine and Stefanos Nikolaidis. 2023. Covariance Matrix Adaptation MAP-Annealing. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '23). Association for Computing Machinery, New York, NY, USA, 456–465. https://doi.org/10.1145/3583131.3590389

Hansen N, Ostermeier A (2001). Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation, 9(2) pp. 159–195.

Bryon Tjanaka, Matthew C Fontaine, David H Lee, Yulun Zhang, Nivedit Reddy Balam, Nathaniel Dennler, Sujay S Garlanka, Nikitas Dimitri Klapsis, and Stefanos Nikolaidis. 2023. Pyribs: A Bare-Bones Python Library for Quality Diversity Optimization. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '23). Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3583131.3590374

Acknowledgements

The authors would like to thank Bryon Tjanaka, Lee Spector, the PUSH Lab at Amherst college and the ICAROS Lab at USC for discussions that helped shape this work. The REU program is supported by the National Science Foundation under Grant No. CNS-2051117.