Ward Gauderis
0588485
01/06/2023

Reinforcement Learning
Faculteit Wetenschappen
Vrije Universiteit Brussel

Fundamental project
Gradient-Free Policy Optimisation

1 Introduction

Introduce problem

Objective of RL: maximize expected rewards

Policy gradient: parametric, maximize objective of reward times probability of action gradient descent

local information: local + learning rate

gradient free: local search with perturbation Less local optima, not sample efficient

zeroth-order optimisation

population-based optimisation: no learning rate

Introduce algorithms

2 Literature Review

HO + RL book

Literature review: alternative solutions

Particle swarm

simulated annealing

stochastic optimisation

genetic algorithms

3 Methods

Input: π_θ: parametric policy
Input: σ: standard deviation
Input: E: number of evaluation episodes
Input: α: learning rate
Loop
  sample $p \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$
  $r^+ \leftarrow$ evaluate$(\pi_{\theta+p}, E)$
  $r^- \leftarrow$ evaluate$(\pi_{\theta-p}, E)$
  $\Delta \leftarrow \frac{r^+ - r^-}{2} p$
  $\pi_\theta \leftarrow \pi_{\theta+\alpha\Delta}$

Algorithm 1: Zeroth-order optimisation

Input: π_θ: parametric policy
Input: σ: standard deviation
Input: E: number of evaluation episodes
Input: N: population size
Loop
  $r_{\text{max}} \leftarrow -\infty$
  repeat N **times**
    sample $p \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$
    $r \leftarrow$ evaluate$(\pi_{\theta+p,E})$
    if $r \geq r_{max}$ **then**
      $r \leftarrow r_{\text{max}}$
      $\pi_\theta \leftarrow \pi_{\theta+p}$

Algorithm 2: Population-based optimisation

Both algorithms were run for 10 000 iterations in the gym environment 'Lunar Lander' with a time limit of 500 steps per episode. This limit was chosen to encourage the agents to learn efficient policies and avoid long simulation times while providing enough to solve every episode. An episode is concidered to be solved by an agent when a cumulative reward above 200 is obtained.

After reasonable fine-tuning of the hyper-parameters, the following values were found to be suitable the zeroth-order method:

- Standard deviation $\sigma = 0.05$
- Number of evaluation episodes $E = 1$

In the experimental results, we investigate the influence of the number of evaluation episodes E on the performance of the agent. For the population-based method, the size of the perturbations was decreased and we will research the combined impact of the population size N and number of evaluation episodes E on the results:

- Standard deviation $\sigma = 0.01$

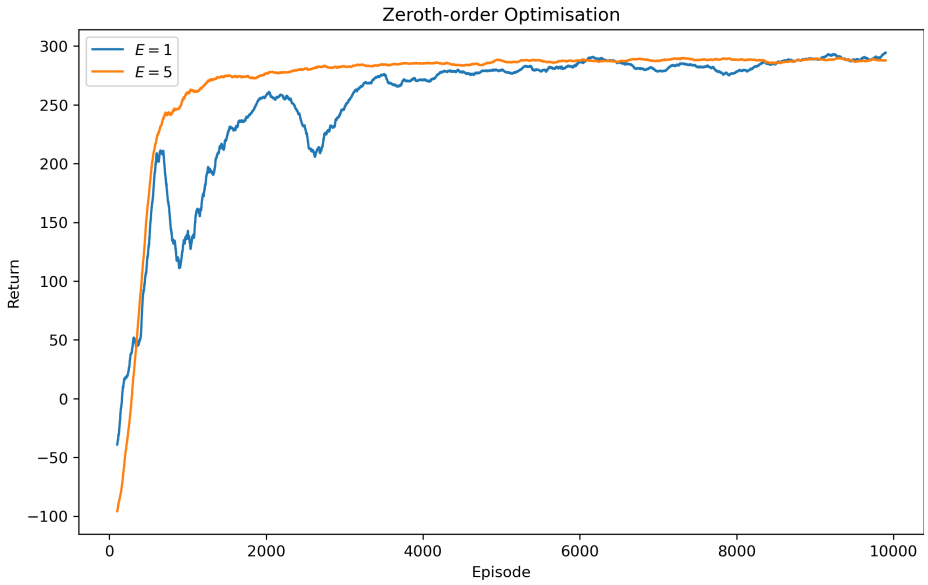

Figure 1: Cumulative reward per episode for the zeroth-order agents (averaged over 200 episodes)

4 Results

For readability the x-axis of the figures displays only the average cumulative reward per iteration of the algorithm. To compare the performance of the different algorithms in terms of the total number of evaluations used, one should

Readability multipliplied by eval

5 Conclusion

References

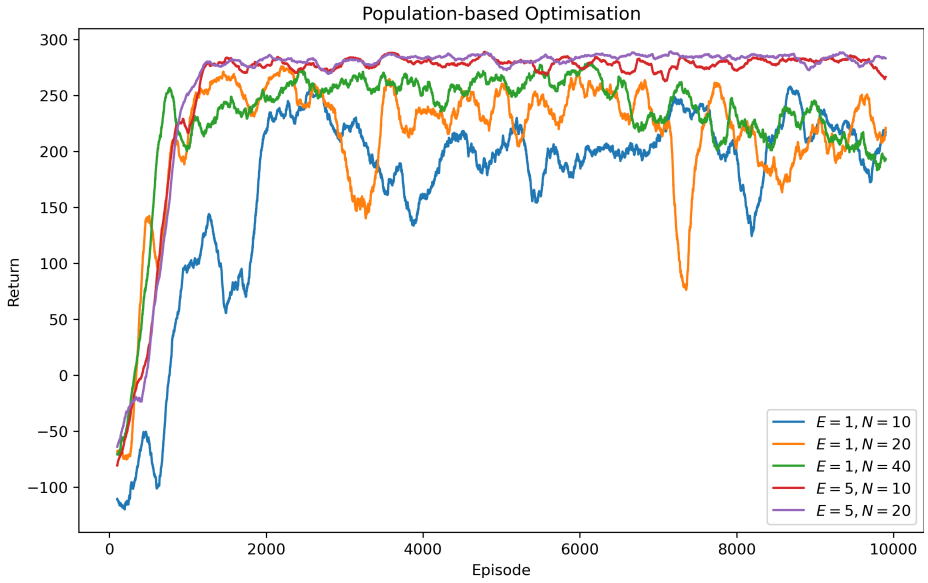

Figure 2: Cumulative reward per episode for the population-based agents (averaged over 200 episodes)