Quality-Diversity Algorithms vs. PPO Robustness Under Adversarial Perturbations in Continuous Control

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does the robustness of policies evolved via quality-diversity algorithms compare to standard PPO when evaluated under adversarial perturbations in continuous control tasks. The increasing importance of robots and automation creates a demand for learnable controllers which can be obtained through various approaches such as Evolutionary Algorithms (EAs) or Reinforcement Learning (RL). Unfortunately, these two families of algorithms have mainly. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Competitiveness of MAP-Elites against Proximal Policy Optimization on locomotion tasks in deterministic simulations. Research question: How does the robustness of policies evolved via quality-diversity algorithms compare to standard PPO when evaluated under adversarial perturbations in continuous control tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

16 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 4.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The comparisons of the different evaluations are based on statistics over the episode reward (i.e., cumulative episode r	×	0.02
The statistical significance of performance difference is verified through Wilcoxon Signed-Rank Test and p-values are re	×	0.01
The open-loop controllers studied in this work take a scalar time-modulo-period as an input and output 18 target angular	×	0.03
The period is always set arbitrarily to one second.	×	0.00
All of the used open-loop architectures have two layers of hidden neurons with two to six neurons in each layer.	×	0.01
Biases were used, which results in total policy parameter count ranging from 64 to 130.	×	0.03
The hyper-parameter selection is done in two phases.	×	0.08
First, 370 unique PPO hyper-parameter configurations are tested for a short horizon (75M frames) with four replications	×	0.06
Then, the four best configurations (according to the median episode reward) are executed for a longer horizon (255M fram	×	0.01
The hyper-parameter configurations sampled for the initial assessment consist of learning rate: [5e-5,1e-2] and clipping	×	0.04
An entropy term is sampled log-uniformly from the range [1e-4,1e-2] in 25% cases or set to zero otherwise in the remaini	×	0.01
Additionally, among uniformly sampled hyper-parameters are mini-batch size, selected from the range of [2,32] Ki, and po	×	0.03
The locomotion task is the same throughout this paper - it is to walk as far as possible along an X-axis within an episo	×	0.03
This walked distance is referred to as the fitness or episode reward.	×	0.04
The environment is deterministic (i.e., the same action always leads to the same result) and simulated with the open-sou	×	0.03

References

- http://arxiv.org/abs/2211.02193v1
- http://arxiv.org/abs/2009.08438v2
- http://arxiv.org/abs/2211.13742v2