Exploring the Scalability and Adaptability of Evolution Strategies in Reinforcement Learning

Reinforcement Learning

Student group:

- Paolo Cursi - 2155622
- Pietro Signorino - 2149741

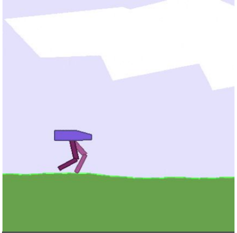
Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

The Task

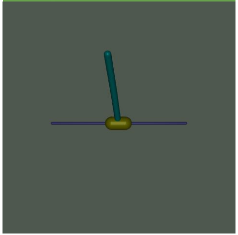
Bipedal Walker

The bipedal walker is a **two-legged robot** designed to mimic human walking, balancing dynamically using actuators and sensors.



Inverted Double Pendulum

The inverted double pendulum is a dynamic system with **two pivoting rods balancing upright**, used to study control and stability.



The Algorithms

Proximal Policy Optimization

Proximal Policy Optimization (PPO) is a reinforcement learning **algorithm** that aims to find an **optimal policy** for an **agent** to interact with its environment. PPO is considered one of the state-of-the-art algorithms in reinforcement learning.

Evolutionary Strategies

Evolutionary Strategies (ES) is a reinforcement learning algorithm that aims to find an optimal policy for an agent to interact with its environment by **evolving the policy parameters**. Unlike PPO, evolutionary strategies' **black box** nature offers greater flexibility by exploring solutions without predefined constraints

The Idea

- The idea is to take a **model** and **train** it on the **two environments**, using both PPO and ES

- These experiment gave us a solid ground to create other experiments!

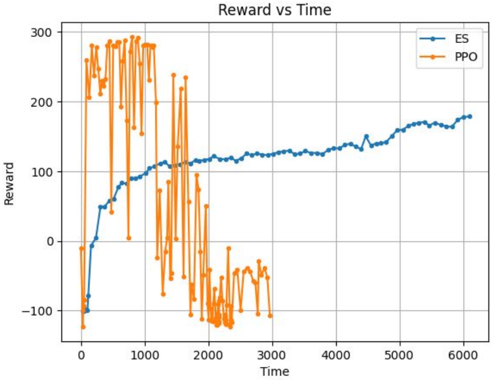
- The model architecture will be explained later

Bipedal Walker

Traning

The plot shows how the best model (at training time) of each generation performs in function of the time of the epoch.

PPO is much faster to converge, **ES** is slower but the training is less noisy.



Bipedal Walker

Results

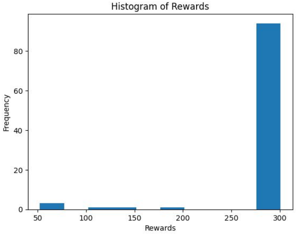
The histograms shows how **100 runs** of the best model performs.

The **PPO** best model performs **better** and is more consistent.

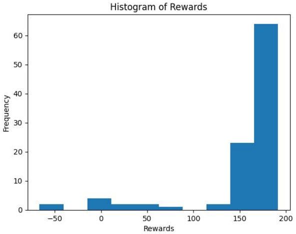
The best scores for both are:

- 300 (PPO)
- 191 (ES)

PPO



ES



Inverted Double Pendulum

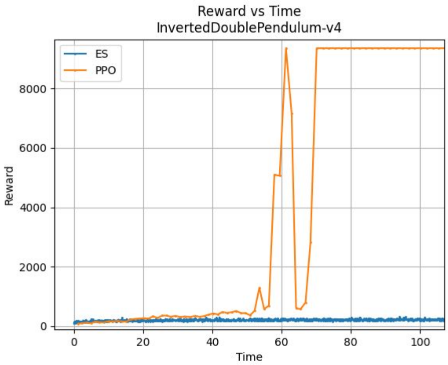
Training

PPO converges to a better solution (the optimum in this case).

ES does not converge to a good solution.

The best scores for both are:

- 9359 (PPO)
- 238 (ES)



ES does not converge in Double Inverted Pendulum

Why does ES not converge for the Double Inverted Pendulum task ?

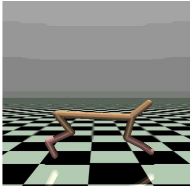
It is probably **related to the nature of the task**:

- In Bipedal Walker the agent earns incremental rewards for forward motion, joint angles, and energy efficiency. Even suboptimal policies yield gradients for improvement.
- In Inverted Double Pendulum it only gets a high reward if both poles are balanced upright. Most perturbations lead to **immediate failure** (near-zero reward). Failed episodes dominate the population, making it hard to estimate a useful search direction.

Is PPO always better?

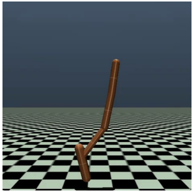
We tried testing PPO and ES on other environments

Half-Cheetah



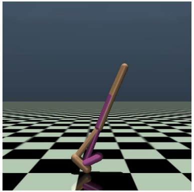
Half-Cheetah is an environment where a **bipedal robot** learns to **move forward** by applying joint torques.

Hopper



Hopper is an environment where a **one-legged robot** learns to **hop forward** by applying joint torques.

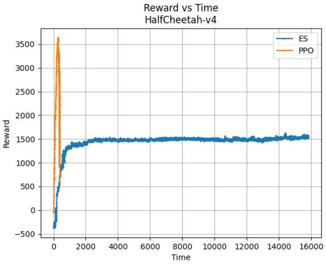
Walker2D



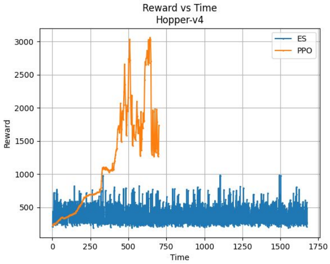
Walker2D is an environment where a **bipedal robot** learns to **walk forward** by applying joint torques.

Results

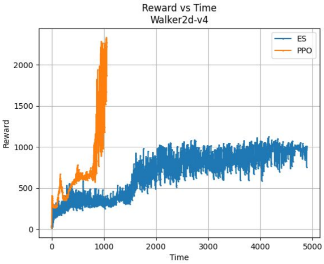
Half-Cheetah



Hopper



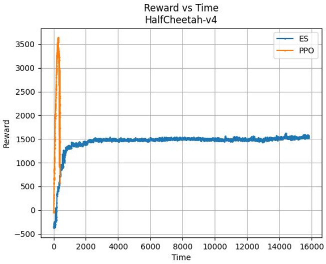
Walker2D



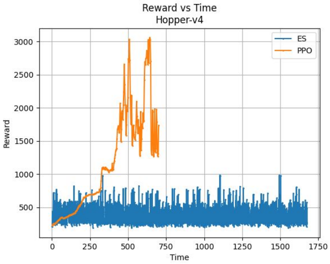
PPO is better in every environment

Results

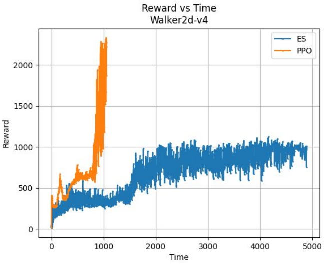
Half-Cheetah



Hopper



Walker2D



PPO is better in every environment

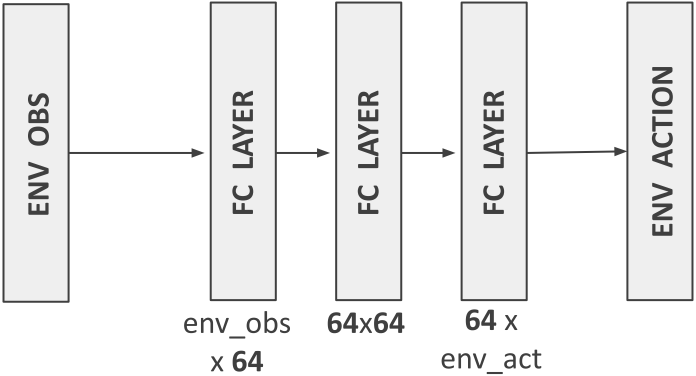
but why?

Results Analysis

- As stated before the nature of the task might be crucial, but…
- The **model choice** is also very **important**, we hypothesize that with different model choices ES would be better
- We want to explore this hypothesis, so we tried to train using different models

Model choice analysis

This is the **model** we **used** to **train** with PPO and ES up until now:



Model choice analysis

The PPO model also has **another component**, the **Critic** (or **Value**) **network**.

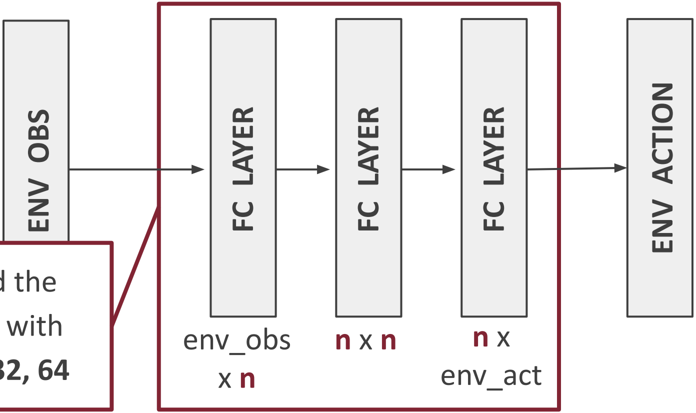
This component is not present in the model trained with ES.

It is composed of three linear layers of dimension

- (env_obs x **64**)
- (**64** x **64**)
- (**64** x **1**)

Model choice analysis

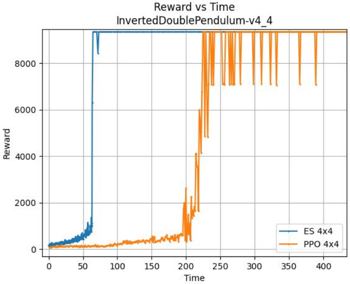
We tried different model dimensions, each model was built like so:



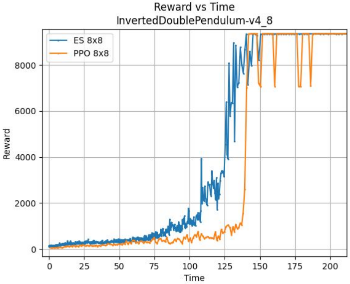
we repeated the experiments with **n** = **4, 8, 16, 32, 64**

Results - Inverted Double Pendulum

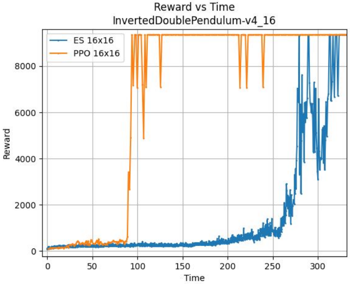
We tested our hypothesis on the **Inverted Double Pendulum** environment



n = 4



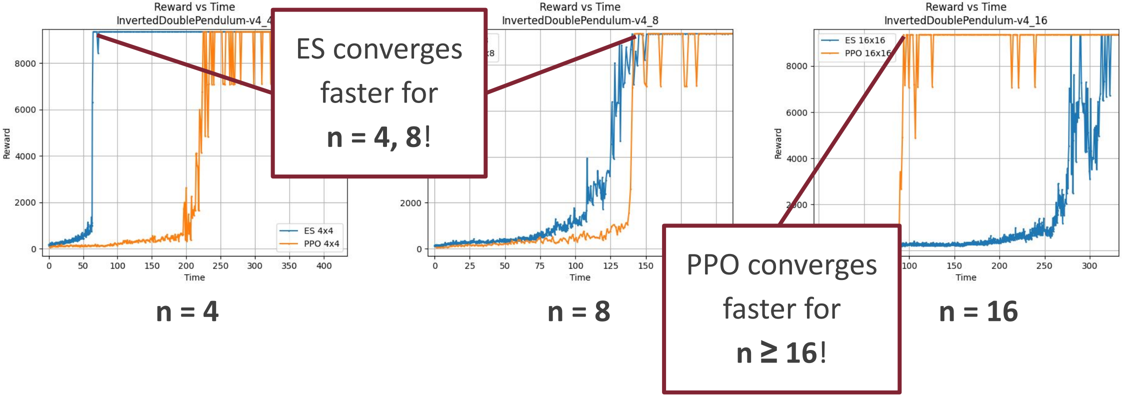
n = 8



n = 16

Results - Inverted Double Pendulum

We tested our hypothesis on the **Inverted Double Pendulum** environment



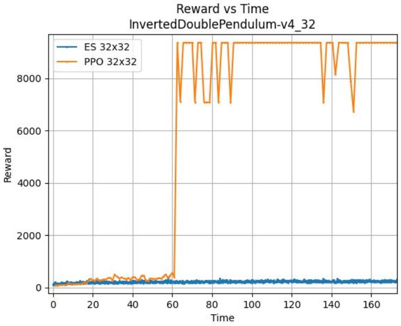
n = 4

n = 8

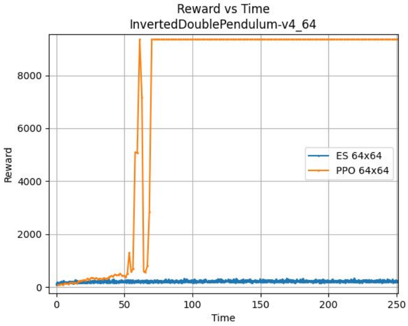
n = 16

Results - Inverted Double Pendulum

We tested our hypothesis on the **Inverted Double Pendulum** environment



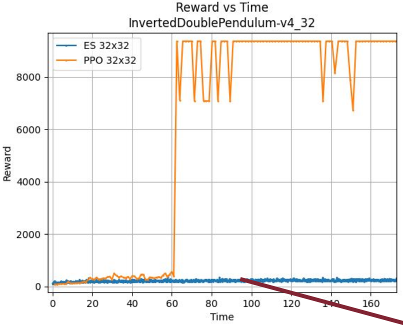
n = 32



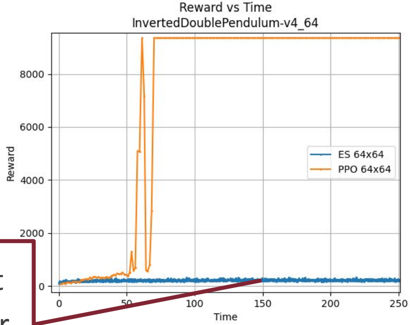
n = 64

Results - Inverted Double Pendulum

We tested our hypothesis on the **Inverted Double Pendulum** environment



n = 32



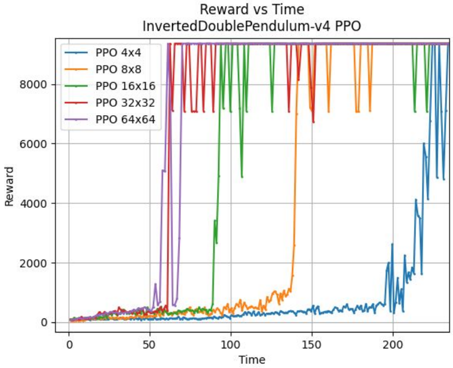
n = 64

ES does not converge for **n = 32, 64**!

Results - Inverted Double Pendulum

The plot shows the reward over time of the models with **n = 4, 8, 16, 32, 64** trained using PPO

Bigger models consistently **converge faster** than smaller ones!

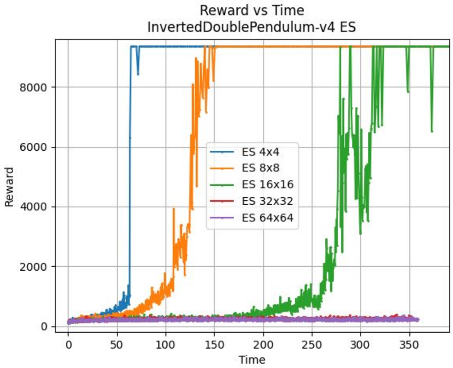


Results - Inverted Double Pendulum

The plot shows the reward over time of the models with **n = 4, 8, 16, 32, 64** trained using ES

We observe the **inverse trend** respect to PPO

Smaller models consistently **converge faster** than bigger ones!

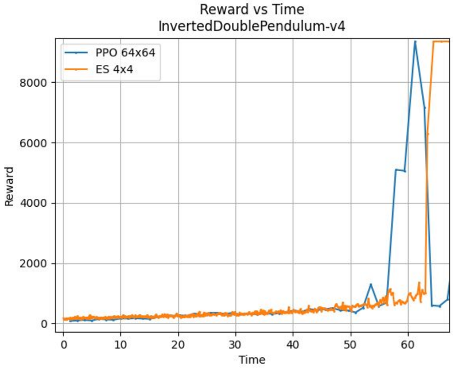


Results - Inverted Double Pendulum

What's the fastest training?

The plot shows the training of the model with **n = 64** using PPO and the one with **n = 4** using ES.

The training is **slightly faster** using PPO, **but** the **model** is **76 times bigger** (not to mention that it also has the Critic network)!



Known facts about the strategies

We studied **these** facts

	Evolution Strategies	Proximal Policy Optimization
PROS	-Black-box optimization -Good in sparse-reward environments -Avoids local optima	-Efficient SGD updates -**Scales well to large networks** -Excels with dense rewards
CONS	-**Cursed by dimensionality** -**Works well for compact policies**	-Struggles with sparse rewards -Prone to local optima

Conclusions

- Our experiments revealed that **Evolution Strategies** exhibit **distinct behavioral** patterns **depending** on the **environment**.

- We also analyzed how the **scalability** of both ES and PPO **is influenced** by the **dimensionality** of the policy **model**:
    - given the same **small model ES converges faster** than PPO
    - **PPO** converges **faster** with **bigger models**

Future Improvements

- **Repeat** these **experiments** with **all** the **other environments** mentioned (Bipedal Walker, Half-Cheetah, Hopper, Walker2D)

- Explore other facts about PPO and ES by creating **other experiments** like
  - using **more complex environments**
  - **modifying** even **more** the network **architecture**

Thank you for your attention!

Questions?