

Sprawozdanie z Projektu 5:

Uczenie przez wzmacnianie w przestrzeniach ciągłych

Kamil Poniewierski

Michał Burda

Grupa: 1 i 2

Maj 2025

1 Opis projektu

Celem projektu było zastosowanie algorytmu uczenia przez wzmacnianie w środowisku o przestrzeni stanów ciągłej. W projekcie wykorzystano bibliotekę `Gymnasium` oraz `Stable-Baselines3`. Przeanalizowano wpływ hiperparametrów na jakość uczenia, przetestowano różne architektury sieci neuronowych, zapisano najlepszego agenta i uruchomiono go w trybie deterministycznym.

2 Środowisko i algorytm

- **Środowisko:** `LunarLanderContinuous-v3` – zadanie polegające na bezpiecznym lądowaniu lądownika księżycowego z wykorzystaniem przestrzeni akcji i stanów ciągłych.
- **Algorytm:** TD3 (Twin Delayed Deep Deterministic Policy Gradient) – nowoczesny algorytm aktor-krytyk dedykowany przestrzeniom ciągłym.

3 Testowane strategie hiperparametrów

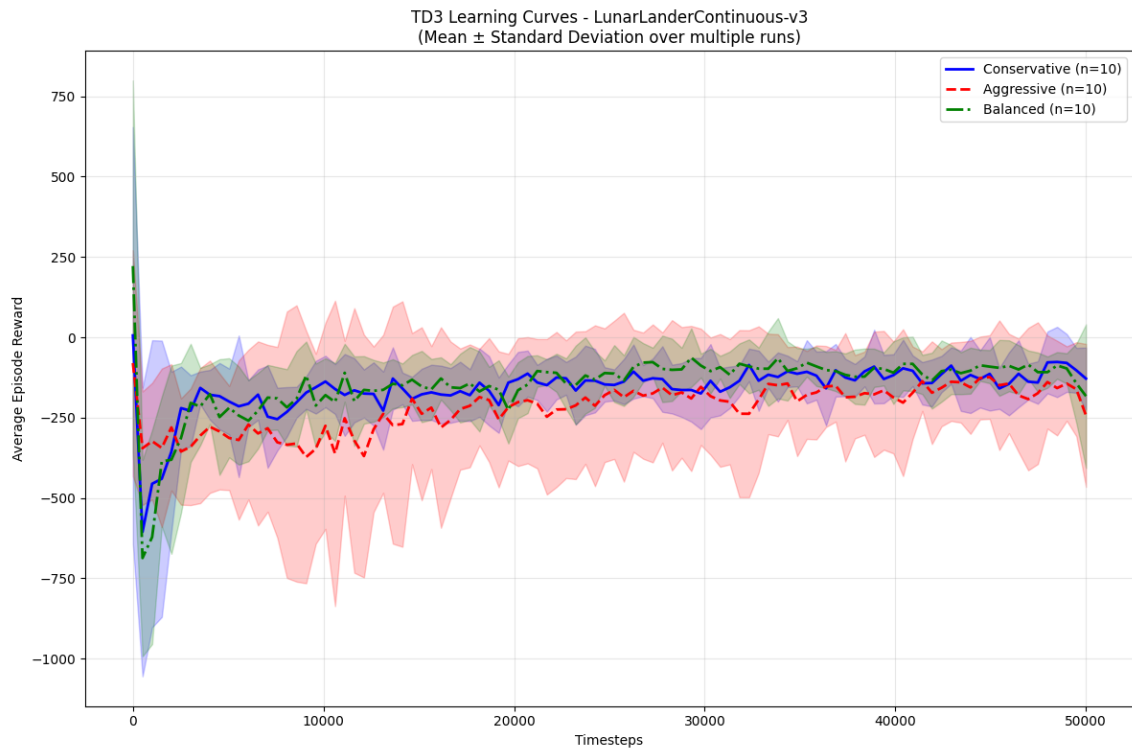
W projekcie przetestowano trzy strategie hiperparametrów: Conservative, Aggressive i Balanced. Ich szczegółowe wartości przedstawiono w tabeli 1.

Strategia	Learning Rate	Buffer Size	Batch Size	Gamma	Tau	Noise Sigma
Conservative	$3 \cdot 10^{-4}$	100000	64	0.99	0.005	0.1
Aggressive	$1 \cdot 10^{-3}$	200000	128	0.95	0.01	0.2
Balanced	$5 \cdot 10^{-4}$	150000	96	0.97	0.007	0.15

Tabela 1: Zestawy hiperparametrów użyte w eksperymencie

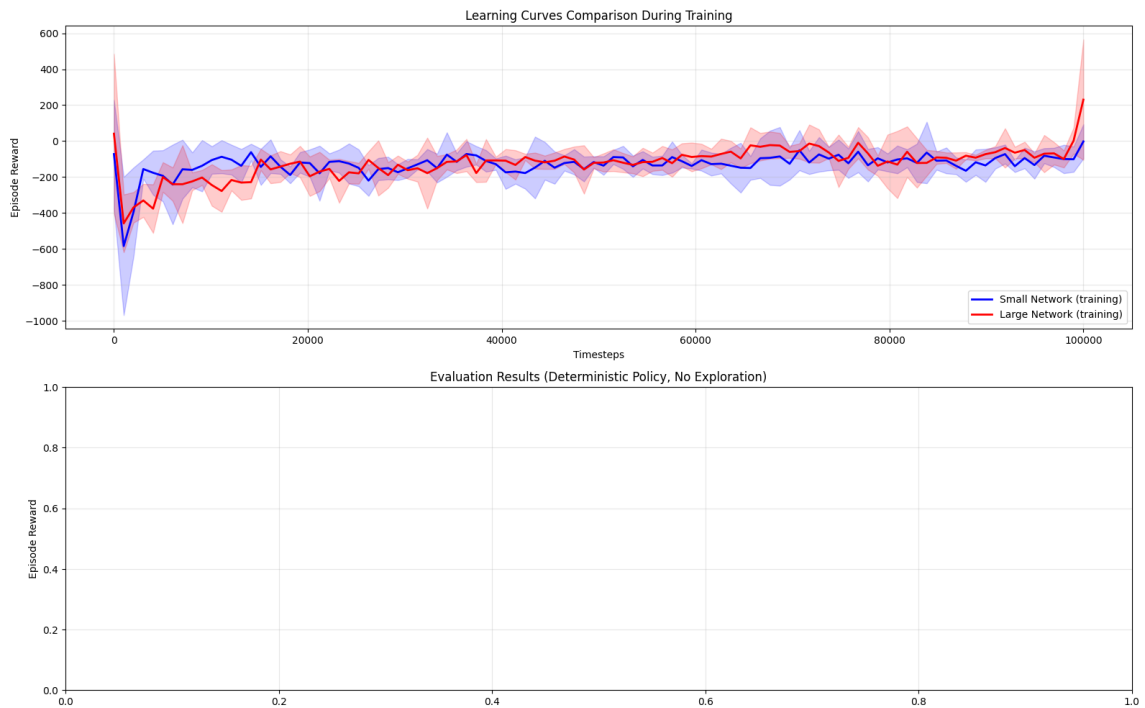
4 Krzywe uczenia

Na rysunku 1 przedstawiono krzywe uczenia algorytmu TD3 w środowisku `LunarLanderContinuous-v3` dla trzech różnych konfiguracji hiperparametrów.



Rysunek 1: TD3 – krzywe uczenia w środowisku LunarLanderContinuous-v3 dla strategii: Conservative, Aggressive, Balanced

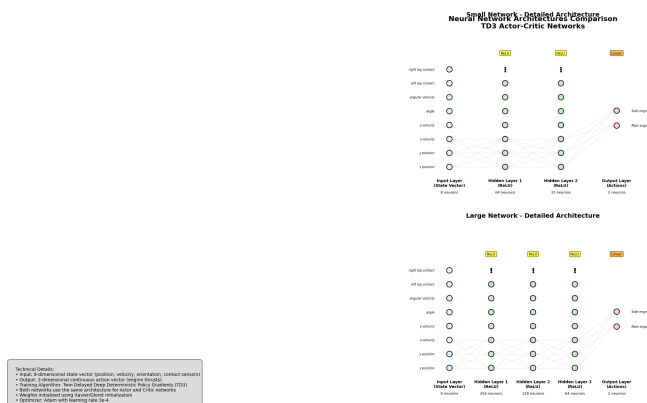
Strategia **Balanced** (zielona) osiąga najwyższe średnie nagrody i najniższą wariancję. Strategia **Aggressive** jest niestabilna, z dużymi odchyleniami.



Rysunek 2: Porównanie wyników treningu dla dwóch architektur sieci: Small vs Large

5 Architektury sieci neuronowych

Dwie testowane architektury przedstawiono na rysunku 3.



Rysunek 3: Schematy sieci: mała (64-32) i duża (256-128-64)

- **Small network:** wejście (8 wymiarów), warstwy: 64, 32, wyjście (2 wymiary)
- **Large network:** wejście (8), warstwy: 256, 128, 64, wyjście (2)

Dodatkowe wyniki uczenia przedstawia rysunek 4.



Rysunek 4: Wydłużony trening dla obu architektur. Dolny wykres wymaga przycięcia.

6 Zapis najlepszego agenta

Zapisano model agenta osiągającego najwyższą średnią nagrodę. Uruchomiono go w trybie deterministycznym (bez eksploracji).

- Średnia nagroda (40 epizodów): 33.38
- Mediana nagrody (40 epizodów): 1.76
- Znacząca różnica między średnią a medianą sugeruje obecność kilku wyjątkowo udanych epizodów oraz dużą wariancję.

Rysunek 5: Działanie najlepszego agenta – tryb deterministyczny

7 Wnioski

- Algorytm TD3 sprawdza się dobrze w środowisku LunarLanderContinuous-v3.
- Najbardziej stabilne wyniki dała strategia Balanced.
- Architektura sieci ma duży wpływ na szybkość i stabilność uczenia – większa sieć wypada lepiej, ale wolniej się uczy.
- Zachowanie agenta deterministycznego potwierdza skuteczność nauki mimo dużej wariancji.