Week 3

Advisor: Dr. Yue Wang

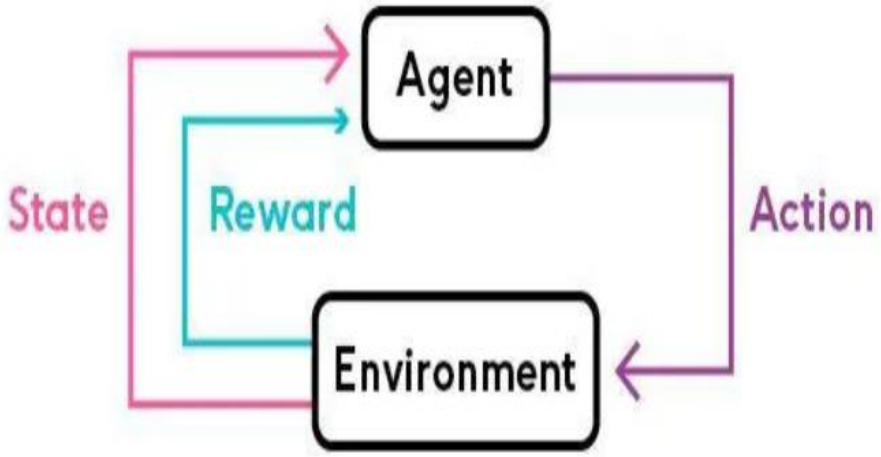
Student: David Tang

Project Overview

Multi-Objective Reinforcement Learning uses scalarized rewards

Methods: Linear | Nash Social Welfare

Goal: Investigate tradeoffs and behavior



$$w_1 r_1 + w_2 r_2 + w_3 r_3$$

$$(r_1)(r_2)(r_3)$$

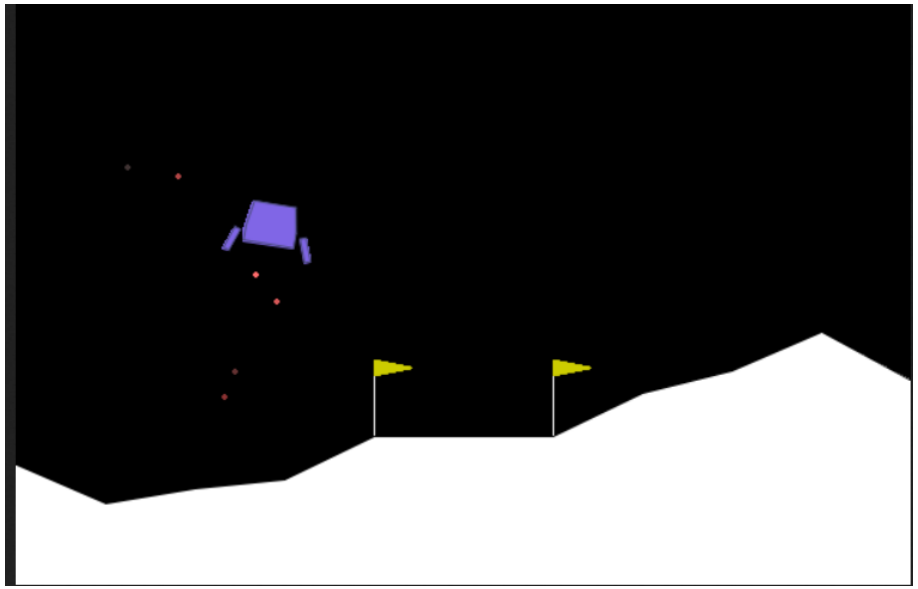
Environment Details

mo-lunar-lander-v3

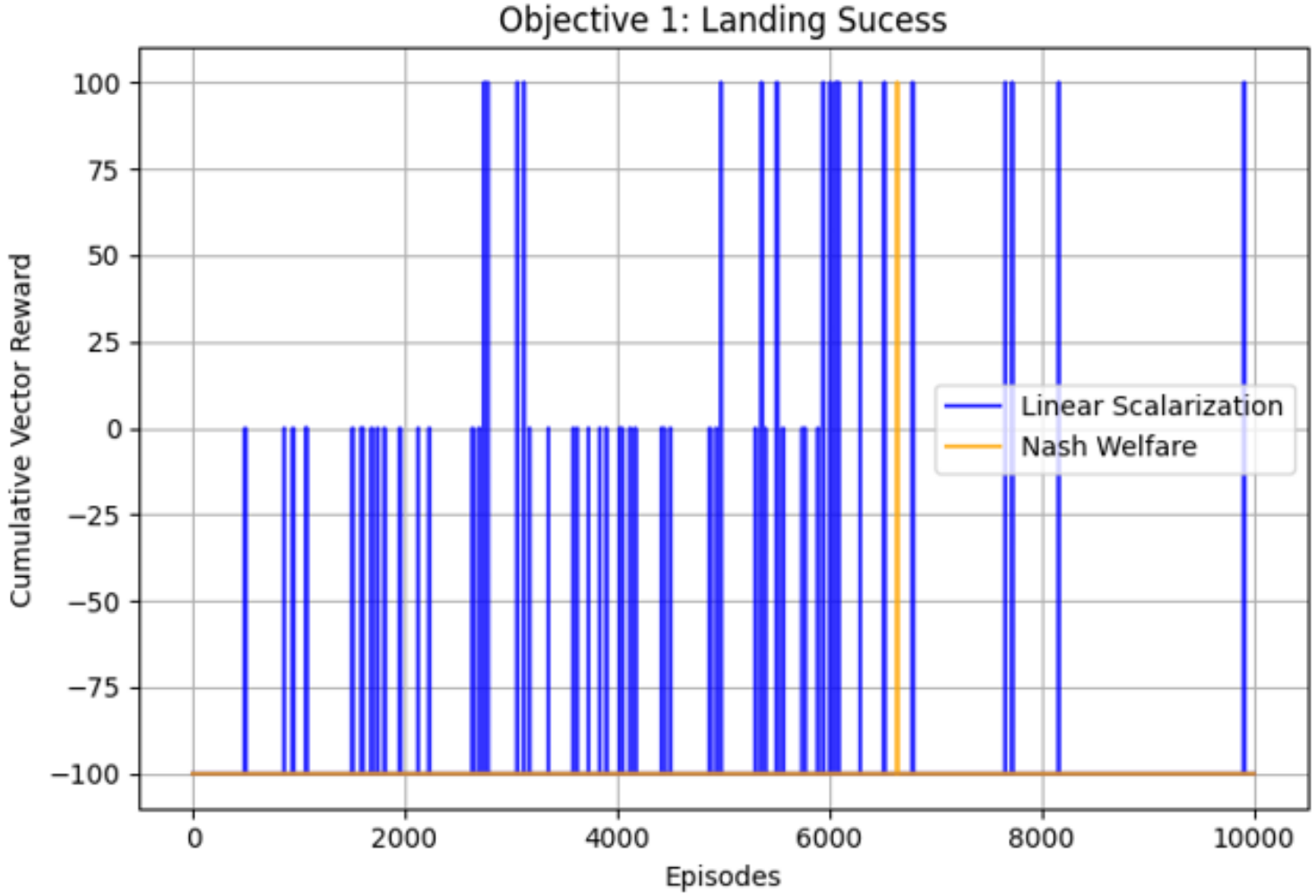
$$\mathbf{s} = [x \quad y \quad v_x \quad v_y \quad \theta \quad \omega \quad c_L \quad c_R] \in \mathbb{R}^8$$

$$\mathcal{A} \in \mathbb{R}^4$$

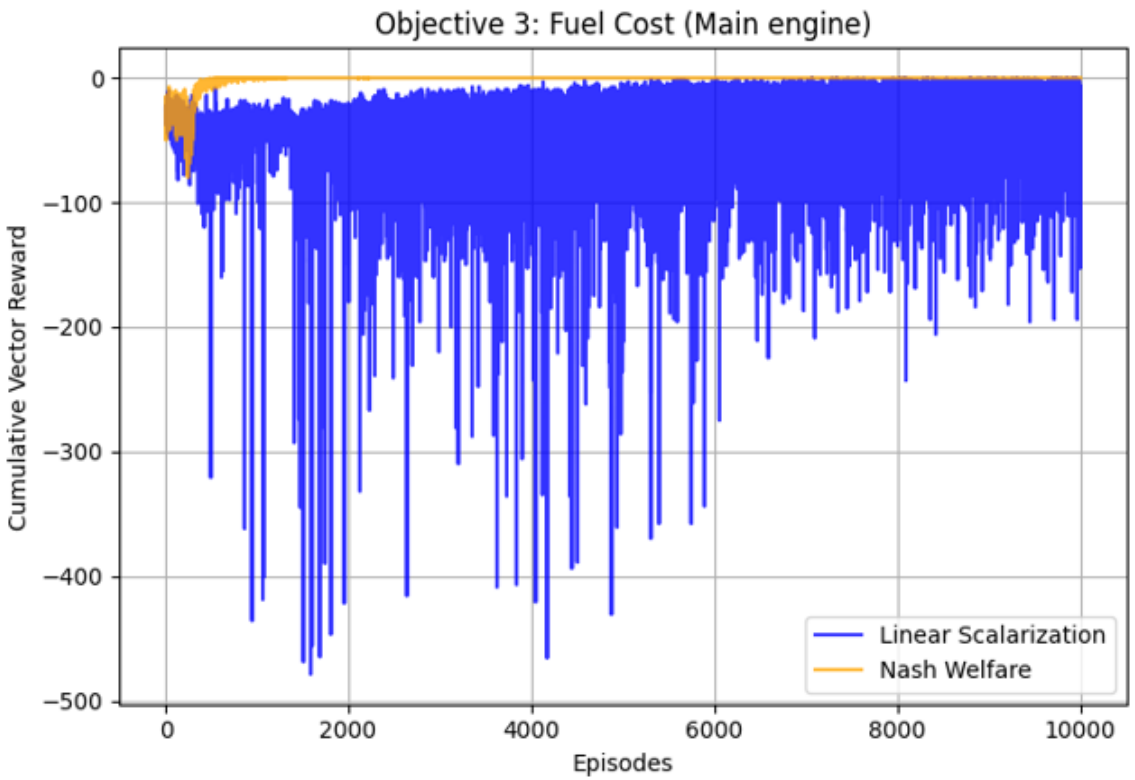
$$\mathbf{r} = \begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} \text{Terminal status } (\pm 100 \text{ for land/crash}) \\ \text{Shaping reward} \\ \text{Fuel cost (main engine)} \\ \text{Fuel cost (side engine)} \end{bmatrix}$$

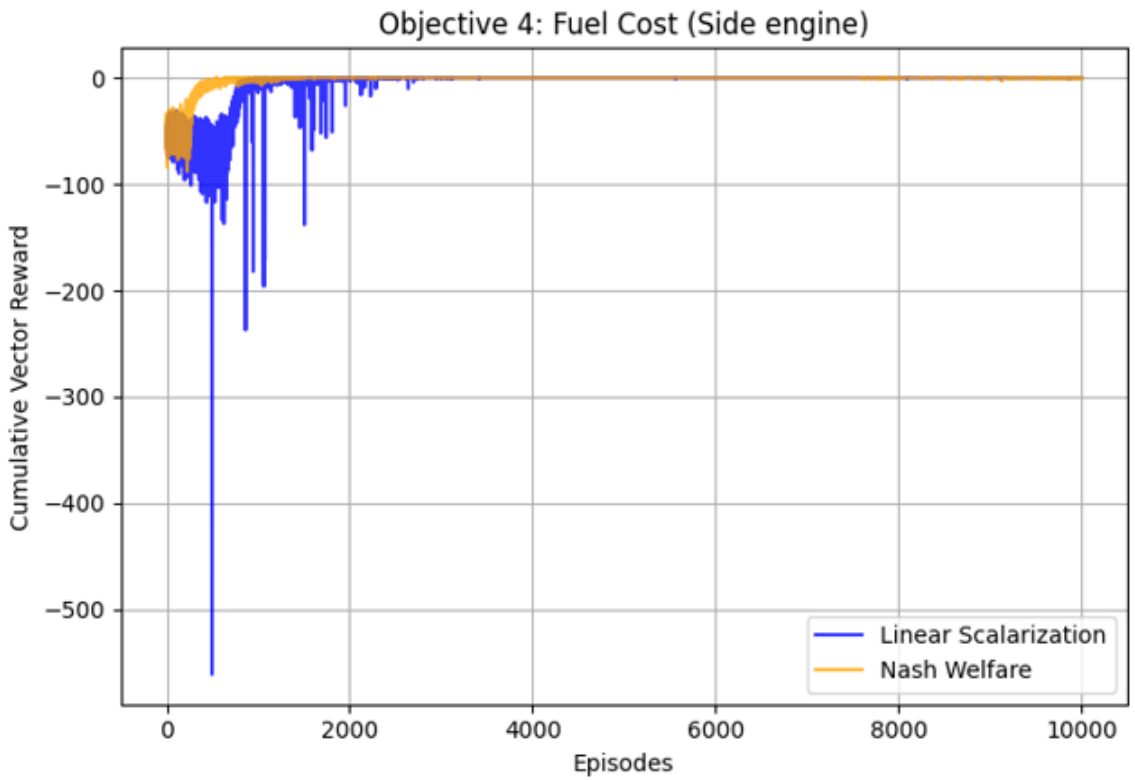


Advantage Actor Critic(A2C) Results

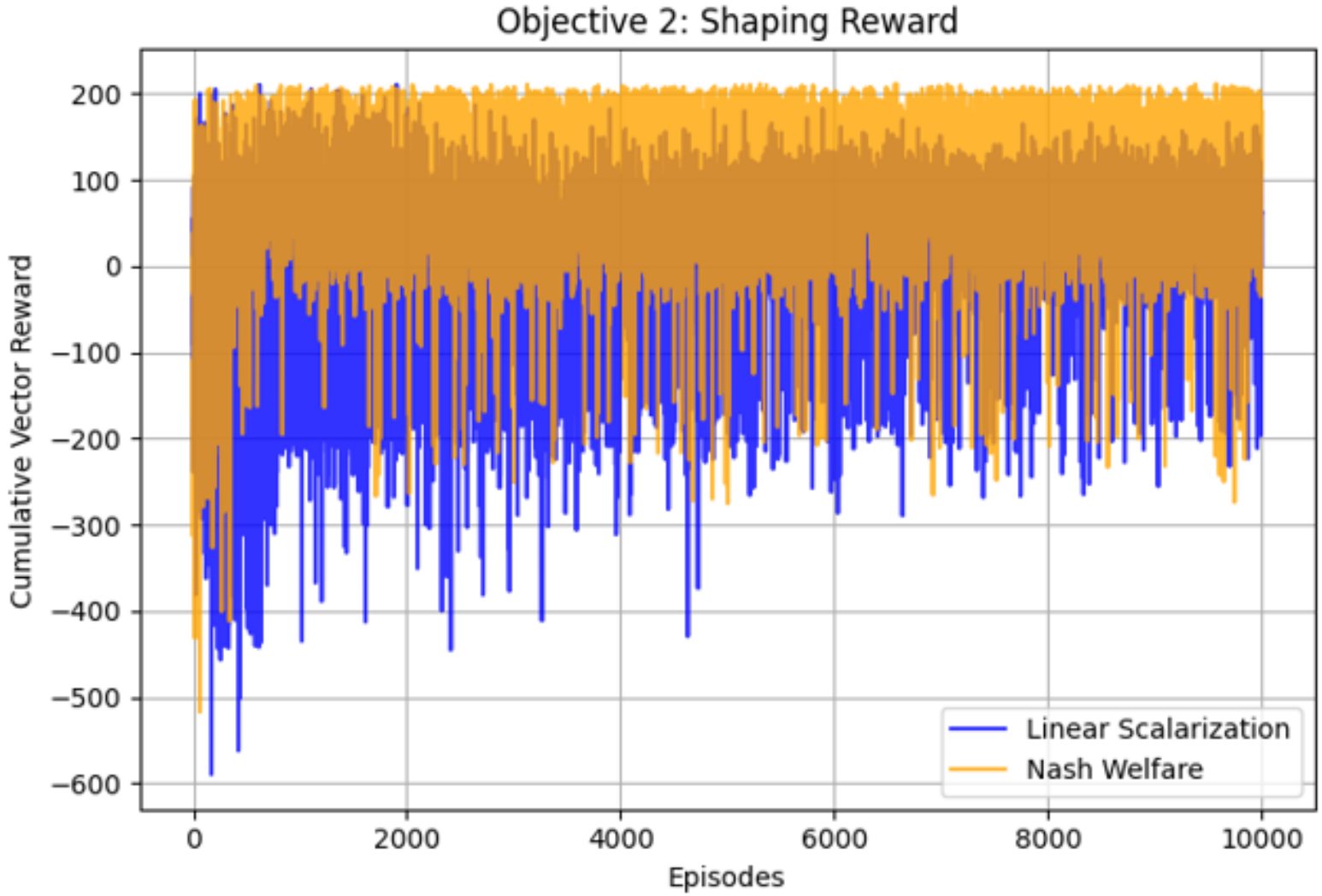


A2C Results Continued



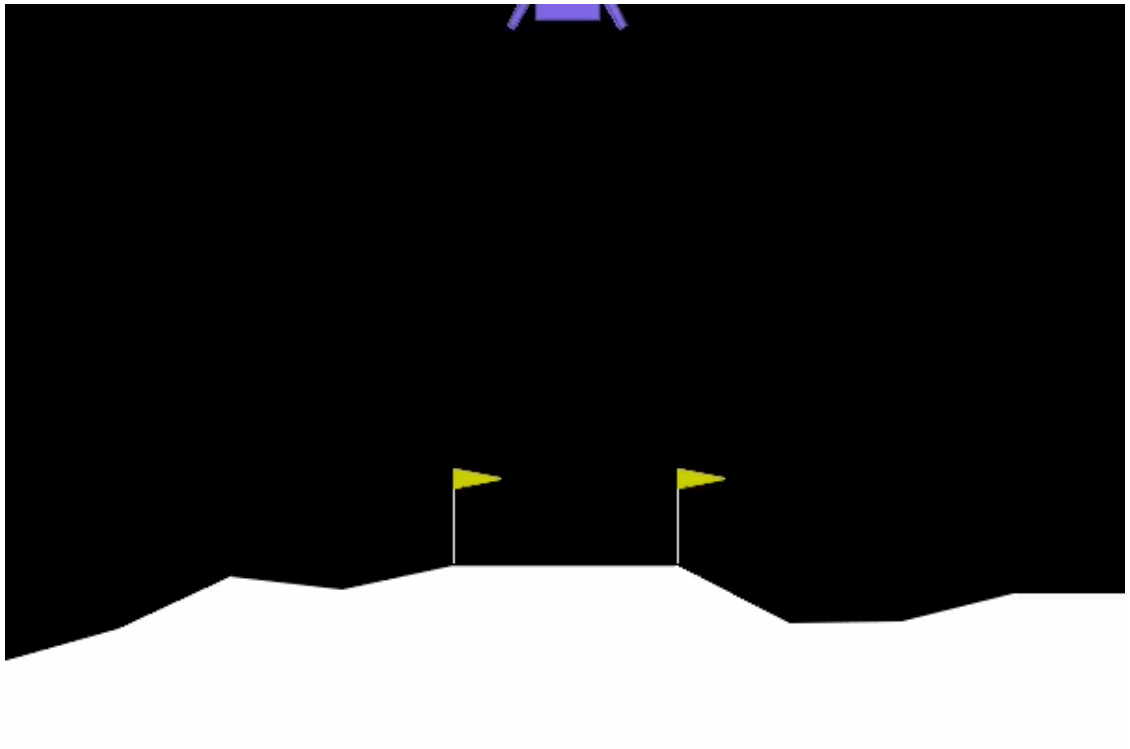


A2C Results Continued.

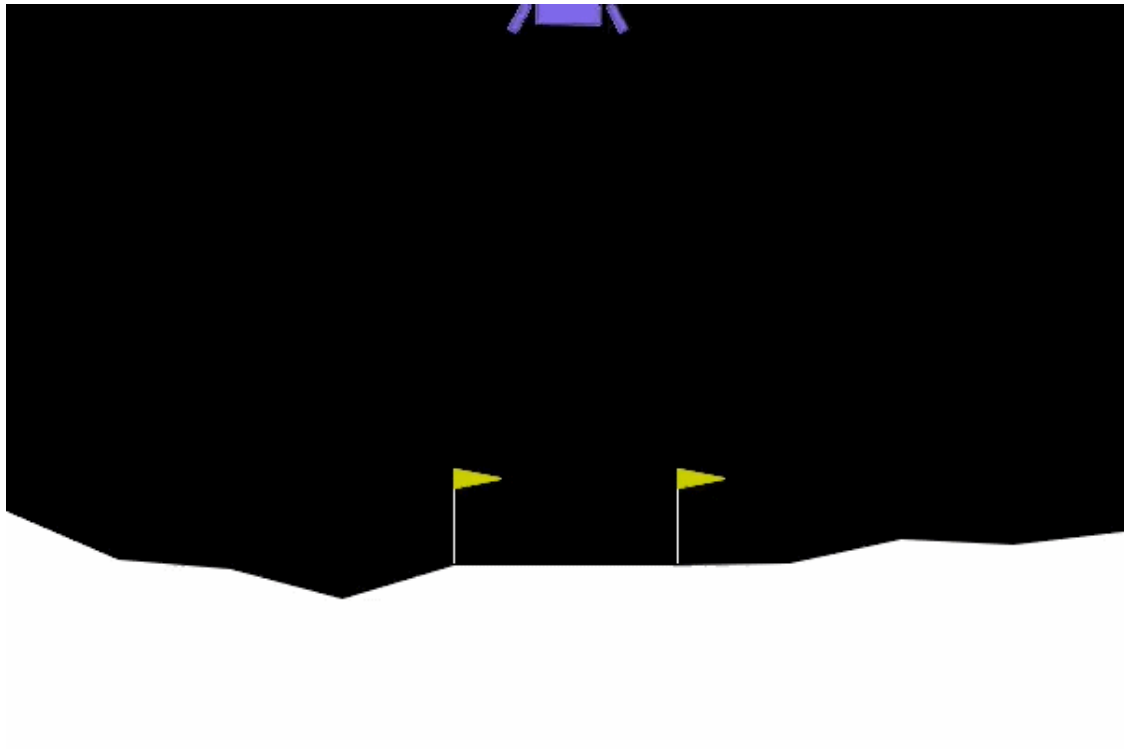


A2C Results Continued..

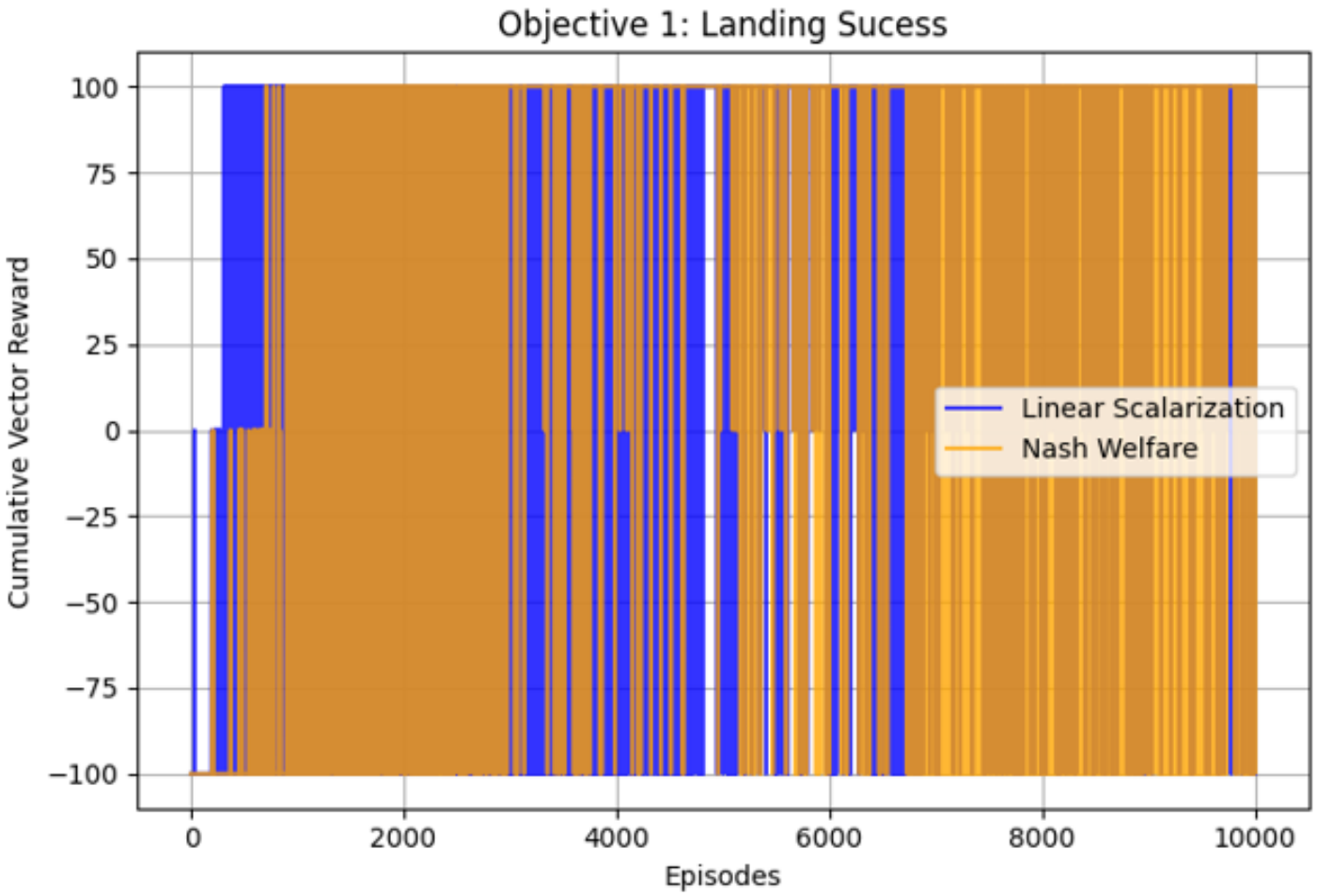
Linear



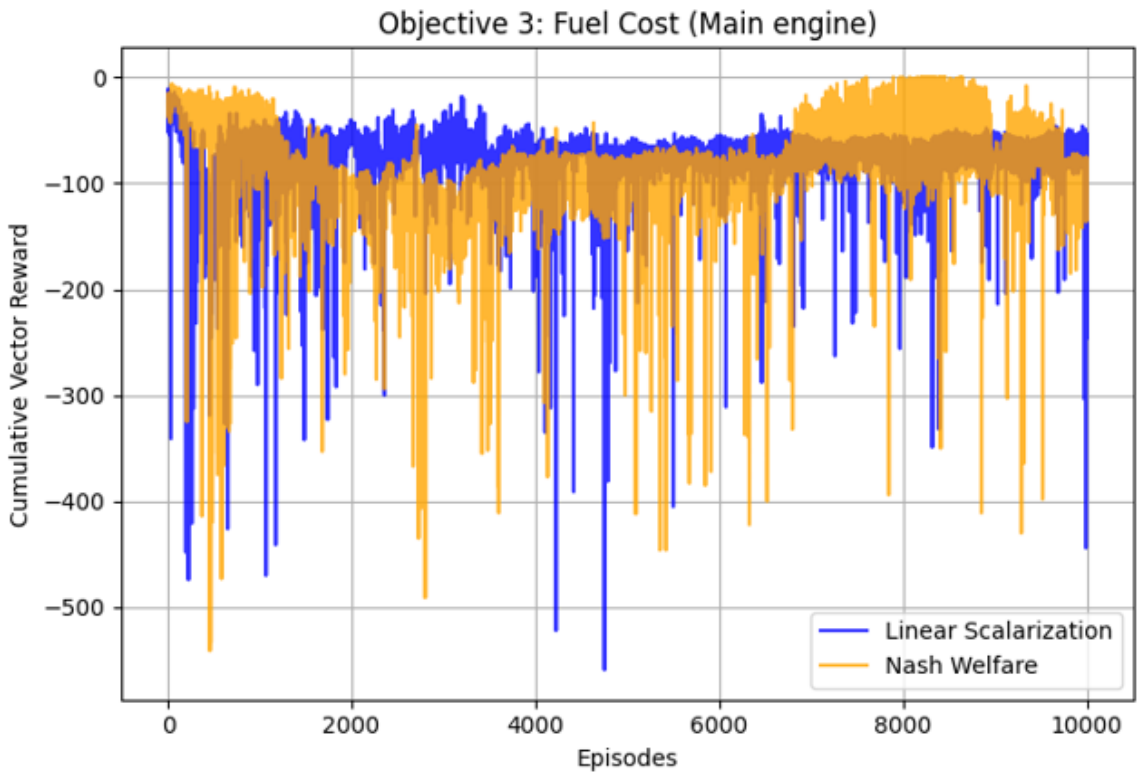
Nash Social Welfare

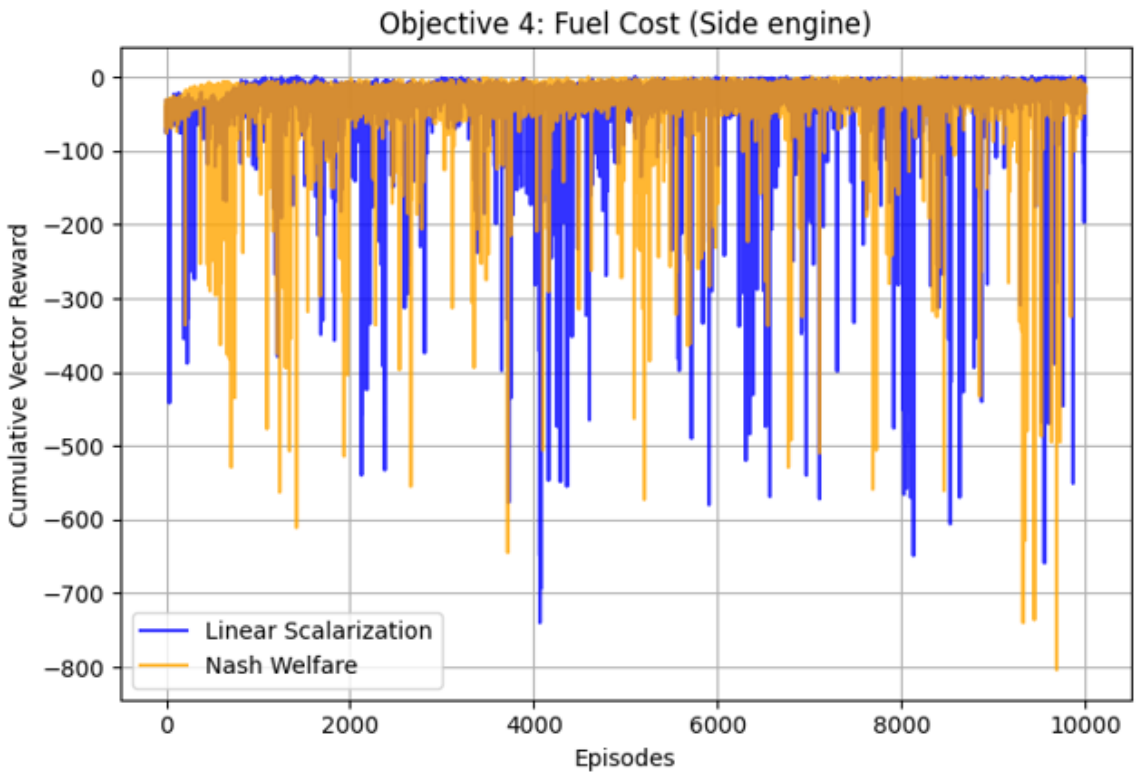


Proximal Policy Optimization(PPO) Results

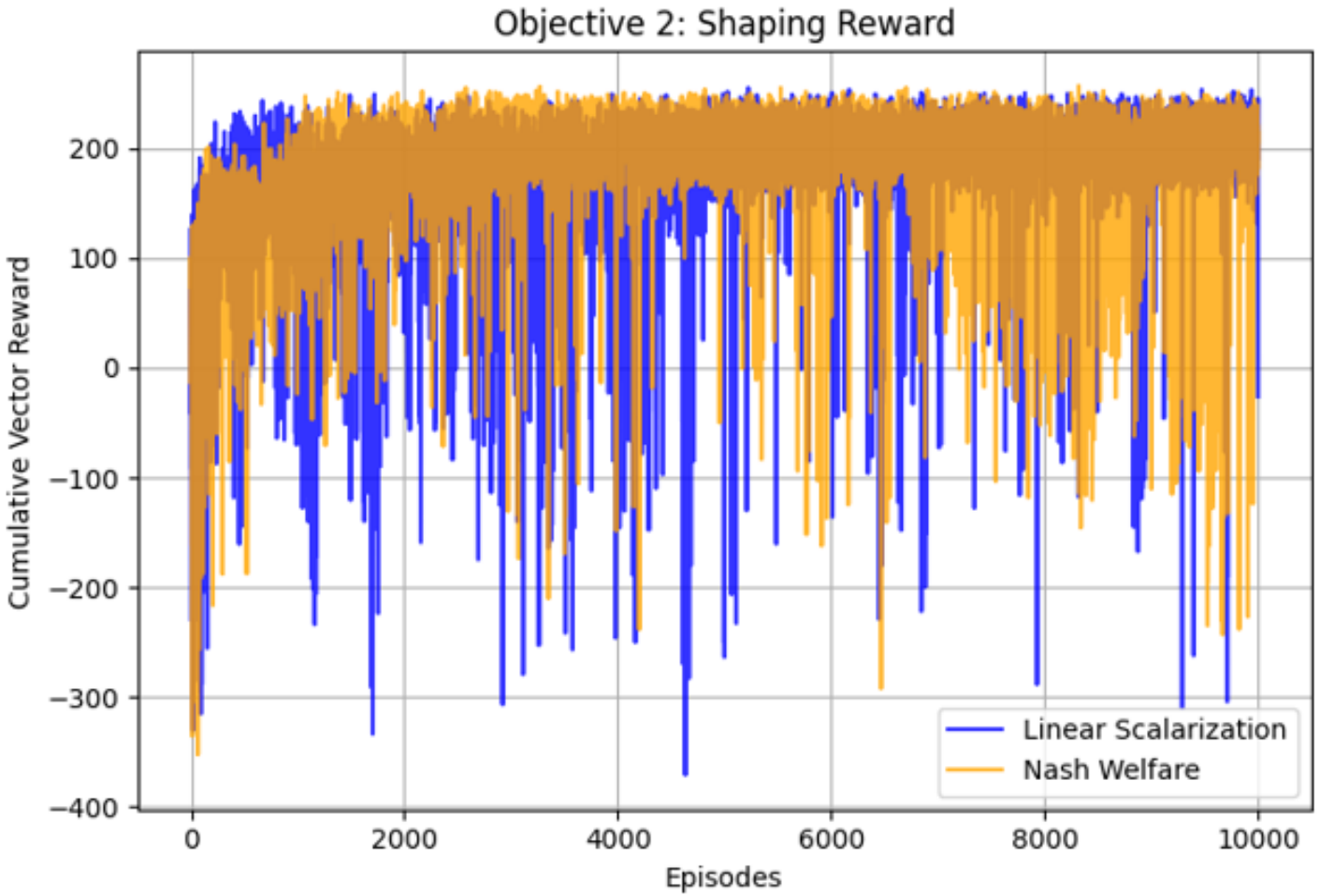


PPO Results Continued



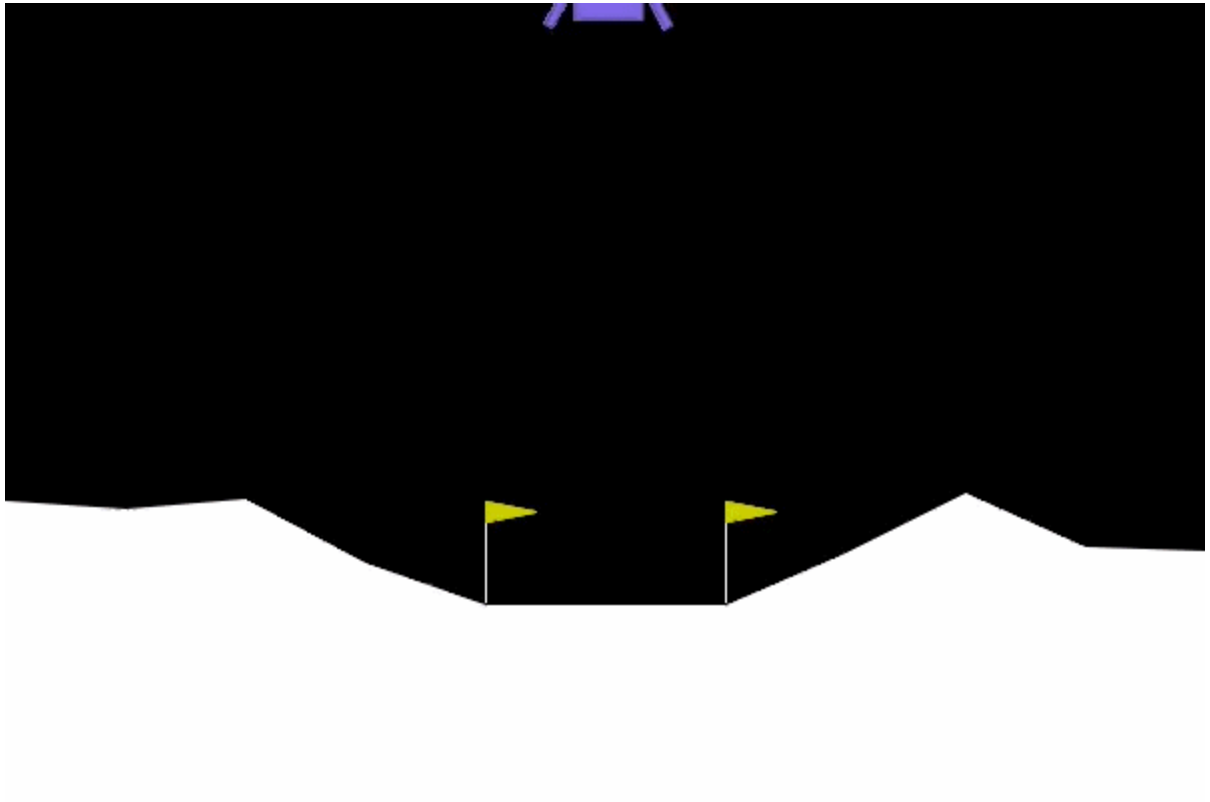


PPO Results Continued.



PPO Results Continued..

Linear



Nash Social Welfare

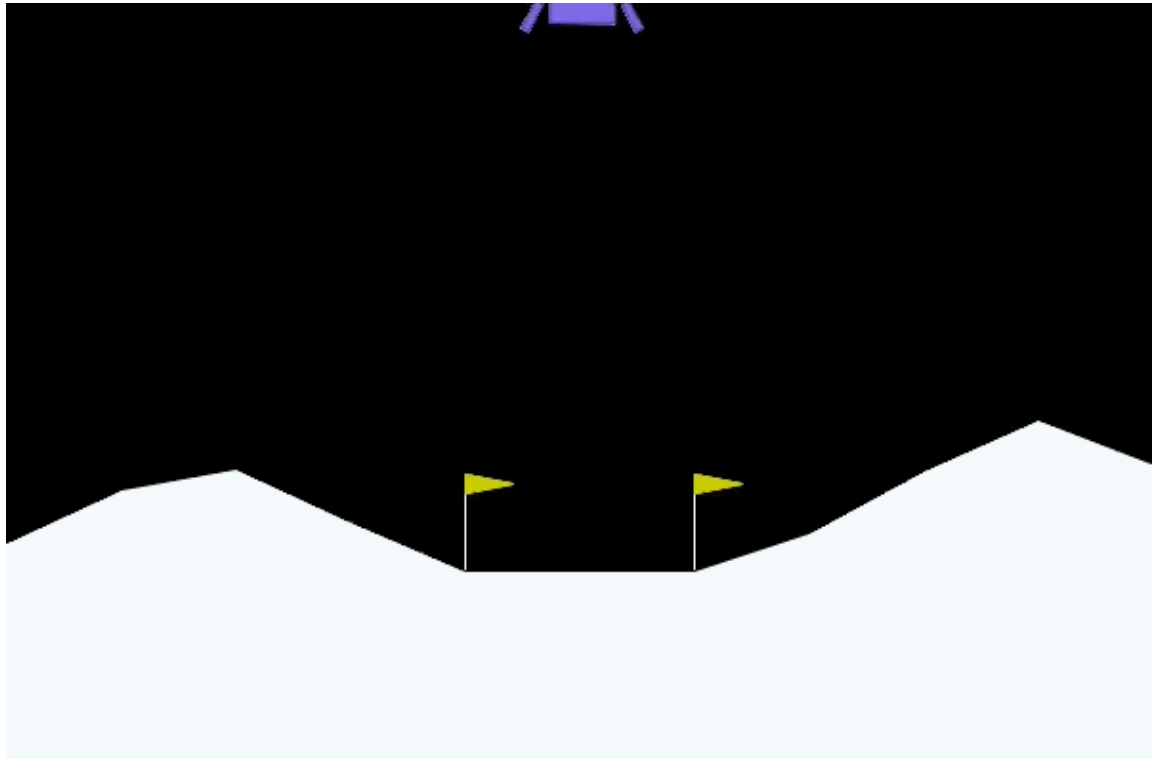


Table Summary

Table 1: Multi-Objective RL Evaluation Summary on `mo-lunar-lander-v3` ($n = 300$)

Algorithm	**Scalarization**	**Reward 1** (Terminal Status)	**Reward 2** (Shaping)	**Reward 3** (Main Fuel)	**Reward 4** (Side Fuel)
A2C	Linear	-100.000 ± 0.000	45.651 ± 64.070	-32.867 ± 36.168	-0.007 ± 0.081
PPO	Linear	75.667 ± 64.611	200.175 ± 25.085	-66.430 ± 21.808	-24.087 ± 53.435
A2C	Nash	-100.000 ± 0.000	88.239 ± 73.544	0.000 ± 0.000	-0.017 ± 0.152
PPO	Nash	81.333 ± 58.180	205.832 ± 21.569	-89.800 ± 15.926	-25.120 ± 12.690

Next Steps

- Implementing Nash Learning

Thank You!