Week 2

Advisor: Dr. Yue Wang
Student: David Tang

About Me

- University of Central Florida
- Mathematics Major, Computer Science Minor
- Rising Senior
- Ongoing Honors Thesis on Multi Objective Nash Alignment
  Advised by Professor Yue Wang

Project: Robust AI Alignment

Context: Builds off of Nash Learning from Human Feedback (NLHF).

Standard Nash Alignment suffers from a "fragile consensus" when preference data is noisy or heterogeneous.

Introduce an uncertainty kernel to bound preference perturbations and handle noisy preferences.

Project Goal: Extend this Robust Nash framework to Multi-Objective Optimization (MOO) environments and empirically benchmark it against standard RL baselines.

(payoff_u1, payoff_u2)	Player 1 (Cooperate)	Player1 (Defect)
Player 2 (Cooperate)	(-1, -1)	(-3, 0)
Player 2 (Defect)	(0, -3)	(-2, -2)

Handling Multiple Objectives

Scenario: Balancing a 2-Objective Environment (e.g., r1=Speed, r2=Safety).

- Arithmetic mean can just focus on easier objective speed
- Potentially ignore harder objective
- If 1 objective is ignored, geometric mean decreases

Let $\mathbf{r} = [r_1, r_2, \ldots, r_K]^T$ be the multi-objective reward vector

$$\mathbf{r}_{\text{unbal}} = [9, 1]^T \quad \mathbf{r}_{\text{bal}} = [5, 5]^T$$

$$S_{\text{Linear}}(\mathbf{r}_{\text{unbal}}) = \frac{1}{2}(9 + 1) = 5$$

$$S_{\text{Linear}}(\mathbf{r}_{\text{bal}}) = \frac{1}{2}(5 + 5) = 5$$

$$S_{\text{Nash}}(\mathbf{r}_{\text{unbal}}) = \sqrt{9 \times 1} = 3$$

$$S_{\text{Nash}}(\mathbf{r}_{\text{bal}}) = \sqrt{5 \times 5} = 5$$

Relation to NSW and Numerical Stability

Geometric Mean is Closely related to Nash Social Welfare (NSW)

$$\text{Nash Social Welfare (NSW)} = \prod_{k=1}^{K} r_k$$

$$\text{Geometric Mean (GM)} = \left(\prod_{k=1}^{K} r_k\right)^{1/K} = (\text{NSW})^{1/K}$$

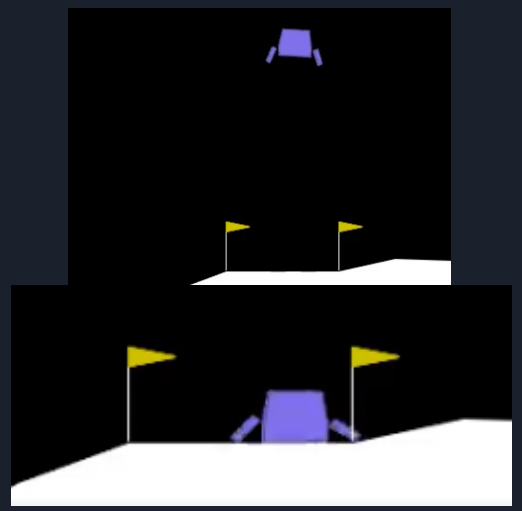
$$\arg\max \log(\text{GM}) = \arg\max \left(\frac{1}{K}\sum_{k=1}^{K} \log(r_k)\right)$$

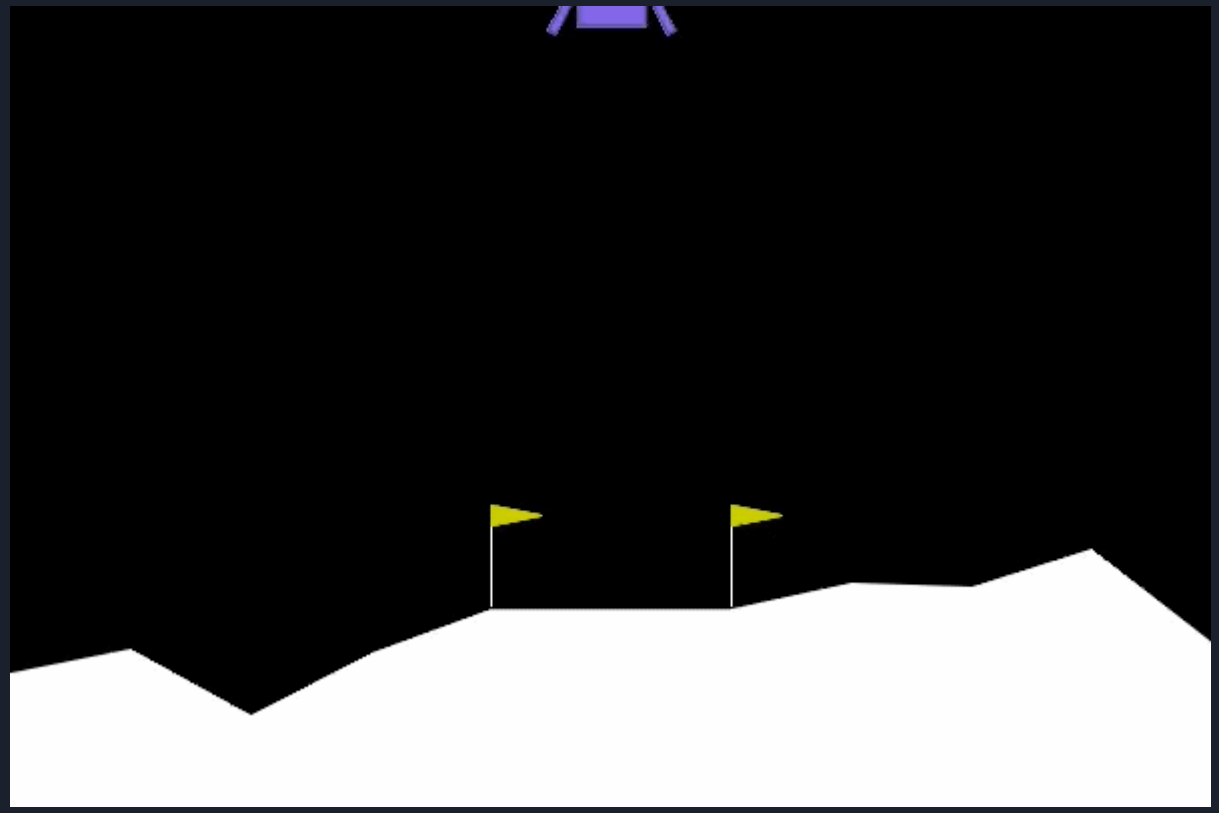
$$\arg\max \log(\text{NSW}) = \arg\max \left(\sum_{k=1}^{K} \log(r_k)\right)$$

Current Progress

Need to offset so each term is strictly positive

Other refinements as incomplete algorithm so far





Timeline

Week 1: Build Foundation for Linear Scalarization vs Nash Social Welfare and A2C baseline

Week 2: Implement PPO baseline

Week 3: Implement Robust Nash Alignment

Week 4: Implement Robust Nash Alignment

Week 5: Test on other environments RL environments

Week 6: Run sweeps across hyperparameters

Week 7: Statistical Analysis and Report Writing

Week 8: Statistical Analysis and Report Writing