Week 9

Student: David Tang

Mentor: Dr. Yue Wang

Poster Progress

Center for Research in Computer Vision
UCF
UNIVERSITY OF CENTRAL FLORIDA

Training Dynamics and Decision Stability in Multi-Objective Reinforcement Learning via Spectral Sensitivity Analysis

Author Name
University of Central Florida, Orlando, FL

Introduction

Problem and Background

- Standard Multi-Objective Reinforcement Learning assumes smooth, convex Pareto frontiers. In practice, discovered empirical frontiers are noisy, non-convex, and discrete.
- Existing selection rules force a choice between either efficiency (Nash Social Welfare) or guaranteeing the worst-case floor (Chebyshev).
- Standard evaluation techniques apply arbitrary additive noise or random point removal, which break fundamental Pareto properties such as weak monotonicity or boundary conditions.

Contributions

- **Hybrid Decision Rule:** Interpolates continuously between Nash efficiency and Chebyshev robustness.
- **Sensitivity Pipeline:** Uses a Type-II Discrete Cosine Transform to systematically measure decision stability.

Preliminaries

For vector $\mathbf{u} \in \mathbb{R}^K$:

- Linear
  $f_{\text{linear}}(\mathbf{u}) = \sum_{k=1}^{K} w_k u_k$
- Nash Social Welfare
  $f_{\text{Nash}}(\mathbf{u}) = \prod_{k=1}^{K} u_k$
- Chebyshev
  $f_{\text{Chebyshev}}(\mathbf{u}) = -\max_k |w_k(z_k^* - u_k)|$

where $\sum_{k=1}^{K} w_k = 1$

N choices $[x_n, y_n]$ sorted ascending by x:

$\mathbf{y} = [y_0, y_1, \ldots, y_{N-1}]^T \in \mathbb{R}^N$

subject to $y_n \geq y_{n+1}$

Method

We define a parametric hybrid decision rule that interpolates between Nash Social Welfare and Chebyshev minimax selection:

$$f_{\text{hybrid}}(\mathbf{u}; \alpha) = \alpha f_{\text{Nash}}(\mathbf{u}) + (1 - \alpha) f_{\text{Chebyshev}}(\mathbf{u}), \quad \alpha \in (0, 1). \tag{4}$$

4

4.1 Sensitivity and the Integrated Sensitivity Score

We consider a two objective setting. Recall from Section 3 that the original empirical Pareto frontier is given by the discrete vector $\mathbf{y}$. We let $\mathbf{y}_E$ represent the simplified frontier at a retained signal energy level of $E \in [0, 1]$, where $E = 1$ yields the original complex frontier and $E \to 0$ represents a heavily smoothed frontier.

A chosen multi-objective selection rule $\mathcal{D}$ (e.g., Nash, Chebyshev, Linear) maps a frontier to an optimal choice vector. We define the selections on the original and simplified frontiers as

$$\mathbf{v}^* = \mathcal{D}(\mathbf{x}, \mathbf{y}) = [x^*, y^*]^T \quad \text{and} \quad \mathbf{v}_E^* = \mathcal{D}(\mathbf{x}_E, \mathbf{y}_E) = [x_E^*, y_E^*]^T. \tag{5}$$

The Sensitivity Curve $S(E)$ tracks how this choice vector shifts as the energy levels vary and is given by the Euclidean distance

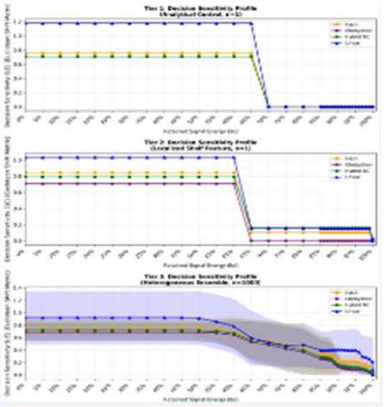
$$S(E) = \|\mathbf{v}^* - \mathbf{v}_E^*\|_2. \tag{6}$$

While it is possible to observe $S(E)$ in isolation, instead of arbitrarily choosing a retained signal energy level E, we instead compute the Integrated Sensitivity Score (A) using trapezoidal integration

$$A = \int_0^1 S(E)\, dE \approx \sum_{j=0}^{M-2} \frac{S(E_j) + S(E_{j+1})}{2} (E_{j+1} - E_j), \tag{7}$$

where M is the number of evaluation intervals and E_j represents the discrete energy threshold at index j. This integral measures the cumulative sensitivity of the selection rule. A lower integrated score ($A \to 0$) indicates an inherently stable, insensitive selection rule, whereas a higher value indicates a less stable selection rule.

Experiment



Conclusion

Future Work

Report Progress

3 Preliminaries

In this section, we describe the mathematical foundation of sequential decision-making in environments with multi-dimensional reward structures and define the notation used in our work.

3.1 Multi-Objective Markov Decision Processes

We model the environment as a Multi-Objective Markov Decision Process (MOMDP), which extends the standard Markov Decision Process framework by mapping transitions to vector-valued feedback signals. Formally, an MOMDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, \boldsymbol{R}, \gamma)$. Here, the state space $\mathcal{S}$ represents the set of all possible states in the environment, and the action space $\mathcal{A}$ is the set of actions the agent can choose from. The transition function P is written as $P(s_{t+1} \mid s_t, a_t)$ and gives the probability of moving to state s_{t+1} when taking action a_t in state s_t. The multi-objective reward function $\boldsymbol{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$ maps a state-action pair to a K-dimensional reward vector $\boldsymbol{r}_t = [r_{t,1}, r_{t,2}, \ldots, r_{t,K}]^T$. The discount factor $\gamma \in [0, 1)$ determines the importance of immediate rewards compared to future rewards.

The behavior of an agent is determined by a policy $\pi(a \mid s)$, which maps states to a probability distribution over actions. The vector value function $\boldsymbol{V}^\pi(s) \in \mathbb{R}^K$ estimates how good it is for the agent to be in a certain state and the expected cumulative future reward. Formally, we define $\boldsymbol{V}^\pi(s)$

as

$$\boldsymbol{V}^\pi(s) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t \;\middle|\; s_0 = s\right], \tag{1}$$

where $\tau = (s_0, a_0, s_1, a_1, \ldots)$ denotes the trajectory generated under policy π within the transition dynamics P. Similarly, the vector action-value function $\boldsymbol{Q}^\pi(s, a) \in \mathbb{R}^K$ denotes the expected return after taking action a in state s and is defined as

$$\boldsymbol{Q}^\pi(s, a) = \boldsymbol{R}(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s, a)}\left[\boldsymbol{V}^\pi(s')\right]. \tag{2}$$

3.2 Multi-Objective Scalarization Methods

Standard deep reinforcement learning objectives require a scalar optimization signal to perform gradient updates. Therefore, vector-valued returns must be mapped to a scalar target using a scalarization function $f : \mathbb{R}^K \to \mathbb{R}$. We consider three primary scalarization methods for a generic vector $\boldsymbol{u} = [u_1, u_2, \ldots, u_K]^T \in \mathbb{R}^K$ representing the multi-objective feedback metrics.

Linear Scalarization The most common approach is linear scalarization which we define as $f_{\text{linear}}(\boldsymbol{u}) = \sum_{k=1}^K w_k u_k$, where $\sum_{k=1}^K w_k = 1$. While straightforward, linear scalarization has a tendency to optimize some objectives while ignoring others and may require rescaling objectives to resolve this [1].

Nash Bargaining Solution (Product Aggregation) Another method to balance competing objectives without requiring manual weight tuning is the cooperative game-theoretic framework originally established by Nash [3]. Assuming a strictly non-negative coordinate space, the aggregation maps multi-dimensional feedback vectors by computing their joint product $f_{\text{Nash}}(\boldsymbol{u}) = \prod_{k=1}^K u_k$. This formulation naturally penalizes policies that heavily favor one objective axis at the expense of another, encouraging joint improvement instead. Following recent advancements in deep multi-task and multi-objective optimization setups [4], we omit the disagreement point and assume symmetric bargaining power (giving equal weight to all objectives).

Chebyshev Scalarization and Maximin The maximin approach optimizes for the weakest element in the vector and helps to guarantee a minimum floor for objective performance. Its general variant is weighted Chebyshev scalarization [5], which is defined as $f_{\text{Chebyshev}}(\boldsymbol{u}) = -\max_{k \in \{1,\ldots,K\}}\left[w_k\left(z_k^\star - u_k\right)\right]$, where the vector $\boldsymbol{z}^\star = [z_1^\star, z_2^\star, \ldots, z_K^\star]^T$ represents a utopian reference point of ideal target maximum for each objective, and w_k denotes the weight assigned to the k-th objective.

3.3 Pareto Efficiency and Empirical Frontiers

As multiple objectives frequently conflict, optimizing every objective simultaneously is often impossible. Instead, evaluation is done using the criterion of Pareto efficiency: where a performance vector $\boldsymbol{a}$ Pareto dominates $\boldsymbol{b}$ if $a_k \geq b_k$ for all k and $a_k > b_k$ for at least one objective. A policy π is Pareto optimal if no feasible policy π' strictly dominates its expected return $\boldsymbol{V}^\pi(s)$.

In practice, reinforcement learning agents approximate this optimal tradeoff space on the empirical frontier consisting of a finite set of N policies on the Pareto frontier. For a two objective setting where each policy is represented by a coordinate pair $[x, y]$, we sort the pairs in ascending order relative to the primary objective x. This maps the empirical frontier to a 1D spatial signal given by

$$\mathbf{y} = [y_0, y_1, \ldots, y_{N-1}]^T \in \mathbb{R}^N \quad \text{subject to } y_n \geq y_{n+1} \quad \forall n \in \{0, \ldots, N-2\}, \tag{3}$$

where y_n is our secondary objective trade-off value evaluated at the primary objective value x_n.

4 Method

4.0.1 Hybrid Convex-Combination Decision Rule

We define a parametric hybrid decision rule that interpolates between Nash Social Welfare and Chebyshev minimax selection:

$$f_{\text{hybrid}}(\mathbf{u}; \alpha) = \alpha f_{\text{Nash}}(\mathbf{u}) + (1 - \alpha) f_{\text{Chebyshev}}(\mathbf{u}), \quad \alpha \in (0, 1). \tag{4}$$

This formulation allows continuous control between efficiency-driven and robustness-driven decision behavior.

Interpretation - $\alpha \to 1$: efficiency-dominated (Nash-like behavior) - $\alpha \to 0$: risk-averse minimax (Chebyshev behavior)

Properties of the Hybrid Rule **Pareto Optimality:** For any $\alpha \in (0, 1)$, any maximizer of f_{hybrid} is strictly Pareto optimal under standard monotonicity assumptions of f_{Nash} and $f_{\text{Chebyshev}}$. **Controlled Risk–Efficiency Trade-off:** The parameter α provides a continuous bound on deviation from the utopian reference point, interpolating between Nash efficiency and Chebyshev robustness. As $\alpha \to 0$, the solution converges to a minimax-style security boundary.

4.1 Sensitivity and the Integrated Sensitivity Score

We consider a two objective setting. Recall from Section 3 that the original empirical Pareto frontier is given by the discrete vector $\mathbf{y}$. We let $\hat{\mathbf{y}}_E$ represent the simplified frontier at a retained signal energy level of $E \in [0, 1]$, where $E = 1$ yields the original complex frontier and $E \to 0$ represents a heavily smoothed frontier.

A chosen multi-objective selection rule $\mathcal{D}$ (e.g., Nash, Chebyshev, Linear) maps a frontier to an optimal choice vector. We define the selections on the original and simplified frontiers as

$$\mathbf{v}^\star = \mathcal{D}(\mathbf{x}, \mathbf{y}) = [x^\star, y^\star]^T \quad \text{and} \quad \mathbf{v}_E^\star = \mathcal{D}(\hat{\mathbf{x}}_E, \hat{\mathbf{y}}_E) = [x_E^\star, y_E^\star]^T. \tag{5}$$

The Sensitivity Curve $S(E)$ tracks how this choice vector shifts as the energy levels vary and is given by the Euclidean distance

$$S(E) = \|\mathbf{v}^\star - \mathbf{v}_E^\star\|_2. \tag{6}$$

While it is possible to observe $S(E)$ in isolation, instead of arbitrarily choosing a retained signal energy level E, we instead compute the Integrated Sensitivity Score (A) using trapezoidal integration

$$A = \int_0^1 S(E)\, dE \approx \sum_{j=0}^{M-1} \frac{S(E_j) + S(E_{j+1})}{2}(E_{j+1} - E_j), \tag{7}$$

where M is the number of evaluation intervals and E_j represents the discrete energy threshold at index j. This integral measures the cumulative sensitivity of the selection rule. A lower integrated score ($A \to 0$) indicates an inherently stable, insensitive selection rule, whereas a higher value indicates a less stable selection rule.

Next Steps

- Finish Poster
- Finish Report