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

David Tang
Department of Mathematics
University of Central Florida

Yue Wang
Department of Electrical & Computer Engineering
University of Central Florida

Abstract

Multi-Objective Reinforcement Learning (MORL) relies on decision selection rules to isolate optimal policies from empirical Pareto frontiers. However, conventional selection criteria remain susceptible to noise from exploration stochasticity and traditional robustness evaluations depend on heuristic, unstructured perturbations. To address these challenges, we introduce a framework for stable multi-objective decision-making and evaluation. First, we propose a parametric hybrid decision rule that interpolates between Nash Social Welfare and Chebyshev scalarization, balancing the efficiency and smoothness of Nash with the robust performance floor of Chebyshev. Second, we introduce a spectral sensitivity framework utilizing the Type-II Discrete Cosine Transform (DCT-II) and Parseval-based energy truncation. By systematically smoothing the Pareto frontier under monotonicity and boundary constraints, we define an Integrated Sensitivity Score that quantifies cumulative selection stability. Evaluations across analytical synthetic and reinforcement learning environments demonstrate that our hybrid decision rule improves empirical training stability and that our framework provides a systematic measure of decision sensitivity.

1 Introduction

Multi-Objective Reinforcement Learning (MORL) has become an essential approach for training agents in environments with conflicting objectives [6]. Unlike traditional single-objective reinforcement learning, which collapses multiple performance metrics into a static scalar reward function, MORL aims to discover a Pareto frontier of policies as it is often impossible to optimize every objective simultaneously. In practical applications, a decision selection rule such as Nash Social Welfare, Weighted Chebyshev, or linear scalarization must be applied to isolate a single optimal vector from this empirical frontier. A limitation of these frameworks is their reliance on selection rules that are sensitive to noise that result from exploration stochasticity inherent to empirical Pareto frontiers.

To move beyond these limitations, we propose a framework for stable multi-objective decision-making and evaluation. Rather than relying on static or sensitive selection criteria, we formulate a parametric hybrid decision rule that bridges Nash Social Welfare and Chebyshev scalarization to balance efficiency and robustness. Furthermore, to systematically analyze how selection rules respond to structural variations, we introduce a pipeline based on the Discrete Cosine Transform (DCT-II) and Parseval's energy truncation, paired with an Integrated Sensitivity Score to quantify cumulative selection stability.

Our research contributions can be summarized as the following:

Preprint.

- **Hybrid Convex Combination Decision Rule:** We propose a parametric hybrid decision rule that interpolates between Nash Social Welfare and Chebyshev scalarization, balancing the efficiency and smoothness of Nash Social Welfare with the robustness of Chebyshev-based decision making.
- **Spectral Sensitivity Framework:** We introduce a spectral framework for evaluating the stability of multi-objective decision rules by measuring how selected solutions change under controlled simplification of the Pareto frontier. We define the Integrated Sensitivity Score, which summarizes cumulative decision sensitivity by integrating the sensitivity curve over all retained signal energy levels.
- **Energy-Controlled DCT Simplification:** We implement the framework using an orthonormal DCT-II simplification operator with Parseval-based energy truncation and geometric post-processing to preserve weak monotonicity and boundary constraints, and evaluate stability across synthetic and MORL environments.

2 Related Work

2.1 Multi-Objective Decision Rules and Scalarization

Selecting a single solution from a Pareto frontier is a fundamental challenge in multi-objective optimization. Classical approaches such as linear scalarization combine objectives using weighted sums, providing a simple and interpretable decision mechanism. In multi-objective reinforcement learning, scalarization methods are commonly used to transform vector-valued rewards into a single optimization objective [6]. Other approaches based on Nash Social Welfare [4] promote balanced solutions across competing objectives. However, these decision rules typically assume reliable estimates of the underlying Pareto frontier and do not explicitly characterize sensitivity to empirical frontier perturbations. In practical deep reinforcement learning settings, discovered Pareto frontiers are often represented as discrete and noisy coordinate sequences rather than clean analytical geometries.

2.2 Chebyshev Scalarization and Non-Smoothness in Multi-Objective Optimization

A well-known limitation of standard Chebyshev scalarization in gradient-based optimization is its non-smoothness due to the $\max(\cdot)$ operator, which can lead to unstable gradient trajectories. While recent works have proposed resolving this by mathematically smoothing the Chebyshev function itself [3], our framework takes an alternative structural approach. Rather than altering the underlying Chebyshev operator, we instead hybridize it with a smooth, naturally well-behaved objective.

2.3 Robustness and Uncertainty in Multi-Objective Systems

Existing sensitivity and robustness analyses in multi-objective optimization typically rely on arbitrary noise injection or random point dropping directly to objective values. Furthermore, recent robustness studies in multi-objective reinforcement learning have focused on preference weight shifts or environmental stochasticity, such as dynamic adversarial games [2], rather than localized, continuous geometric distortions like function approximation noise and undulating plateaus.

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

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

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 [4]. 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 [5], 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 (derivation of this equivalence is in the Appendix 8) [7], which is defined as $f_{\text{Chebyshev}}(\boldsymbol{u}) = -\max_{k \in \{1,\ldots,K\}} \left[w_k \left(z_k^* - u_k \right) \right]$, where the vector $\boldsymbol{z}^* = [z_1^*, z_2^*, \ldots, z_K^*]^T$ 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

4.1 Hybrid Convex Combination Decision Rule

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

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

This formulation allows us control the balance between efficiency and robustness for decision making. While straightforward, our hybrid inherits the strengths of its components while compensating for its individual weaknesses: it preserves strict Pareto optimality and smoothness from Nash Social Welfare, while providing a strict performance floor and robust optimization behavior from Chebyshev scalarization.

4.2 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

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

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

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

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.

4.3 Requirements for Admissible Simplification Operators

To map the original frontier $\mathbf{y}$ to its simplified counterpart $\hat{\mathbf{y}}_E$, we define a generalized simplification operator $\mathcal{P}_E$:

$$\hat{\mathbf{y}}_E = \mathcal{P}_E(\mathbf{y}),$$

where the parameter $E \in [0,1]$ represents the exact proportion of retained signal energy used as the invariant control parameter. Throughout this work, the signal energy refers to the squared L_2 norm of the frontier representation ($\|\mathbf{y}\|_2^2$). This will serve as an objective proxy for geometric complexity; reducing this energy eliminates the high-frequency variations.

To be considered an admissible simplification operator within this framework, the following requirements should be satisfied:

1. **Weak Monotonicity Preservation:** Since Pareto frontiers are non-increasing with respect to competing objectives, the simplification operator must never introduce artificial upward slopes. The simplified front must strictly preserve weak monotonicity where

$$\Delta y_n \leq 0 \implies \Delta \hat{y}_{E,n} \leq 0 \quad \forall n \in \{0,\ldots,N-2\}.$$

2. **Boundary Value Preservation:** The extreme endpoints where one objective is entirely maximized at the expense of the other must remain fixed. This anchors the Pareto boundary constraints of the domain where

$$\hat{y}_{E,0} = y_0 \quad \text{and} \quad \hat{y}_{E,N-1} = y_{N-1}.$$

3. **Nesting Property:** In addition, the underlying simplification operator should ideally satisfy

$$\mathcal{P}_{E_1}(\mathcal{P}_{E_2}(\mathbf{y})) = \mathcal{P}_{E_1}(\mathbf{y}) \quad \text{for } E_1 < E_2.$$

While practical nonlinear corrections (such as enforcing weak monotonicity) can relax this exact spatial composition, the operator should fundamentally prioritize a structured, monotonic reduction of detail. This guarantees that lower energy levels reduce fine-scale geometric variations without artificially generating structural anomalies.

4.4 Type-II Discrete Cosine Transform (DCT) as a Simplification Operator

In order to put the general simplification properties defined in Section 4.2 into practice, we construct $\mathcal{P}_E$ using an orthonormal Type-II Discrete Cosine Transform (DCT-II) with localized geometric post-processing. The forward transform projects the empirical spatial frontier vector $\mathbf{y}$ into the frequency domain, mapping the sequence to a vector of spectral coefficients $\mathbf{c} = [c_0, c_1, \ldots, c_{N-1}]^T$ according to

$$c_k = \beta_k \sum_{n=0}^{N-1} y_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \quad k = 0, \ldots, N-1, \tag{8}$$

where β_k is the orthonormal scaling factor given by

$$\beta_k = \begin{cases} \sqrt{\frac{1}{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & k > 0. \end{cases} \tag{9}$$

Because the forward transform is strictly orthonormal, it preserves the total norm of the vector. By Parseval's identity, the total signal energy is perfectly conserved between domains where

$$\|\mathbf{y}\|_2^2 = \sum_{n=0}^{N-1} |y_n|^2 = \sum_{k=0}^{N-1} |c_k|^2 = \|\mathbf{c}\|_2^2. \tag{10}$$

This property allows us to use E as a controlled spectral complexity parameter so we can isolate a target coefficient cutoff index k matched to our desired energy threshold where

$$E(k) = \frac{\sum_{i=0}^{k-1} c_i^2}{\sum_{i=0}^{N-1} c_i^2}. \tag{11}$$

We then set all coefficients beyond this cutoff to zero, giving us the filtered vector of frequency coefficients $\mathbf{c}_{\text{filt}} = [c_0, c_1, \ldots, c_{k-1}, 0, \ldots, 0]^T$. To reconstruct the smoothed spatial frontier $\hat{\mathbf{y}}_{\text{recon}}$, we apply the Type-III Discrete Cosine Transform (the Inverse DCT):

$$\hat{y}_{\text{recon},n} = \sum_{i=0}^{k-1} \beta_i c_i \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)i\right], \quad n = 0, \ldots, N-1. \tag{12}$$

While the core spectral truncation stage acts as a strict orthogonal projection in L^2 space, raw spectral truncation can introduce localized Gibbs oscillations or minor boundary translations that violate the monotonicity and boundary conditions. To resolve this, our implementation combines spectral filtering with sequential geometric corrections. The raw sequence is transformed into a monotone non-increasing array and anchored at its boundaries:

$$\begin{aligned} \tilde{y}_{N-1} &= y_{N-1}, \\ \tilde{y}_n &= \max\left(\hat{y}_{\text{recon},n}, \tilde{y}_{n+1}\right) \quad \forall n \in \{N-2, \ldots, 0\}, \\ \hat{y}_{E,n} &= y_{N-1} + (y_0 - y_{N-1}) \cdot \left(\frac{\tilde{y}_n - \tilde{y}_{N-1}}{\tilde{y}_0 - \tilde{y}_{N-1}}\right). \end{aligned} \tag{13}$$

The resulting sequence $\hat{\mathbf{y}}_E$ thus is a valid Pareto frontier. As the discovered empirical Pareto frontier can be sparse or non-uniformly distributed, we first sort the coordinate pairs in ascending order of the primary objective $\mathbf{x}$. To satisfy the uniform spaced assumption required by the DCT-II operator, we define a uniformly spaced grid over the domain $[x_0, x_{N-1}]$ and interpolate the secondary objectives $\mathbf{y}$ onto this grid using linear interpolation.

Algorithm 1 Sensitivity Evaluation Pipeline

Input: Raw empirical coordinates $\mathbf{x}_{\text{raw}} \in \mathbb{R}^P$, raw empirical objectives $\mathbf{y}_{\text{raw}} \in \mathbb{R}^P$, decision operator $\mathcal{D}(\cdot, \cdot)$, target energies $\mathbf{E} = [E_1, \ldots, E_m]$
Initialize orthonormal scale factor: $\beta_k = \begin{cases} \sqrt{1/N}, & k = 0 \\ \sqrt{2/N}, & k > 0 \end{cases}$
Sort $(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}})$ and interpolate onto a uniform grid of size N to yield $\mathbf{x}_{\text{grid}}, \mathbf{y}_{\text{grid}} \in \mathbb{R}^N$
Redefine normalized coordinates in $[0, 1]$: $\mathbf{x} \leftarrow \text{Normalize}(\mathbf{x}_{\text{grid}})$, $\mathbf{y} \leftarrow \text{Normalize}(\mathbf{y}_{\text{grid}})$
$\mathbf{v}^* \leftarrow \mathcal{D}(\mathbf{x}, \mathbf{y})$ ◁ *Reference selection*
$\mathbf{c} \leftarrow \text{DCT}(\mathbf{y})$ where $c_k = \beta_k \sum_{n=0}^{N-1} y_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right) k\right]$
for $j = 1, 2, \ldots, m$ **do**
  Find minimal $k \in \mathbb{N}$ such that: $\frac{\sum_{i=0}^{k-1} c_i^2}{\sum_{i=0}^{N-1} c_i^2} \geq E_j$ ◁ *Energy cutoff*
  $\mathbf{c}_{\text{filt}} \leftarrow [c_0, c_1, \ldots, c_{k-1}, 0, \ldots, 0]^T$
  $\hat{\mathbf{y}}_{\text{recon}} \leftarrow \text{IDCT}(\mathbf{c}_{\text{filt}})$ where $\hat{y}_{\text{recon},n} = \sum_{i=0}^{k-1} \beta_i c_i \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right) i\right]$
  $\tilde{y}_{N-1} \leftarrow y_{N-1}$
  $\tilde{y}_n \leftarrow \max\left(\hat{y}_{\text{recon},n}, \tilde{y}_{n+1}\right) \quad \forall n \in \{N-2, \ldots, 0\}$ ◁ *Enforce monotonicity*
  if $\tilde{y}_0 - \tilde{y}_{N-1} \neq 0$ **then**
    $\hat{y}_{E_j,n} \leftarrow y_{N-1} + (y_0 - y_{N-1}) \cdot \frac{\tilde{y}_n - \tilde{y}_{N-1}}{\tilde{y}_0 - \tilde{y}_{N-1}} \quad \forall n \in \{0, \ldots, N-1\}$ ◁ *Boundary anchoring*
  else
    $\hat{\mathbf{y}}_{E_j} \leftarrow [y_{N-1}, y_{N-1}, \ldots, y_{N-1}]^T$ ◁ *Fallback baseline*
  end if
  $\mathbf{v}^*_{E_j} \leftarrow \mathcal{D}(\mathbf{x}, \hat{\mathbf{y}}_{E_j})$ ◁ *Perturbed selection*
  $S(E_j) \leftarrow \left\| \mathbf{v}^* - \mathbf{v}^*_{E_j} \right\|_2$
end for
$A \leftarrow \sum_{j=1}^{m-1} \frac{S(E_j) + S(E_{j+1})}{2} (E_{j+1} - E_j)$ ◁ *Trapezoidal integration*
return $\{S(E_j)\}_{j=1}^m, A$

5 Experiment

To validate our framework, we test it in both synthetic analytical environments and MORL environments.

5.1 Synthetic Examples and Analytical Benchmarking

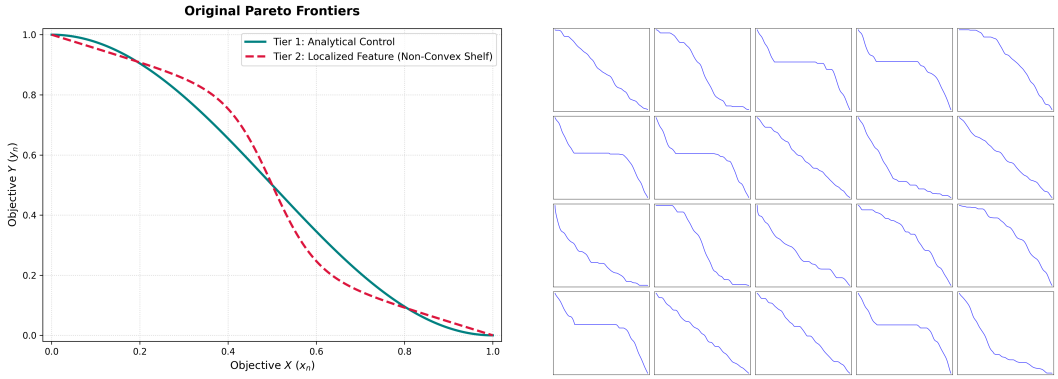
We evaluate the performance using three synthetic setups. Each of these is illustrated in Figure 1.

1. Analytical Smooth Control Baseline. This serves as a control group to establish a performance baseline using a perfectly smooth, convex empirical Pareto front grid defined over $x_n \in [0, 1]$ via a low-frequency cosine profile: $y_n = \frac{1}{2}\left(1 + \cos(\pi x_n)\right)$.

2. Non-Convex Stress Test (The Shelf Feature). To isolate how selection criteria respond to localized non-convexities, this test embeds a high-frequency localized structural anomaly ("the shelf") into the discrete coordinate generation: $y_n = 1.0 - 0.5x_n - 0.3\tanh\left(10(x_n - 0.5)\right)$, normalized to bound the unit domain: $y_n = \frac{y_n - y_{N-1}}{y_0 - y_{N-1}}$.

3. Stochastic Ensemble Evaluation over Diversified Geometries. To establish statistical significance beyond hand-crafted cases, we perform a Monte Carlo simulation over an ensemble of heterogeneous empirical Pareto fronts ($n = 1000$). The generator randomly seeds diverse struc-

tures with certain global properties (*linear, convex, concave, sigmoidal*) alongside highly localized anomalies (*sharp knees, undulating plateaus*). Smooth stochastic perturbations are introduced via a localized Gaussian filter ($\sigma = 20$) and we ensure weak monotonicity.



(a) Tier 1 and Tier 2 baseline curves.

(b) Tier 3 stochastic structural variations.

Figure 1: Synthetic Pareto Frontiers

The primary metric results across all testing tiers are presented in Figure 2, 3, and 4. The corresponding aggregated performance metric—the Integrated Sensitivity Score (A) calculated via Equation 7—is summarized in Table 1.

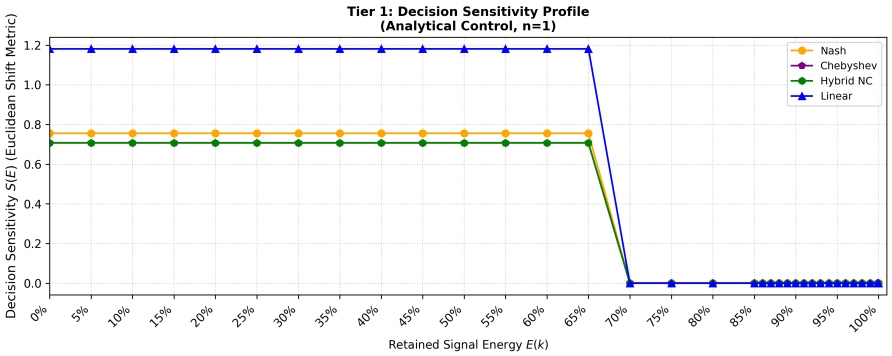


Figure 2: Sensitivity curves $S(E)$ evaluated across varying energy thresholds for Tier 1: Analytical Control Baseline.

The results we observe from Figure 2 are the hybrid convex combination and Chebyshev perform indistinguishably from each other here. This is not surprising as the our Pareto Frontier for Tier 1 is considered well-behaved. As for Figure 3, the slight shelf present in the Pareto frontier reveals that Chebyshev remains the most stable. However, at the around 65% retained signal energy, the hybrid and linear perform around the same.

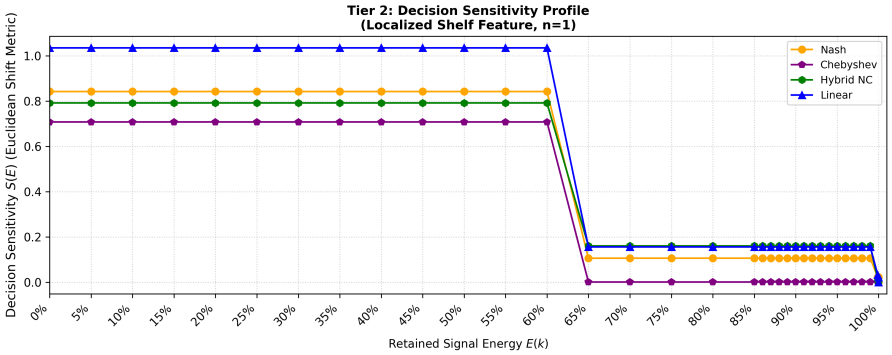


Figure 3: Sensitivity curves $S(E)$ evaluated across varying energy thresholds for Tier 2: Non-Convex Stress Test (Shelf Feature).

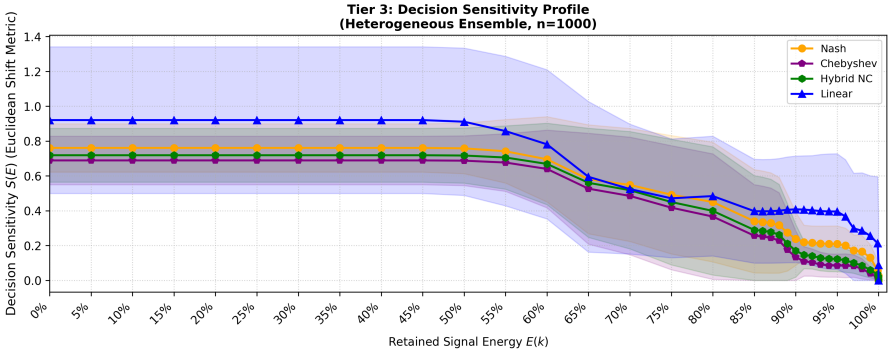


Figure 4: Sensitivity curves $S(E)$ evaluated across varying energy thresholds for Tier 3: Stochastic Ensemble Evaluation.

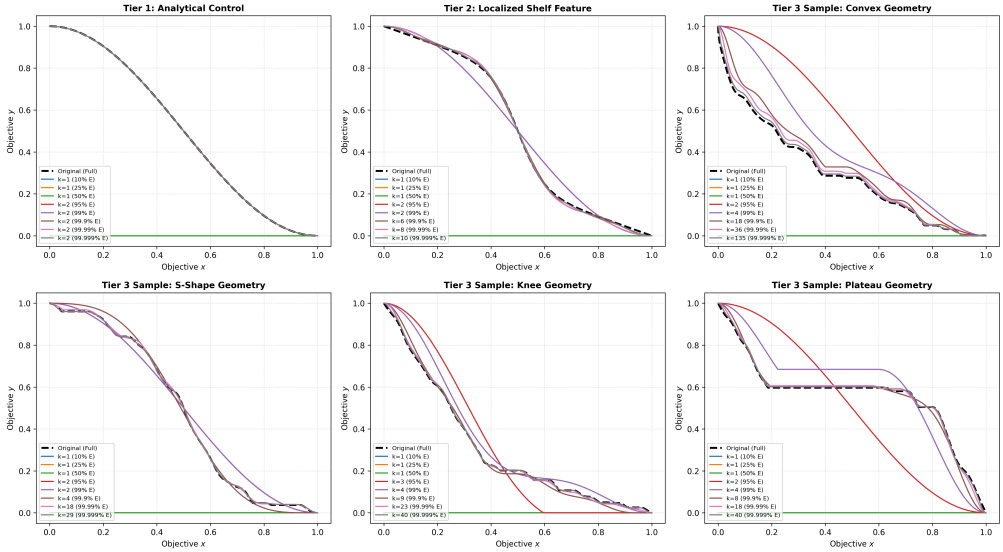


Figure 5: Visual progression of empirical Pareto frontiers across Tiers 1, 2, and 3 under varying spectral energy thresholds (E). Notice how high-frequency anomalies (like the shelf feature in Tier 2) are progressively smoothed while hard boundary anchors remain preserved.

Table 1: Integrated Sensitivity Score Summary

Decision Metric	**Tier 1** ($n = 1$) Mean ($\pm$ Std)	**Tier 2** ($n = 1$) Mean ($\pm$ Std)	**Tier 3 Ensemble** ($n = 1000$) Mean ($\pm$ Std)
MaxMin	0.4774 ($\pm$0.0000)	0.4427 ($\pm$0.0000)	0.5430 ($\pm$0.1549)
Chebyshev	0.4774 ($\pm$0.0000)	0.4427 ($\pm$0.0000)	0.5430 ($\pm$0.1549)
Hybrid NMM	0.4774 ($\pm$0.0000)	0.4675 ($\pm$0.0000)	0.5598 ($\pm$0.1716)
Hybrid NC	0.4774 ($\pm$0.0000)	0.5543 ($\pm$0.0000)	0.5736 ($\pm$0.1696)
Nash	0.5107 ($\pm$0.0000)	0.5657 ($\pm$0.0000)	0.6175 ($\pm$0.1535)
Linear	0.7972 ($\pm$0.0000)	0.7047 ($\pm$0.0000)	0.7336 ($\pm$0.1867)

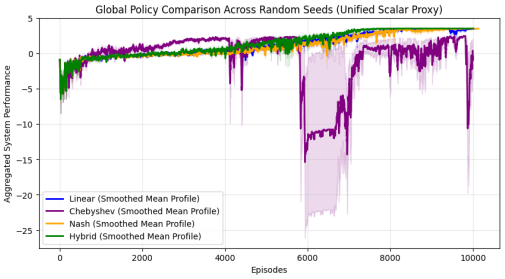
We can see that the Chebyshev and Maximin variants are the most stable. The hybrid variants benefit stability wise by having the component of Chebyshev or Maximin even when the other component is Nash. Finally, we can see overall linear scalarization seems to be the least stable compared to the other scalarization methods. An additional thing to note is that the stability of all of the decision criteria seem to be the least stable in Tier 3. This is consistent with the increased complexity of the Pareto frontiers in Tier 3. On the other hand, some of the decision criteria were more stable for the Tier 2 Pareto Frontier with the shelf. However, if we look closely at the Tier 1 curves and Tier 2 curves side by side and start at 65%, the decision sensitivity for all variants in Tier 2 are greater than that of all variants in Tier 1. This illustrates how the choice of integration bounds can change the stability metric. We will see a similar phenomenon in the multi-objective reinforcement learning example as well.

5.2 Multi-Objective Reinforcement Learning Example

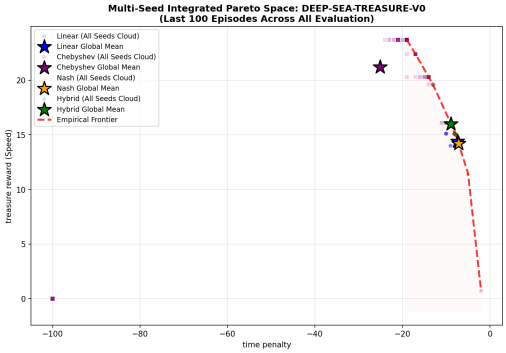
We also test our framework in a multi-objective reinforcement learning environment. Training here was done using Proximal Policy Optimization (PPO) with Generalized Advantage Estimate and vector-valued critics. Scalarization was done right before the policy gradient update for the actor whereas the critics used mean squared error for each objective added together. Here we focus on deep-sea-treasure-v0, which is a grid-based environment with two objectives: treasure and speed. The illustrations of mo-lunar-lander-v3 can be found in Appendix 9.

As we analyze the multi-objective stability for each method in Figure 6a, Chebyshev in purple stands out as collapsing seemingly during the midst of training. This can be likely attributed to the non-differentiable nature of the max operator. On the other hand, our hybrid method increases smoothly rather than having sudden sharp unstable performance.

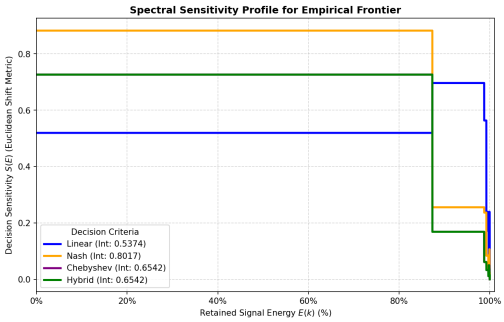
Our discovered empirical Pareto Frontier shown in Figure 6b by the agent is convex and relatively simple. An interesting thing to wonder is considering why the mean policy of Chebyshev with the purple star is not on the discovered Pareto frontier. This is simply because if one looks closely at the bottom left of the graphic with the Pareto Frontier, one of the policies from Chebyshev essentially hit the time limit. It may appear at first glance that linear is the most stable here in Figure 6c as it has the least area. However, at around 90%, the sensitivity of the linear increases whereas the others drop. Suppose we were to only evaluate the Integrated Sensitivity Score (A) starting from 80%. Suddenly, linear becomes the least stable. This implies considering where to begin calculating the area under the curve can make a difference in which results in the most or least stable.



(a) Decision trajectory stability during training.



(b) Discovered Pareto frontier



(c) Spectral sensitivity (Treasure vs Time Penalty).

Figure 6: Deep Sea Treasure (DST) evaluation.

6 Conclusion

To conclude, we propose a hybrid convex combination decision rule that inherits the strengths of Nash Social Welfare and Chebyshev scalarization while compensating for its weaknesses. We also propose a framework for evaluating how stable a multi-objective decision rule using spectral sensitivity analysis.

7 Future Work

It is possible some sort of spectral regularization or augmentation can be added to inherently encourage navigating towards a stable Pareto frontier or ensure the decision rule is more stable. However, this is only one possible approach that may help with stability and there are likely others approaches as well.

References

[1] Ziming Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 1991–1999, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[2] Mingxi Hu and Meiling Yu. Time-consistent robust multi-objective reinforcement learning via a bellman–isaacs weight-adversary recursion. In *Forty-third International Conference on Machine Learning*, 2026.

[3] Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In Ruslan Salakhutdinov, Zico Kolter,

Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30479–30509. PMLR, 21–27 Jul 2024.

[4] John F. Nash. The bargaining problem. *Econometrica: Journal of the Econometric Society*, 18(2):155–162, 1950.

[5] Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16428–16446. PMLR, 17–23 Jul 2022.

[6] Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Mach. Learn.*, 84(1–2):51–80, July 2011.

[7] Kristof Van Moffaert, Madalina M. Drugan, and Ann Nowé. Scalarized multi-objective reinforcement learning: Novel design techniques. In *2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 191–199, 2013.

Appendix

8 Chebyshev as a Generalized Maximin

To see how weighted Chebyshev scalarization reduces to the classic maximin criterion, consider the minimization of the distance to the reference point (or equivalently, maximizing its negative). Let the weights be uniform, $w_k = 1$ for all $k \in \{1, \ldots, K\}$, and assume the utopian reference point approaches infinity, $z_k^* \to \infty$.

The scalarized objective can be rewritten by pushing the negative sign inside giving us

$$f_{\text{Chebyshev}}(\boldsymbol{x}) = -\max_{k \in \{1,\ldots,K\}} \left[1 \cdot (z_k^* - x_k)\right] \tag{14}$$

$$= \min_{k \in \{1,\ldots,K\}} \left[-(z_k^* - x_k)\right] \tag{15}$$

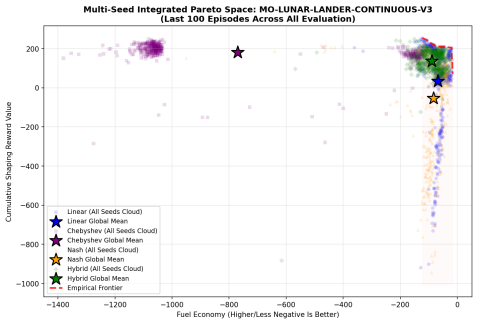
$$= \min_{k \in \{1,\ldots,K\}} \left[x_k - z^*\right]. \tag{16}$$

Since $z_k^* = z^*$ is assumed to be a uniform constant across all objectives k, subtracting it shifts the entire performance spectrum uniformly but does not alter the location of the optimum. Maximizing this function yields:

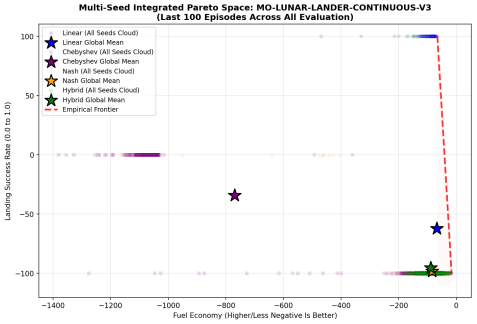
$$\arg\max_{\boldsymbol{x}} f_{\text{Chebyshev}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} \left[\min_{k \in \{1,\ldots,K\}} x_k\right], \tag{17}$$

which is exactly the standard maximin decision criterion.

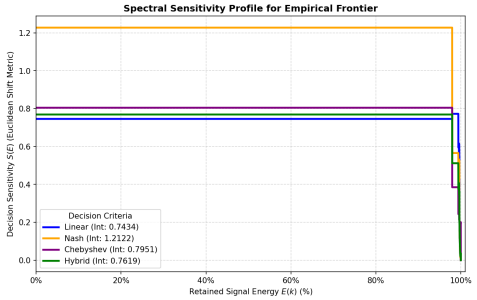
9 Lunar Lander



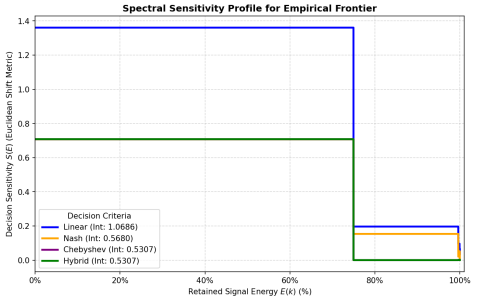
(a) Shaping vs. Fuel trade-off frontier.



(b) Success Rate vs. Fuel trade-off frontier.



(c) Spectral sensitivity (Shaping vs. Fuel).



(d) Spectral sensitivity (Success vs. Fuel).

Figure 7: Multi-Objective Lunar Lander performance metrics and corresponding spectral sensitivity analysis across continuous trajectory-dependent control dynamics.