Navigating the Social Welfare Frontier: Portfolios for Multi-objective Reinforcement Learning

Cheol Woo Kim[*] Harvard University

Jai Moondra[*] Georgia Institute of Technology

Shresth Verma Harvard University

Madeleine Pollack Massachusetts Institute of Technology

Lingkai Kong Harvard University

Milind Tambe Harvard University

Swati Gupta Massachusetts Institute of Technology

Abstract

In many real-world applications of reinforcement learning (RL), deployed policies have varied impacts on different stakeholders, creating challenges in reaching consensus on how to effectively aggregate their preferences. Generalized p-means form a widely used class of social welfare functions for this purpose, with broad applications in fair resource allocation, AI alignment, and decision-making. This class includes well-known welfare functions such as Egalitarian, Nash, and Utilitarian welfare. However, selecting the appropriate social welfare function is challenging for decision-makers, as the structure and outcomes of optimal policies can be highly sensitive to the choice of p. To address this challenge, we study the concept of an α-approximate portfolio in RL, a set of policies that are approximately optimal across the family of generalized p-means for all $p \leq 1$. We propose algorithms to compute such portfolios and provide theoretical guarantees on the trade-offs among approximation factor, portfolio size, and computational efficiency. Experimental results on synthetic and real-world datasets demonstrate the effectiveness of our approach in summarizing the policy space induced by varying p values, empowering decision-makers to navigate this landscape more effectively.

Keywords: Social Welfare, Multi-objective Reinforcement Learning, Portfolio

[*]Equal contribution

1 Introduction

In this paper, we study a reinforcement learning (RL) setting where a deployed policy impacts multiple stakeholders in different ways. Each stakeholder is associated with a unique reward function, and the goal is to train a policy that adequately aggregates their preferences. This setting, which is often modeled using multi-objective reinforcement learning (MORL), arises in many RL applications, such as fair resource allocation in healthcare [1], cloud computing [2, 3] and communication networks [4, 5]. Recently, with the rise of large language models (LLMs), reinforcement learning from human feedback (RLHF) techniques that reflect the preferences of heterogeneous individuals have also been explored [6, 7, 8].

Preference aggregation in such scenarios is often achieved by choosing a social welfare function, which takes the utilities of multiple stakeholders as input and outputs a scalar value representing the overall welfare [9, 7, 8, 6]. However, selecting the appropriate social welfare function is a nontrivial task. Different functions encode distinct fairness criteria, and the resulting policies can lead to vastly different outcomes for stakeholders depending on the choice of the social welfare function.

In this work, we focus on a class of social welfare functions known as generalized p-means, a widely used class of social welfare functions in algorithmic fairness and social choice theory. Each choice of p represents a distinct notion of fairness, and the p-means unify commonly used welfare functions such as Egalitarian welfare ($p = -\infty$), Nash welfare ($p = 0$) and Utilitarian welfare ($p = 1$), providing a smooth transition between these principles. Notably, this is known to be the only class of social welfare functions that satisfy several key axioms of social welfare, such as monotonicity, symmetry, and independence of scale [10, 11, 12, 13].

In practice, the right choice of p is often unclear in advance, and the decision-maker must understand how policies vary with p in order to make informed choices about which p (and thus which policy) to adopt. Small changes in p can sometimes lead to dramatically different policies, and selecting a policy optimized for an arbitrary p can lead to poor outcomes under a different p value. Despite these challenges, much of the existing work assumes a fixed social welfare function — and hence a fixed value of p — is given [14].

To address this challenge, we propose a method to compute a small yet representative set of policies that covers the entire spectrum of fairness criteria represented by $p \leq 1$. Our main algorithm, p-MEANPORTFOLIO, sequentially selects finite p values starting from $-\infty$ to 1. These values are chosen so that the optimal policies at these points sufficiently cover the entire range of $p \leq 1$ for a given approximation factor α. We also propose a computationally efficient heuristic algorithm, which adaptively selects the next p value from the intervals formed by previously chosen p values. The portfolios provide a structured summary of approximately optimal policies, allowing decision-makers to navigate the implications of different p values efficiently and make an informed choice about which policy to deploy.

2 Problem Setup

We consider multi-objective MDPs with finite horizon H, where N stakeholders have distinct reward functions $R_i, i \in [N]$. The key challenge in this setting is training a policy that effectively aggregates heterogeneous preferences across all stakeholders.

The generalized p-means provides a principled method to train a singly policy that aggregate heterogeneous preferences by assigning a scalar objective value to any policy π. The generalized p-mean welfare function $f(\cdot, p) : \mathbb{R}^N_{>0} \to \mathbb{R}_{>0}$ for $p \in [-\infty, 1]$ is defined as: $f(\mathbf{x}, p) = \min_{i\in[N]} x_i$ if $p = -\infty$; $\left(\frac{1}{N}\sum_{i\in[N]} x_i^p\right)^{1/p}$ if $p \notin \{-\infty, 0\}$; $\left(\prod_{i\in[N]} x_i\right)^{1/N}$ if $p = 0$.

For trajectory τ with total rewards $\mathbf{G}(\tau) = [G_1(\tau), \ldots, G_N(\tau)]$ where $G_i(\tau) = \sum_{h=1}^{H} R_i(s_h, a_h)$, we consider two aggregation rules from prior work: $v^{(1)}(\pi, p) = \mathbb{E}_{\tau\sim\pi}[f(\mathbf{G}(\tau), p)]$ and $v^{(2)}(\pi, p) = f(\mathbb{E}_{\tau\sim\pi}[\mathbf{G}(\tau)], p)$. For a fixed p, aggregation function $v^{(\ell)}(\cdot, p), \ell \in \{1, 2\}$, and a given set of policies Π, the welfare maximizing policy can be obtained by solving the problem: $\max_{\pi\in\Pi} v^{(\ell)}(\pi, p)$. The algorithms we develop assume that this problem can be solved, and we refer to each instance of solving this problem as an "oracle".

A fundamental challenge emerges when seeking to find a welfare maximizing policy: we must fix a specific value of p. However, the appropriate choice of p is often unclear to decision-makers, as it requires balancing competing notions of fairness without clear guidance on which perspective should dominate in a given context. To address this challenge, our approach leverages the concept of an $\alpha-$approximate portfolio, which provides robust performance guarantees across the entire spectrum of fairness criteria.

Definition 2.1 (α-approximate Portfolio)**.** Given aggregation rule $\ell \in \{1, 2\}$ and approximation factor $\alpha \in (0, 1)$, a set of policies $\Pi' \subseteq \Pi$ is called an α-approximate portfolio for aggregation functions $v^{(\ell)}$ if for each $p \leq 1$, there exists a policy π in Π' that achieves $v^{(\ell)}(\pi, p) \geq \alpha \times \max_{\pi'\in\Pi} v^{(\ell)}(\pi', p)$.

Table 1: Comparison of actual approximation factors for p-MEANPORTFOLIO and natural baselines.

Size	p-MEANPORTFOLIO	Random Policy	Random p
1	**0.66**	0.24	**0.66**
2	**0.61**	0.31	**0.61**
3	**0.97**	0.54	0.65
8	**0.87**	0.71	0.67
10	**0.99**	0.60	0.65

Table 2: Comparison of oracle calls and actual approximation factors.

Size	p-MEANPORTFOLIO		BUDGET-CONSTRAINED	
	Oracle	Approx.	Oracle	Approx.
1	17	**0.66**	**1**	**0.66**
2	30	**0.61**	**2**	**0.61**
3	44	**0.97**	**3**	0.92
8	118	0.87	**8**	**0.90**
10	144	**0.99**	**10**	0.92

3 Proposed Method and Results

Our approach is to compute an α-approximate portfolio that is both small and diverse, which provides effective coverage across all $p \leq 1$. Such a portfolio allows decision-makers to efficiently compare policy alternatives over the full spectrum of fairness criteria, without needing to commit to a specific value of p in advance. This eliminates the reliance on ad hoc parameter choices or costly trial-and-error exploration.

3.1 Algorithm p-MEANPORTFOLIO

Our main algorithm p-MEANPORTFOLIO is guaranteed to construct an α-approximate portfolio Π' for all $v^{(\ell)}(\cdot, p),\ p \leq 1$. The algorithm iteratively chooses an increasing sequence of p values $p_0 < p_1 < \ldots < p_K = 1$ using a line search subroutine, and adds the corresponding optimal policy $\pi_{p_k} = \arg\max_{\pi \in \Pi} v^{(\ell)}(\pi, p_k)$ to the portfolio. It ensures that π_{p_k} is α-approximate for all $p \in [p_k, p_{k+1}]$, i.e., $v^{(\ell)}(\pi_{p_k}, p) \geq \alpha \times \max_{\pi' \in \Pi} v^{(\ell)}(\pi', p)$ for all $p \in [p_k, p_{k+1}]$. In other words, each policy π_{p_k} in the portfolio is approximately optimal for any p in the interval $[p_k, p_{k+1}]$ to its right.

Theoretical guarantees. We show that the portfolio size is bounded by $|\Pi'| = O\Big(\frac{\ln \kappa}{\ln(1/\alpha)}\Big)$, and the number of oracle calls is bounded by $\widetilde{O}\Big(\frac{(\ln \kappa)^2 \ln \ln N}{\ln(1/\alpha)}\Big)$, where κ is a problem-dependent parameter representing the ratio of the maximum to minimum achievable rewards.

Key insight. The optimal value function $\max_{\pi \in \Pi} v^{(\ell)}(\pi, p)$ exhibits favorable monotonicity in p, which allows the optimal policy at a given p to remain approximately optimal over an interval of larger p values to its right.

3.2 Heuristic under a Budget Constraint

In p-MEANPORTFOLIO, we provide guarantees on the approximation factor α. However, achieving these guarantees may require a large number of oracle calls across different values of p, which can be impractical when computational resources are severely limited.

To address this limitation, we propose a heuristic BUDGET-CONSTRAINEDPORTFOLIO, designed for scenarios with a strict budget K on oracle calls. Unlike p-MEANPORTFOLIO, which selects p values in a monotonically increasing order from $-\infty$ to 1, this algorithm dynamically refines the search space based on observed approximation performance. It greedily targets regions where the approximation factor is likely to be weakest, ensuring efficient use of the limited oracle budget.

3.3 Experimental Results

We evaluate our approach by measuring the actual approximation factors achieved by the portfolios generated by the proposed methods across varying portfolio sizes. We conduct experiments on three domains, spanning both synthetic and real-world settings. Tables 2 and 1 present representative results from one domain studied previously in [15]. Table 1 shows that a small portfolio can achieve near-optimal approximation quality across all $p \leq 1$. For example, a portfolio of size 3 constructed via p-MEANPORTFOLIO achieves near-optimal approximation quality (97% optimal) across all $p \leq 1$, significantly outperforming baseline methods with the same portfolio size. This shows that our algorithm allows decision-makers to select from a compact set of policies without needing to worry about other potential choices of p. Table 2 demonstrates that BUDGET-CONSTRAINEDPORTFOLIO achieves similar approximation quality to p-MEANPORTFOLIO—sometimes even outperforming it—while requiring only a fraction of the oracle calls.

Overall, our work introduces the first algorithms for computing portfolios that summarize the policy space across fairness criteria represented by $p-$means, offering both theoretical guarantees and practical efficiency.

References

[1] Shresth Verma, Niclas Boehmer, Lingkai Kong, and Milind Tambe. Balancing act: Prioritization strategies for LLM-designed restless bandit rewards, 2024.

[2] Julien Perez, Cécile Germain-Renaud, Balázs Kégl, and Charles Loomis. Responsive elastic computing. In *Proceedings of the 6th International Conference Industry Session on Grids Meets Autonomic Computing*, GMAC '09, page 55–64. Association for Computing Machinery, 2009.

[3] Hao Hao, Changqiao Xu, Wei Zhang, Shujie Yang, and Gabriel-Miro Muntean. Computing offloading with fairness guarantee: A deep reinforcement learning method. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):6117–6130, 2023.

[4] Qingqing Wu, Yong Zeng, and Rui Zhang. Joint trajectory and communication design for multi-uav enabled wireless networks. *IEEE Transactions on Wireless Communications*, 17(3):2109–2121, 2018.

[5] Jingdi Chen, Yimeng Wang, and Tian Lan. Bringing fairness to actor-critic reinforcement learning for network utility optimization. In *IEEE Conference on Computer Communications*, pages 1–10, 2021.

[6] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 6116–6135. PMLR, 2024.

[7] Huiying Zhong, Zhun Deng, Weijie J. Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback, 2024.

[8] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation, 2024.

[9] Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with generalized gini welfare functions. In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers*, pages 3–29, Cham, 2024. Springer Nature Switzerland.

[10] Kevin W. S. Roberts. Interpersonal Comparability and Social Choice Theory. *The Review of Economic Studies*, 47(2):421–439, January 1980.

[11] Hervé Moulin. *Fair Division and Collective Welfare*. The MIT Press, January 2003.

[12] Cyrus Cousins. Revisiting fair-pac learning and the axioms of cardinal welfare. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6422–6442, 2023.

[13] Kanad Shrikar Pardeshi, Itai Shapira, Ariel D. Procaccia, and Aarti Singh. Learning social welfare functions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[14] Conor F. Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, April 2022.

[15] Zimeng Fan, Nianli Peng, Muhang Tian, and Brandon Fain. Welfare and fairness in multi-objective reinforcement learning. *arXiv preprint arXiv:2212.01382*, 2022.