Learning Fair Pareto-Optimal Policies in Multi-Objective Reinforcement Learning

Umer Siddique, Peilang Li, Yongcan Cao

Keywords: Multi-objective reinforcement learning, Deep reinforcement learning, Fair optimization, Welfare functions

Summary

Fairness is important in multi-objective reinforcement learning (MORL), where policies must balance optimality and equity across objectives. While *single-policy* MORL methods can learn fair policies for fixed user preferences using welfare, they fail to generalize for different user preferences. To address this limitation, we propose a novel framework for fairness in *multi-policy* MORL, which learns a set of fair policies. Our theoretical analysis establishes that for concave and piecewise-linear welfare functions, fair policies remain in the convex coverage set (CCS). Additionally, we demonstrate that non-stationary and stochastic policies improve fairness over stationary and deterministic policies. Building on our theoretical analysis, we introduce three scalable methods: an extension of Envelope for fair stationary policies, a non-stationary counterpart using state-augmented accrued rewards, and a novel extension for learning stochastic policies. We validate our methods through extensive experiments across three domains and show that our methods fairer solutions as compared to MORL baselines.

Contribution(s)

1. We introduce a novel framework for fairness in multi-policy MORL, which enables learning a set of fair policies for varying user preferences.
   Context: Prior work on fairness in MORL has mainly focused on a single policy for predefined preference weights via some welfare functions. Our framework generalizes fairness across multiple policies, which allow end users to select any policy provided by their preference weights.
2. We provide theoretical analysis demonstrating that for concave, piecewise-linear welfare functions, fair policies remain in the convex coverage set (CCS). Additionally, we establish that non-stationary and stochastic policies can enhance fairness over stationary and deterministic policies, respectively.
   Context: Existing work has explored fairness in RL for predefined preference weights but has not theoretically analyzed how non-stationary and stochastic policies can improve fairness for varying preference weights.
3. We propose three scalable methods for learning fair policies in MORL using a single parameterized network: (i) an extension of Envelope (Yang et al., 2019) for learning fair policies, (ii) a non-stationary extension that incorporates state-augmented accrued rewards to adaptively improve fairness, and (iii) a novel stochastic policy learning method that further enhances fairness.
   Context: Unlike prior work on MORL, which typically learns Pareto optimal policies, our methods efficiently learn a set of fair policies while maintaining scalability.

Learning Fair Pareto-Optimal Policies in Multi-Objective Reinforcement Learning

Umer Siddique, Peilang Li, Yongcan Cao
muhammadumer.siddique@my.utsa.edu,
peilang.li@my.utsa.edu, yongcan.cao@utsa.edu

Department of Electrical and Computer Engineering,
University of Texas at San Antonio

Abstract

Fairness is an important aspect of decision-making in multi-objective reinforcement learning (MORL), where policies must ensure both optimality and equity across multiple, potentially conflicting objectives. While *single-policy* MORL methods can learn fair policies for fixed user preferences using welfare functions such as the *generalized Gini welfare function* (GGF), they fail to provide the diverse set of policies necessary for dynamic or unknown user preferences. To address this limitation, we formalize the fair optimization problem in *multi-policy* MORL, where the goal is to learn a set of Pareto-optimal policies that ensure fairness across all possible user preferences. Our key technical contributions are threefold: (1) We show that for concave, piecewise-linear welfare functions (e.g., GGF), fair policies remain in the *convex coverage set* (CCS), which is an approximated Pareto front for linear scalarization. (2) We demonstrate that non-stationary policies, augmented with accrued reward histories, and stochastic policies improve fairness by dynamically adapting to historical inequities. (3) We propose three novel algorithms, which include integrating GGF with multi-policy multi-objective Q-Learning (MOQL), state-augmented multi-policy MOQL for learning non-statoinary policies, and its novel extension for learning stochastic policies. We evaluate our algorithms across various domains and compare our methods against the state-of-the-art MORL baselines. The empirical results show that our methods learn a set of fair policies that accommodate different user preferences.

1 Introduction

Multi-objective reinforcement learning (MORL) is an important topic in the area of reinforcement learning (RL) that focuses on designing control policies to optimize multiple objectives simultaneously. While traditional MORL methods focus on learning Pareto optimal solutions—ensuring no objective can be improved without sacrificing another—they often neglect fairness, which requires equitable treatment of all objectives or users in our context. For example, in healthcare, a policy may aim to maximize overall patient outcomes (optimality) while ensuring equal treatment across different demographic groups (fairness). A common approach to solving fairness in MORL is to use *utilitarian* welfare functions, where user utilities are aggregated, typically via weighted sum, into a scalarized objective. Despite its simplicity, this approach struggles with fairness, as some users' utilities may be significantly reduced to achieve overall efficiency. An alternative approach is to employ an *egalitarian* welfare function, which prioritizes the least advantaged user by maximizing the minimum utility. While this approach improves fairness, it often leads to inefficient solutions overall, as it optimizes only the lowest utility without ensuring fairness across all objectives.

Several works have explored fairness in the *single-policy* RL setting (Weng, 2019; Siddique et al., 2020; Zimmer et al., 2021; Chen & Hooker, 2021; Do & Usunier, 2022; Fan et al., 2022; Yu et al., 2023b; Nashed et al., 2023), where a single fair policy is learned. For instance, Siddique et al. (2020) enforces fairness using the GGF as a scalarized function and assigning appropriate weights to different objectives to ensure their equitable treatment. Extensions have been explored in multi-agent RL (Zimmer et al., 2021; Siddique et al., 2024b) and preferential treatment under known preference weights (Yu et al., 2023b). Recently, fairness has been studied in multi-policy MORL (Cimpeana et al., 2023; Michailidis et al., 2024) where Cimpeana et al. (2023) defined several fairness notions, while (Michailidis et al., 2024) proposed the Lorenz Condition Network (LCN), an extension of the Pareto Conditioned Network (PCN), which trains a policy network in a supervised manner to map states to desired returns. Despite these works, the investigation of fairness in RL still poses some limitations, including (1) learning a *single* fair policy, (2) required knowledge of the welfare function (e.g., scalarized function) with preference weights a prior, and (3) training a conditioning network on specific return targets, limiting their ability to generalize to unseen preferences. Hence, existing methods operate under fixed preferences and cannot be generalized for all possible preferences.

To address these limitations, we propose a novel framework for addressing fairness in *multi-policy* MORL, rather than the traditional *single-policy* MORL that is the focus of existing work. Our methods are highly scalable as they leverage a single parameterized network to learn an undominated set of policies, specifically a convex coverage set (CCS), by sampling the entire preference space in MORL. In particular, to address fairness, we apply the welfare function (e.g., GGF) during learning for each sampled preference weight to ensure that each learned policy treats its objectives fairly. We further introduce non-stationary action selection using the state-augmented accrued rewards to enhance fairness by effectively utilizing historical information. We further demonstrate the benefits of learning stochastic policies for fairness. Motivated by hindsight experience replay (Andrychowicz et al., 2017), we incorporate resampling of random preference weights across different preference conditions to improve sample efficiency in MORL, as it is done in (Yang et al., 2019).

The main contributions of this paper are as follows:

1. We introduce a novel framework for fairness in multi-policy MORL, enabling users to select any fair policy based on their specific preferences, thereby enhancing user satisfaction(Section 3.2).
2. We provide theoretical analysis establishing that for concave, piecewise-linear welfare functions (e.g., GGF), fair policies remain in CCS. Additionally, we demonstrate that non-stationary policies can improve fairness by adapting to historical disparities and that stochastic policies further improve fairness over deterministic policies(Section 4).
3. Building on our theoretical insights, we propose three scalable methods for learning fair policies in MORL using a single parameterized network: (i) an extension to Envelope (Yang et al., 2019) for learning fair stationary policies, (ii) a non-stationary counterpart that incorporates state-augmented accrued rewards to improve fairness over time adaptively, and (iii) a novel extension for learning stochastic policies, which further enhances fairness(Section 5).
4. We experimentally validate our methods and demonstrate their effectiveness compared to state-of-the-art MORL and fairness methods across three different domains(Section 6).

2 Related Work

Fairness in machine learning (ML) has become a significant research direction (Dwork et al., 2012; Zafar et al., 2017; Sharifi-Malvajerdi et al., 2019; Singh & Joachims, 2019; Chierichetti et al., 2017; Busa-Fekete et al., 2017; Agarwal et al., 2018; Nabi et al., 2019; Zhang & Liu, 2021). Several studied have addressed fairness in model predictions (Speicher et al., 2018), recommender systems (Leonhardt et al., 2018), classification (Dwork et al., 2012; Zafar et al., 2017; Agarwal et al., 2018; Kim et al., 2019), and ranking (Singh & Joachims, 2019). While much of the literature focuses on the principle of "equal treatment of equals", other aspects, such as proportionality (Bei et al., 2022) or envy-freeness (Chevaleyre et al., 2006) and its multiple variants (e.g., (Beynier et al.,

2019; Chakraborty et al., 2021)), have been considered in ML. In contrast, our work is grounded in distributive justice (Rawls, 1971; Brams & Taylor, 1996; Moulin, 2004), with a focus on optimizing a welfare function for fairness considerations. This principled approach has also been recently advocated in several papers (Heidari et al., 2018; Speicher et al., 2018; Cousins, 2021).

Recently, fairness in RL has gained significant attention with the work by (Jabbari et al., 2017), which ensures fairness in state visitation using scalar rewards. The work of (Jiang & Lu, 2019) proposed FEN, a hierarchical decentralized approach using gossip algorithms to ensure fairness among agents. Similarly, Chen et al. (2021) proposed to incorporate fairness into actor-critic RL algorithms, optimizing general fairness utility functions for real-world network optimization problems. Considering the multi-objective nature of many RL problems, the study of fairness in MORL has been widely studied. In particular, Siddique et al. (2020) proposed multiple adaptations to deep RL algorithms that optimize the GGF. Zimmer et al. (2021); Siddique et al. (2024a) extended this to the decentralized cooperative MARL. Fan et al. (2022) proposed to optimize the Nash welfare function using scalarized expected return criterion, while Do & Usunier (2022) proposed to optimize GGF in rankings. Yu et al. (2023b); Qian et al. (2025) proposed methods that learn a fair policy providing preferential treatment to some users while ensuring equal treatment of all others under the assumption that these preferential weights are known in advance. Siddique et al. (2023) proposed FPbRL, which learns fair preference-based policies without true rewards. Recently, fairness has been considered in multi-policy MORL with Michailidis et al. (2024) propose learning Lorenz Condition networks, which ensures fairness through Lorenz domination and adds an extra parameter λ, however, we use the welfare function to learn a set of fair optimal policies.

Despite the significant successes achieved in the field of deep RL and MORL, existing methods heavily rely on scalarization functions to learn a *single policy* with fixed preference weights. However, such single-policy methods do not work when preferences are unknown or user-specific solutions are required. To address this limitation, several works have been proposed to accommodate user-specific preferences, including but not limited to those proposed by (Barrett & Narayanan, 2008; Van Moffaert et al., 2013; Moffaert & Nowé, 2014; Yang et al., 2019; Alegre et al., 2023; Reymond et al., 2022). Notably, these methods aim to learn a set of policies that approximate the Pareto frontier of optimal solutions. For instance, (Barrett & Narayanan, 2008) and (Moffaert & Nowé, 2014) proposed methods to compute policies on the Pareto front's convex hull, while (Yang et al., 2019) introduced envelope Q-learning, learning policies from the convex coverage set (CCS). These approaches, however, do not address fairness, which is the focus of this paper.

3 Preliminaries

3.1 Multi-Objective Markov Decision Process

A multi-objective Markov Decision Process (MOMDP) extends the classical MDP framework to scenarios where an agent must optimize multiple objectives simultaneously. An MDP (Puterman, 1994) is defined by the tuple, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions available to the agent, $\mathcal{P}_{a,s,s'} \in [0, 1]$ is the probability of transition from state s to state s' after taking action a, *i.e.*, $\mathcal{P}(s'|s, a) = \mathcal{P}[S_{t+1} = s'|S_t = s, A_t = a]$, $r(s, a) : s \times a \mapsto r$ is the immediate reward obtained by taking action a at state s, and $\gamma \in [0, 1)$ is the discount factor. An MOMDP can be represented by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \boldsymbol{r}, \gamma, \Omega, f_\Omega)$, in which the definitions of $\mathcal{S}, \mathcal{A}, \mathcal{P}$, and γ are the same as in MDP except that the reward $\boldsymbol{r}$ is now a vector, with each component corresponding to an objective that the agent seeks to optimize. Here, the additional Ω represents the entire space of preferences, and f_Ω is the preference function which takes a linear form, producing a single utility $f_{\boldsymbol{\omega}}(\boldsymbol{r}) = \boldsymbol{\omega}^T \boldsymbol{r}(s, a)$, where $\boldsymbol{\omega}$ is a vector representing the preference weights for different objectives. In MOMDPs, the objectives may be conflicting, and hence it is often difficult to optimize all objectives simultaneously.

The goal of an agent in an MOMDP is to either learn a single policy that balances multiple objectives or a set of policies that optimize different trade-offs among objectives. These approaches are referred

to as *single-policy* MORL and *multi-policy* MORL, respectively. A policy π is a strategy that maps states to actions, which can be deterministic (i.e., $\forall s, \pi(s) \in \mathcal{A}$) or stochastic (i.e., $\forall s, a, \pi(a|s)$ denotes the probability of selecting a in s). In MOMDPs, policies are typically *stationary* (Markovian), with action probabilities depending only on the current state, while a non-stationary (adaptive) policy $\pi(a|\tau, s)$ may also depend on the history τ. Standard definitions in MDPs, such as the return $G(\tau)$ and the value functions V or Q, extend naturally to MOMDPs, albeit represented as vectors and matrices respectively. The vector return in an MOMDP is expressed as $\boldsymbol{G}(\tau) = \sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{r}_t$, where τ is a trajectory comprising a sequence of states, actions, and rewards following the policy, and $\boldsymbol{r}_t$ is a vector reward obtained at time step t. The state value function of a policy π in an MOMDP is defined as $\boldsymbol{V}^{\pi}(s) = [V_i^{\pi}(s)] = \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{r}_t \mid S_0 = s \right]$, where all operations (addition, product) are applied component-wise.

In MOMDPs, value functions do not offer a complete ordering over the policy space. This means it is possible to encounter scenarios wherein, e.g., $V_i^{\pi}(s) > V_i^{\pi'}(s)$ for objective i, while $V_j^{\pi}(s) < V_j^{\pi'}(s)$ for objective j. Hence, value functions in MOMDPs induce only a partial ordering within the policy space, necessitating additional information into objective prioritization for policy ordering.

Envelope Multi-Objective Q-Learning. The Envelope algorithm (Yang et al., 2019) learns a convex coverage set (CCS) by sampling preference weights $\boldsymbol{\omega} \in \Omega$ and optimizing linearly scalarized Q-values: $Q(s, a, \boldsymbol{\omega}) = \boldsymbol{\omega}^T \boldsymbol{Q}(s, a)$, where $\boldsymbol{Q}(s, a) \in \mathbb{R}^N$ is the vector of Q-values for N objectives. The Bellman optimality equation for Envelope algorithm is: $\boldsymbol{Q}^*(s, a, \boldsymbol{\omega}) = \boldsymbol{r}(s, a) + \gamma \max_{a'} \boldsymbol{\omega}^T \boldsymbol{Q}^*(s', a')$. A single neural network parameterizes $\boldsymbol{Q}(s, a, \boldsymbol{\omega})$ by concatenating $\boldsymbol{\omega}$ to the state s, enabling efficient learning across all preferences. Despite its scalability, Envelope lacks explicit fairness guarantees, as linear scalarization may prioritize dominant objectives.

3.2 Fairness Formulation

In MORL, fairness, rooted in distributive justice (Moulin, 2004), is crucial for ensuring equitable distribution of rewards. Prior studies in fair optimization within MORL have primarily focused on learning a *single-policy*, commonly referred to as an average policy (Siddique et al., 2020; Fan et al., 2022; Yu et al., 2023a). In this paper, we adopt a more inclusive view of fairness, including *efficiency*, *equity*, and *impartiality* to generate fair optimal solutions for user-specific preferences. For discussion on fairness and welfare function, please refer to the Appendix.

Definition 3.1. *Efficiency states that among two solutions, if one solution is (weakly or strictly) preferred by all users, then it should be preferred to the other one, e.g., $\boldsymbol{V} \succ \boldsymbol{V}' \Rightarrow \phi(\boldsymbol{V}) > \phi(\boldsymbol{V}')$, where $\phi(\boldsymbol{V})$ is the scalar utility function by using the ϕ that specifies the value of a solution.*

The efficiency property specifies that given all else equal, one prefers to increase a user's utility. In the MORL setting, the efficiency property simply means Pareto dominance. More specifically, a solution is considered efficient if it is not dominated by any other solution for all objectives.

Definition 3.2. *For a given pair of solutions $\boldsymbol{V}, \boldsymbol{V}' \in \mathbb{R}^N$, $\boldsymbol{V}$ weakly Pareto-dominates $\boldsymbol{V}'$ if $\forall i, V_i \geq V_i'$, $\forall i \in \{1, \cdots, N\}$, where N is the total number of objectives. Besides, $\boldsymbol{V}$ Pareto-dominates $\boldsymbol{V}'$ if $V_i \geq V_i', \forall i$ and $\exists j, V_j > V_j'$. For brevity, we denote Pareto dominance as $\geq$ for the weak form and $>$ for the strict form.*

Essentially, a solution $\boldsymbol{V}$ (weakly) Pareto-dominates another solution $\boldsymbol{V}'$ if the former's value $\phi(\boldsymbol{V})$ (weakly) Pareto-dominates that of the latter $\phi(\boldsymbol{V}')$. A solution $\boldsymbol{V}^*$ is said to be *Pareto-optimal* if no other solution $\boldsymbol{V}$ Pareto-dominates it. *Pareto front* $(\mathcal{F})$ is defined as the set of Pareto-optimal solutions, which may consist of infinitely many solutions, especially when policies can be stochastic. A typical way to approximate $(\mathcal{F})$ is to compute the convex coverage set (CCS), defined below.

Definition 3.3. *A solution in CCS has a maximal scalarized value in a weighted sense if there exists a weight vector $\boldsymbol{\omega} \in \Omega$ such that the scalarized utility $\boldsymbol{\omega}^T \boldsymbol{V}$ is weakly preferred to the scalarized utility $\boldsymbol{\omega}^T \boldsymbol{V}'$ for all other solutions $\boldsymbol{V}'$ in the Pareto front. Formally speaking, $\boldsymbol{V} \in$ CCS $\iff$ $\exists\, \boldsymbol{\omega} \in \Omega$ s.t. $\boldsymbol{\omega}^T \boldsymbol{V} \geq \boldsymbol{\omega}^T \boldsymbol{V}', \forall\, \boldsymbol{V}' \in \mathcal{F}$.*

Next, we discuss the significance of the *equity* property, a stronger property than efficiency and often associated with distributive justice, as it refers to the fair distribution of resources or opportunities. This property ensures that a fair solution follows the *Pigou-Dalton principle* (Moulin, 2004), which states the transferring of rewards from more advantaged users to less advantaged users.

Definition 3.4. *A solution satisfies the Pigou-Dalton principle if for all $\boldsymbol{V}$, $\boldsymbol{V}'$ equal except for $V_i = V'_i + \delta$ and $V_j = V'_j - \delta$ where $V'_i - V'_j > \delta > 0$, $\phi(\boldsymbol{V}) > \phi(\boldsymbol{V}')$.*

Finally, the *impartiality* property, which is rooted in the principle of "equal treatment of equals" states that individuals sharing similar characteristics should be treated similarly.

Definition 3.5. *In a system, individuals with similar characteristics should be treated similarly, i.e., the solution should be independent of the order of its arguments $\phi(\boldsymbol{V}) = \phi(\boldsymbol{V}_\sigma)$, where σ is a permutation and $\boldsymbol{V}_\sigma$ is the vector obtained from vector $\boldsymbol{V}$ permuted by σ.*

To ensure fairness that satisfies the above three properties, we use a well-known generalized Gini welfare function (GGF) (Weymark, 1981), which can be defined as:

$$\phi_{\text{GGF}}(\boldsymbol{u}) = \sum_{i \in N} \omega_i u_i^{\uparrow}, \tag{1}$$

$\boldsymbol{u} \in \mathbb{R}^N$ represents the utility vector of a size N for N objectives, $\boldsymbol{\omega} \in \mathbb{R}^N$ is a fixed weight vector with positive components that strictly decrease (i.e., $\omega_1 > \ldots > \omega_N$) with $\sum_i w_i = 1$, and $\boldsymbol{u}^{\uparrow}$ denotes the vector by sorting the components of $\boldsymbol{u}$ in an increasing order (i.e., $u_1^{\uparrow} \leq \ldots \leq u_N^{\uparrow}$). GGF satisfies the aforementioned three fairness properties. As the weights are positive, it is monotonic with respect to Pareto dominance, thus satisfying the efficiency property. Since the utility vector is reordered, it is also symmetric and therefore satisfies the impartiality property. Furthermore, the positive and decreasing weights ensure that GGF is Schur-concave, i.e., monotonic with respect to Pigou-Dalton transfers, therefore satisfies the impartiality property.

GGF has been studied and used in MORL extensively (Siddique et al., 2020; Mandal & Gan, 2022; Yu et al., 2023a; Qian et al., 2025), however, prior works have focused exclusively on the single-policy setting. To our knowledge, we are the first to apply GGF in a multi-policy MORL context. In multi-policy MORL, the standard approach is to identify all Pareto non-dominated solutions (Mukai et al., 2012; Van Moffaert & Nowé, 2014); however, this is impractical for large-scale problems, as the Pareto front grows exponentially. A more scalable alternative is to approximate the CCS, which forms the convex envelope of optimal trade-offs

4 Fairness in MORL

Since we are in a multi-policy MORL setting, where an agent learns a set of Pareto optimal policies, fairness becomes more important as different stakeholders may have different preferences, and during inference, any solution can be used from the Pareto non-dominated solutions given the stakeholder preferences. We formalize this sophisticated multi-policy fair optimization problem as:

$$\forall \boldsymbol{\omega} \in \Omega, \quad \max_{\pi \in \Pi} \ \phi_{\text{GGF}}(\boldsymbol{J}(\pi)), \tag{2}$$

where Ω is the set of valid preference weights sorted in descending order, $\boldsymbol{J}(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t]$ is the expected discounted return, and $\phi_{\text{GGF}}(\boldsymbol{J}) = \sum_{i=1}^{N} w_i J_{(i)}$ with $J_{(1)} \leq \cdots \leq J_{(n)}$. The concavity of GGF makes problem (2) as convex optimization problem, enabling efficient solutions within the CCS. Below, we establish three foundational results, which show that it is always feasible to obtain optimal solutions in the CCS corresponding to GGF fair optimization. Next, we demonstrate that a non-stationary policy based on accrued rewards is beneficial in yielding improved fairness when compared with its stationary counterpart. Here, a policy yields improved fairness or is fairer if a higher welfare score, defined in (1), is achieved. Lastly, we show that a stochastic policy may yield fairer solutions than a deterministic one.

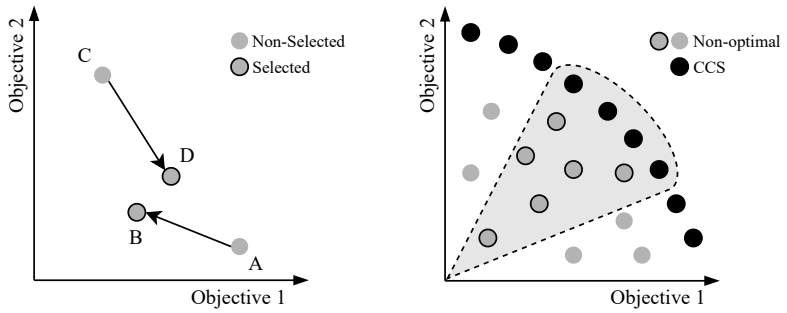


Figure 1: Examples of 2-objective MOMDP where GGF leads to fairer outcomes.

Sufficiency of Optimal Solutions in the CCS. The first question relates to the learning of fair policies in a multi-policy MORL setting is which subset of policies may be optimal among the set of all (possibly non-stationary) policies. Indeed, for linear scalarization function, CCS contains the set of Pareto front solutions. Below, we formally state it:

Lemma 4.1. *For any MOMDP with linear preferences over objectives, the CCS contains an optimal policy for any linear combination of the objectives.*

While GGF introduces non-linear fairness objectives, its piecewise linearity and concavity allow it to be expressed as a maximum over linear functions, which ensures that optimal solutions lie within the CCS. The following proposition establishes the sufficiency of the CCS in representing optimal policies for ϕ_{GGF} preference weights.

Proposition 4.1. *For any $s \in \mathcal{S}$ in an MOMDP and a piecewise-linear concave welfare function ϕ_{GGF} (e.g., GGF) that can be represented as, $\phi_{GGF}(\boldsymbol{V}^{\pi}(s)) = \min_{\sigma \in \mathbb{S}_N} \{\boldsymbol{\omega}_{\sigma}^{\top} \boldsymbol{V}^{\pi}(s)\}$, there exists a policy $\pi^* \in$ CCS such that $\phi_{GGF}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{GGF}(\boldsymbol{V}^{\pi}(s)), \quad \forall \pi \in \Pi$.*

Example 4.1 *To illustrate how the GGF function ensures fairness in MORL, consider a two-objective MOMDP with objective values $\boldsymbol{V}_1 = (3,1)$ and $\boldsymbol{V}_2 = (2,3)$ and weights $(1,2)$. For $\boldsymbol{V}_1$, two weighted combinations are possible: **A)** $(3,1) \cdot (2,1) = (6,1)$ with scalar sum $6+1=7$, **B)** $(3,1) \cdot (1,2) = (3,2)$ with scalar sum $3+2=5$. Since the GGF is defined as $\phi_{GGF}(\boldsymbol{V}^{\pi}(s)) = \min_{\sigma \in \mathbb{S}_N} \{\boldsymbol{\omega}_{\sigma}^{\top} \boldsymbol{V}^{\pi}(s)\}$, it selects the lower scalar value, preferring point B over A (see left figure of Figure 1). Similarly, for $\boldsymbol{V}_2$: **C)** $(2,3) \cdot (1,2) = (2,6)$ with scalar sum $2+6=8$, **D)** $(2,3) \cdot (2,1) = (4,3)$ with scalar sum $4+3=7$. Here, point D is preferred over C. This mechanism directs the solutions toward the fairer region (grey dotted area in the right figure of Figure 1), demonstrating that maximizing the GGF leads to fair Pareto-optimal solutions.*

Fairness of Non-Stationary Policies. In fair MORL, learning non-stationary policies can be beneficial, as they use historical information to make more informed decisions and adapt over time.

Proposition 4.2. *Let the reward $\boldsymbol{r}$ be nonnegative, and Π_S and Π_{NS} be the sets of stationary and non-stationary policies, respectively. For any $s \in \mathcal{S}$ in an MOMDP and a given ϕ_{GGF}, there exists a non-stationary policy $\pi_{NS} \in \Pi_{NS}$ that achieves a higher welfare score than any stationary policy $\pi_S \in \Pi_S$, i.e., $\exists \pi_{NS} \in \Pi_{NS} : \phi_{GGF}(\boldsymbol{V}^{\pi_{NS}}(s)) \geq \max_{\pi_S \in \Pi_S} \phi_{GGF}(\boldsymbol{V}^{\pi_S}(s))$.*

Example 4.2 *To illustrate the value of learning a non-stationary policy, consider a 2-objective MOMDP, shown in Fig. 2. At timestep $t > 0$, the agent has accrued a vector reward $\boldsymbol{r}_{acc} = (10,0)$ for two objectives. The preference weights, encapsulated within the welfare function ϕ, denote decreasing weights, such as $(0.8, 0.2)$. With two potential actions, each leading to a final state, action a_1 yields a reward of $(0,10)$, while action a_2 yields $(5,5)$. Since s_t is the absorbing state, we can set the discount factor $\gamma = 1$. Under the given welfare function ϕ defined in 1, executing a_1 yields a welfare score of 2, whereas executing a_2 yields a score of 5 if only future rewards are*

considered. However, considering historical data, i.e., $\mathbf{r}_{acc}$, a_1 yields a higher accrued episodic return of $(10, 10)$ and a welfare score of 10. Similarly, a_2 yields $(15, 5)$ and 7 episodic return and welfare scores, respectively. Note that action a_1 is a fairer choice in this case since it balances the two objectives, unlike action a_2, which fails to achieve a more equitable outcome. Hence, employing historical data, namely, accrued rewards in this case, is critical to enable fair policy learning.

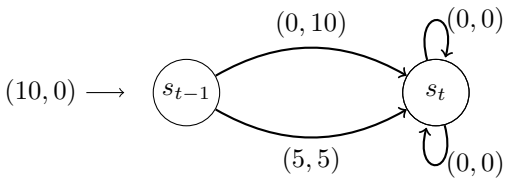


Figure 2: Example of MOMDP where actions lead to different rewards.

Optimality of Stochastic Policies for Fairness Unlike single-objective RL, in MORL, a deterministic policy may not be optimal. A fairer solution can often be achieved through randomization.

Proposition 4.3. *Let Π_{ST} be the set of stochastic policies and Π_D be the set of deterministic policies. For an MOMDP $\mathcal{M}$ and a concave welfare function such as ϕ_{GGF}, there exists a stochastic policy $\pi_{ST} \in \Pi_{ST}$ such that $\phi_{GGF}(\boldsymbol{V}^{\pi_{ST}}) \geq \max_{\pi_D \in \Pi_D} \phi_{GGF}(\boldsymbol{V}^{\pi_D})$.*

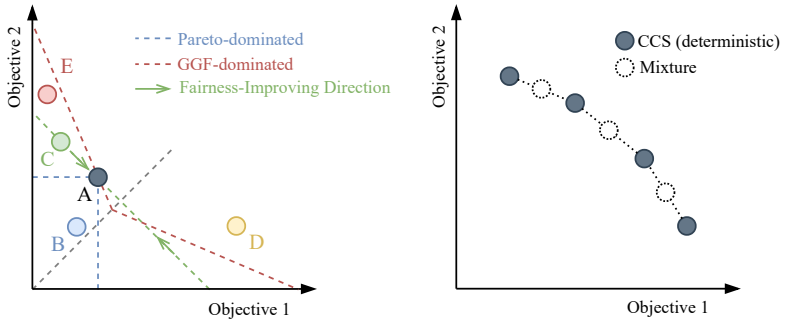


Figure 3: Left Figure: Point A Pareto-dominates B and is preferred to C by the Pigou–Dalton transfer(fairer solution). Depending on GGF weights, D and E may be dominated or non-dominated by A (w.r.t. GGF); for weights $(0.3, 0.7)$, A is preferred to E but not D. Right: Black points denote deterministic policies in the CCS; mixing these yields stochastic policies (dotted points), which can achieve fairer solutions unattainable by any single deterministic policy.

Proofs of the above lemma and propositions are provided in Section 8. The left figure of Figure 3 illustrates GGF in a two-objective task. The optimality of stochastic policies implies that restricting the search to deterministic policies is insufficient, and stochastic policies expand the solution space and can better capture trade-offs, thus improving overall fairness, as shown in Figure 3.

5 Proposed Algorithms

In this section, we introduce three novel algorithms that incorporate fairness into MORL based on the technical analysis in the previous section. These algorithms optimize the GGF function (1) to ensure fairness across N fixed users with varying preferences. These methods are scalable and sample-efficient as they utilize a single parameterized network to estimate Q-values for all objectives while maintaining a diverse set of Pareto-optimal policies. Specifically, we introduce Fair Multi-Objective Deep Q-Learning (F-MDQ), its non-stationary extension (FN-MDQ), and a novel extension incorporating stochastic policies (FNS-MDQ). This progression from stationary to non-stationary to stochastic and non-stationary policies demonstrates our systematic approach to enhancing fairness in MORL algorithms, with each method building upon the previous one.

F-MDQ. F-MDQ builds on the Envelope algorithm (Yang et al., 2019) by replacing the linear scalarization function with the GGF welfare function ϕ. This ensures fairness while learning policies

across all preferences $\boldsymbol{\omega} \in \Omega$. The Bellman optimality equation for F-MDQ is given by:

$$\boldsymbol{Q}^*(s, a, \boldsymbol{\omega}) = \mathbb{E}[\boldsymbol{r}(s, a) + \gamma \boldsymbol{Q}^*(s', \sup_{a' \in \mathcal{A}} \phi_{\text{GGF}}(\boldsymbol{r}(s, a) + Q^*(s', a', \boldsymbol{\omega}), \boldsymbol{\omega}) \mid s, a],$$

where $\boldsymbol{Q}^\pi(s, a, \boldsymbol{\omega})$ represents the expected return vector for policy π, conditioned on preference $\boldsymbol{\omega}$. As the MO Q-function is parameterized, it can be learned by minimizing the loss function $\mathcal{L} = \mathbb{E}_{(s,a,\boldsymbol{r},s',\omega)\sim\mathcal{D}} \left[\|\boldsymbol{y} - \boldsymbol{Q}(s, a, \boldsymbol{\omega})\|_2^2\right]$, where the expectation is taken over experiences sampled from the replay buffer $\mathcal{D}$. Given that the loss function includes an expectation over $\boldsymbol{\omega}$, the preference weights are sampled randomly and are decoupled from the transitions, allowing increased sample efficiency through a resampling scheme similar to Hindsight Experience Replay (HER) (Andrychowicz et al., 2017). The target $\boldsymbol{y}$ is F-MDQ is computed as

$$\boldsymbol{y} = \boldsymbol{r}(s, a) + \gamma \boldsymbol{Q}'(s', \sup_{a' \in \mathcal{A}} \phi_{\text{GGF}}(\boldsymbol{r}(s, a) + \gamma Q(s', a', \boldsymbol{\omega})), \boldsymbol{\omega}),$$

where Q' represents the target multi-objective Q-function, and the supremum is applied over the GGF welfare function ϕ_{GGF} instead of a linear weighted sum. This ensures that actions are selected based on higher welfare scores rather than simply maximizing Q-values.

FN-MDQ. FN-MDQ extends F-MDQ by incorporating accrued rewards into the state to learn non-stationary policies, as discussed in Proposition 8.2. It augments the state with accrued rewards, allowing the agent to balance reward distribution across users (as demonstrated in Example 2). The augmented state is defined as $\mathfrak{s}_t = (s_t, \boldsymbol{r}_{\text{acc}})$, where $\boldsymbol{r}_{\text{acc}} = \sum_{i=1}^{t-1} \gamma^{i-1} \boldsymbol{r}_i$ is the discounted reward received in the current trajectory. The regression target for FN-MDQ is given by

$$\boldsymbol{r}(s_t, a_t) + \gamma \boldsymbol{Q}'(\mathfrak{s}_{t+1}, \sup_{a' \in \mathcal{A}} \phi_{\text{GGF}}(Q(\mathfrak{s}_{t+1}, a', \boldsymbol{\omega})), \boldsymbol{\omega}).$$

Here, the immediate reward $\boldsymbol{r}(s_t, a_t)$ is excluded from the optimal action since this is already included in the augmented state as part of the discounted total reward. This extension enables the agent to identify and prioritize users who have received insufficient rewards within a trajectory.

FNS-MDQ. Given that stochastic policies can outperform deterministic ones (as established in Proposition 8.3), the performance of FN-MDQ can be enhanced by incorporating stochastic policies. We now explain how stochastic policies can be integrated into the FN-MDQ algorithm.

Under the stochastic policies, the target Q-value is adjusted to account for the expected Q-values, which reformulates the update as

$$\boldsymbol{r}(s_t, a_t) + \gamma \boldsymbol{Q}'(\mathfrak{s}_{t+1}, \sum_{a' \in \mathcal{A}} \phi_{\text{GGF}}(\pi(a' \mid \mathfrak{s}_{t+1}) Q(\mathfrak{s}_{t+1}, a', \boldsymbol{\omega})), \boldsymbol{\omega}),$$

where $\pi(a' \mid \mathfrak{s}_{t+1})$ is the probability of taking action a' given the augmented state $\mathfrak{s}_{t+1}$. This reformulation considers the distribution of possible actions rather than selecting a single best deterministic action, aligning with our theoretical insights.

Unlike F-MDQ and FN-MDQ, which rely on deterministic action selection, FNS-MDQ samples actions from a probability distribution over Q-values. This stochastic action selection improves fairness by enabling more balanced policy exploration and reducing biases that arise from always selecting the highest Q-value action. Note that, during the training phase, all algorithms employ an ϵ-greedy policy during training, however, FNS-MDQ differs in its action-selection strategy by using the best learned stochastic policy rather than a deterministic greedy approach. This increased flexibility and randomness can lead to more equitable solutions.

6 Experiments

To evaluate the proposed methods, we conduct experiments across three domains—each characterized by varying levels of complexity in terms of the number of objectives. These domains, ranging from low to high in terms of the number of objectives, include species conservation, resource

gathering, and multi-product web advertising. Each environment presents unique challenges where fairness plays a critical role. We first briefly describe each environment (details are available in Appendix B) and then present our experimental results.

6.1 Environments

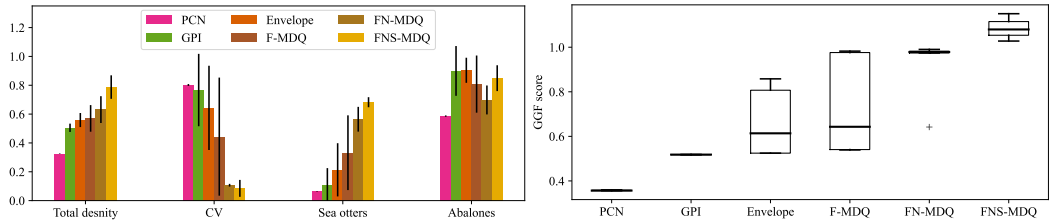
Our first domain is a species conservation (SC) environment, which addresses a critical ecological challenge: balancing the populations of two highly interacting endangered species, the sea otter and the northern abalone. Both species are at risk of extinction, requiring sophisticated management strategies to ensure their survival. We adopt the model proposed by (Chadès et al., 2012), which simulates the predation relationship between the species, where sea otters prey on abalones. This dynamic presents a unique preservation challenge, as the survival of one species could potentially drive the other to extinction if not properly managed. The state space is composed of the current population sizes of sea otters and northern abalones. The action space includes introducing sea otters, enforcing anti-poaching measures, controlling sea otter populations, implementing a combination of half-antipoaching and half-controlled sea otters, or taking no action. Each action has significant ecological implications. For instance, introducing sea otters may help balance the abalone population, but if mismanaged, could lead to abalone extinction. The reward function is defined by the population densities of both species, i.e., $N = 2$. Fairness in this context is interpreted as achieving a balanced distribution of species densities to ensure their preservation.

Our second environment is a resource-gathering (RG) problem, which is a 5×5 grid world that contains three types of resources: gold, gems, and stones. These resources are randomly positioned on the grid and regenerate randomly upon consumption. The main challenge here is to collect these resources, where each resource has a different value: gold and gems are valued at 1, while stones have a lower value of 0.4. This creates an intentionally uneven resource distribution, with two stones, one gold, and one gem. In this environment, the state is defined by the agent's current location on the grid and the cumulative count of each resource collected during its trajectory. The agent can take four actions: up, down, left, and right. The reward is a vector representing the resources collected for each type. In this environment, fairness is defined as the equitable collection of resources, despite their differing values. Note that this problem is particularly important for validating whether the proposed methods can achieve fairer solutions while still reaching Pareto optimal solutions.

Our third domain is a multi-product web advertising (MWP) problem that involves an online store offering $N = 7$ distinct products. Here, the agent decides which advertisement to display: a product-specific advertisement for one of the products $i \in [0, ..., N-1]$, or a general advertisement that is not tailored to any specific product. In this environment, the state space includes the number of products available in the store, as well as the number of visits, purchases, and exits. The action space is $N+1$, where actions 0 through $N-1$ correspond to displaying advertisements for specific products, and action N involves showing a general advertisement. This additional action adds complexity, requiring the agent to decide the optimal moment to transition between states. The reward function is designed so that the agent receives a reward of 1 in the i^{th} dimension of the reward vector if a product of the type i is sold after displaying its advertisement. In this environment, fairness is defined as balancing the frequency of advertisements shown for each product, ensuring no single product is overly prioritized. The challenge lies in increasing overall rewards while maintaining a fair distribution of advertisement exposure across all products.

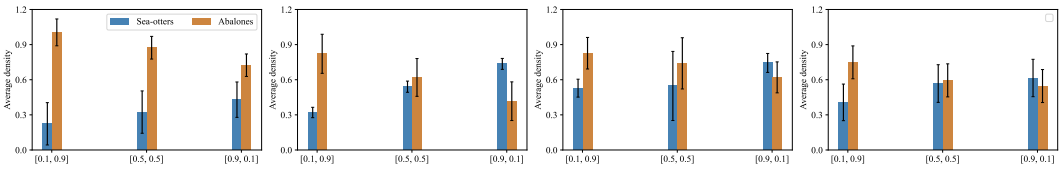
6.2 Baselines

We compare our proposed methods against several multi-policy MORL baselines. Generalized Policy Improvement Linear Support (GPI-LS) (Alegre et al., 2023) employs GPI (Barreto et al., 2017) to combine policies within its learned CCS and prioritize the weight vectors on which agents should train at each moment. The Envelope (Yang et al., 2019) uses a single neural network conditioned on a weight vector to approximate the CCS. PCN (Reymond et al., 2022) utilizes a neural network conditioned on a desired return per objective and is trained via supervised learning to predict actions



(a) Total density, CV, min density, and max density. (b) GGF scores.

Figure 4: Performances of multi-policy MORL baselines and our methods in species conservation.



(a) Envelope. (b) F-MDQ. (c) FN-MDQ. (d) FNS-MDQ.

Figure 5: Individual densities of Envelope, and our proposed methods during testing with unseen preferences in species conservation.

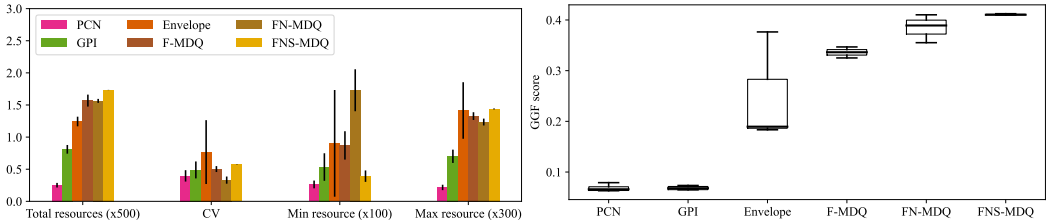
that yield the desired return. Hyperparameters for each method were optimized, and experiments were run for five different seeds, with average results reported. Further details on experimental configurations and hyperparameters are provided in Appendix C.

6.3 Results

In this section, we present the experimental results across the three environments presented above. The primary objective of these experiments is to assess the effectiveness of our proposed methods by addressing the following key research questions: **(A)** How effective are our methods in learning fairer solutions compared to multi-policy MORL baselines? **(B)** Can our methods generate fair solutions across different preference settings during inference? **(C)** To what extent can our proposed algorithms achieve comparable performance in terms of hypervolume and cardinality relative to multi-policy MORL approaches? **(D)** What is the impact of our approach on the diversity and quality of non-dominated solutions that satisfy fairness criteria? **(E)** Does the incorporation of stochastic policies in MO Q-learning based algorithms contribute to improved fairness or overall performance?

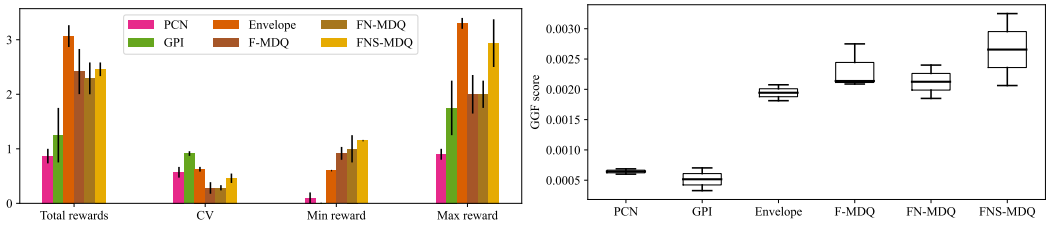
Question (A) To evaluate how effective our methods are in learning fair solutions, we conducted experiments in the SC, RG, and MWP domains, as shown in Figures 4a, 6a and 7a. We compare our proposed methods (F-MDQ, FN-MDQ, and FMS-MDQ) with multi-policy MORL baselines such as PCN, GPI, and Envelope during the training phase. We choose these baselines as they are the current state-of-the-art MORL baselines. The Key evaluation metrics used include total rewards, Coefficient of Variation (CV) indicating the variations in different objectives' utilities, and the minimum and maximum objective utilities. Moreover, GGF welfare scores were computed to quantify fairness. As we are in a multi-policy MORL, an agent learns a set of Pareto optimal policies during learning. To show the results, we computed these metrics over the last 50 trajectories for all the Pareto optimal policies and reported their normalized scores. Note that, during the last 50 trajectories, all the agents are converged so it ensures a fair comparison for multi-policy MORL methods.

As shown in Figure 4a, PCN performs the worst. GPI outperforms PCN, likely due to its TD3-based (Fujimoto et al., 2018) architecture and efficient prioritization scheme in learning the Pareto front $\mathcal{F}$. The Envelope algorithm performs better than PCN and GPI as it achieves higher total density and, interestingly, lower CV. However, our proposed algorithms outperform all other methods by achieving the lowest CV and highest welfare scores Figure 4b, with FN-MDQ outperforming



(a) Total resources, CV, min and max resources. (b) GGF scores.

Figure 6: Performances of multi-policy MORL baselines and our methods in resource gathering.



(a) Total rewards, CV, min reward, and max reward. (b) GGF scores.

Figure 7: Performances of multi-policy MORL baselines and our proposed methods in the MPW.

F-MDQ, underscoring the value of non-stationary policies. Furthermore, FNS-MDQ outperforms both F-MDQ and FN-MDQ as it maximizes the minimum objective utility and demonstrates better fairness through optimizing the welfare function ϕ_{GGF}. Similar results are observed in RG Figure 6a, where PCN performs the worst as it collects the least resources, likely due to its limitations in deterministic environments (Reymond et al., 2022). Although GPI performs better than PCN, both exhibit low CV alongside poor overall performance and GGF welfare utility Figure 6b. The Envelope algorithm achieves better performance in terms of rewards but suffers from the highest CV and lower GGF utility scores. In contrast, our proposed methods attain a lower CV compared to all baselines, and they achieve the highest GGF scores, highlighting their effectiveness in identifying fair policies through welfare function optimization. Interestingly, FNS-MDQ exhibits a higher CV due to its higher maximum objective and the total resources collected. Nevertheless, it also achieves the highest welfare scores. Consistent with our previous results, our proposed methods in MVP environment Figure 7a achieve the highest welfare scores, indicating their capacity to ensure an equitable distribution of rewards across all objectives. Moreover, they maintain the lowest CV, highlighting their robustness in learning fair policies, even in highly stochastic environments with a higher number of objectives. Once again, PCN, and GPI perform the worst, further underscoring the efficacy of our methods in this context.

Question (B) To check whether our methods can generate fair solutions across different preference settings, we evaluated our algorithms with unseen preferences during testing in the SC environment. As shown in Figure 5, which presents the individual species densities (sea otters and abalones) for preference configurations $(0.1, 0.9), (0.5, 0.5), (0.9, 0.1)$, the Envelope algorithm fails to produce fair solutions, suggesting its limitation in generating fair optimal policies across varying preferences. In contrast, F-MDQ generates more balanced solutions, while FN-MDQ and FNS-MDQ achieve even fairer outcomes, further validating our earlier findings.

Question (C) To answer this question, we evaluate algorithms in terms of MORL metrics, such as cardinality and hypervolume (HV). A higher cardinality indicates greater policy diversity within $\mathcal{F}$, while HV measures both the convergence rate and policy diversity (Laumanns et al., 2002). Recall that, HV is defined as for any given $\mathcal{F}'$ an approximation of $\mathcal{F}$ and a reference point (the worst-possible return), it measures the volume of the hypercube spanned by the reference point and

Table 1: Hypervolume (HV) and Cardinality (CD) of various MORL methods on SC, RC, and MWP.

Methods	SC		RC		MWP	
	HV $(10^4)^{\uparrow}$	CD$^{\uparrow}$	HV (10^5)	CD	HV (10^9)	CD
PCN	1.81 ± 0.14	19.67 ± 2.99	11.69 ± 0.90	6.0 ± 1.27	10.17 ± 0.22	43.5 ± 1.06
GPI	2.82 ± 0.03	12.0 ± 2.05	7.33 ± 0.19	43.0 ± 2.62	10.44 ± 0.86	41.0 ± 2.83
Envelope	2.35 ± 0.18	5.6 ± 1.04	17.51 ± 3.73	19.75 ± 6.79	10.55 ± 1.96	51.5 ± 1.06
F-MDQ	2.22 ± 0.19	6.6 ± 1.31	16.92 ± 1.63	31.33 ± 7.84	10.45 ± 2.40	48.0 ± 2.12
FN-MDQ	2.34 ± 0.07	11.68 ± 1.05	20.38 ± 1.49	33.54 ± 8.29	10.51 ± 2.42	52.2 ± 2.44
FNS-MDQ	2.91 ± 0.20	15.38 ± 1.10	24.40 ± 2.22	36.11 ± 8.96	10.62 ± 2.45	51.05 ± 2.30

estimated return in a trajectory. Table 1 presents the HV and cardinality in all environments. These results show that our proposed methods perform on par with the considered baselines.

Question (D) The results discussed in previous questions suggest that our methods can generate a range of Pareto optimal solutions across varied preference configurations, which indicates better coverage of the objective space, thus leading to improved performance across multiple objectives. For quality, our proposed algorithms consistently achieve the lowest CV and highest GGF welfare scores across SC, RG, and MVP domains, indicating that our solutions exhibit more equitable distribution of objective utilities while maintaining Pareto optimality compared to baseline methods (PCN, GPI, and Envelope). These outcomes align with our theoretical justifications (see Section 4).

Question (E) Finally, to assess the impact of incorporating stochastic policies in MO Q-learning algorithms, we refer to the results in Figures 4a, 6a and 7a, where stochastic policies consistently improve both efficiency and fairness. Moreover, as shown in Table 1 incorporating stochastic policies also enhances MORL metrics, including HV and cardinality, validating the contribution of stochasticity to both fairness and overall performance.

7 Conclusions and Limitations

In this paper, we presented a novel approach to addressing fairness in the context of multi-policy MORL. Our proposed methods leverage a single parameterized network to learn optimized policies across the entire space of possible preferences. Both theoretical and empirical analyses demonstrate that learning a non-stationary policy significantly improves fairness. Additionally, we highlighted the importance of stochastic policies in achieving fair outcomes. Experimental evaluations in three domains validated the effectiveness of our approach in yielding more equitable policies compared to state-of-the-art MORL and fair baselines.

Our approach also has some limitations. First, it is limited to MOMDPs with discrete action spaces. Second, it assumes that preference weights are linear to learn the CCS, which may not capture the concave regions of the Pareto front. Third, the current formulation is focused on individual fairness. Given that optimizing a welfare function is a broad framework applicable to various real-world MORL problems involving general utilities, an important direction for future research is to extend this approach to accommodate more sophisticated objective functions, particularly those related to group-level fairness, safety, and risk sensitivity.

Acknowledgments

This work was supported by the Office of Naval Research under Grant N000142412405 and the Army Research Office under Grants W911NF2110103 and W911NF2310363.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

Lucas N Alegre, Ana LC Bazzan, Diederik M Roijers, Ann Nowé, and Bruno C da Silva. Sample-efficient multi-objective learning via generalized policy improvement prioritization. *arXiv preprint arXiv:2301.07784*, 2023.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *NeurIPS*, 30, 2017.

André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *NeurIPS*, 30, 2017.

Leon Barrett and Srini Narayanan. Learning all optimal policies with multiple criteria. In *ICML*, 2008.

X. Bei, S. Liu, C.K. Poon, and H. Wang. Candidate selections with proportional fairness constraints. In *AAMAS*, 2022.

Aurélie Beynier, Yann Chevaleyre, Laurent Gourvès, Ararat Harutyunyan, Julien Lesca, Nicolas Maudet, and Anaëlle Wilczynski. Local envy-freeness in house allocation problems. *AAMAS*, 2019.

Steven J. Brams and Alan D. Taylor. *Fair Division: From Cake-Cutting to Dispute Resolution*. Cambridge University Press, March 1996.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *ICML*, pp. 625–634, 2017.

Iadine Chadès, Janelle MR Curtis, and Tara G Martin. Setting realistic recovery targets for two interacting endangered species, sea otter and northern abalone. *Conservation Biology*, 26(6): 1016–1025, 2012.

M. Chakraborty, A. Igarashi, W. Suksompong, and Y. Zick. Weighted envy-freeness in indivisible item allocation. *TEAC*, 9(3):1–39, 2021.

Satya R. Chakravarty. *Ethical Social Index Numbers*. Springer Verlag, 1990.

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

Violet Xinying Chen and JN Hooker. A guide to formulating equity and fairness in an optimization model. *Preprint*, pp. 162–174, 2021.

Yann Chevaleyre, Paul E Dunne, Michel Lemaître, Nicolas Maudet, Julian Padget, Steve Phelps, and Juan A Rodríguez-aguilar. Issues in Multiagent Resource Allocation. *Computer*, 30:3–31, 2006.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *NeurIPS*, 30, 2017.

Alexandra Cimpeana, Catholijn Jonkerb, Pieter Libina, and Ann Nowéa. A multi-objective framework for fair reinforcement learning. In *Multi-Objective Decision Making Workshop 2023*, 2023.

Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In *NeurIPS*, 2021.

Virginie Do and Nicolas Usunier. Optimizing generalized gini indices for fairness in rankings. *arXiv preprint arXiv:2204.06521*, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, January 2012.

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

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *ICML*, pp. 1582–1591, 2018.

Hoda Heidari, Claudio Ferrari, Krishna P. Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *NeurIPS*, 2018.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *ICML*, pp. 1617–1626, 2017.

N. Jensen. An introduction to bernoullian utility theory, I: utility functions. *Swedish Journal of Economics*, 69:163–183, 1967.

Jiechuan Jiang and Zongqing Lu. Learning Fairness in Multi-Agent Systems. In *NeurIPS*, 2019.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

David Kurokawa, Ariel D. Procaccia, and Nisarg Shah. Leximin Allocations in the Real World. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pp. 345–362, June 2015. DOI: 10.1145/2764468.2764490.

M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multiobjective optimization. *Evolutionary Computation*, 10(3):263–282., 2002.

Jurek Leonhardt, Avishek Anand, and Megha Khosla. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*, pp. 101–102, 2018.

Debmalya Mandal and Jiarui Gan. Socially fair reinforcement learning. *arXiv preprint arXiv:2208.12584*, 2022.

Dimitris Michailidis, Willem Röpke, Diederik M Roijers, Sennay Ghebreab, and Fernando P Santos. Scalable multi-objective reinforcement learning with fairness guarantees using lorenz dominance. *arXiv preprint arXiv:2411.18195*, 2024.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *JMLR*, 15:3663–3692, 2014.

H. Moulin. *Fair Division and Collective Welfare*. MIT Press, 2004.

Yusuke Mukai, Yasuaki Kuroe, and Hitoshi Iima. Multi-objective reinforcement learning method for acquiring all pareto optimal policies simultaneously. In *IEEE International Conference on Systems, Man, and Cybernetics*, 2012.

Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *ICML*, 2019.

Samer B Nashed, Justin Svegliato, and Su Lin Blodgett. Fairness and sequential decision making: Limits, lessons, and opportunities. *arXiv preprint arXiv:2301.05753*, 2023.

Patrice Perny, Paul Weng, Judy Goldsmith, and Josiah Hanna. Approximation of Lorenz-optimal solutions in multiobjective Markov decision processes. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

M.L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. Wiley, 1994.

Junqi Qian, Umer Siddique, Guanbao Yu, and Paul Weng. From fair solutions to compromise solutions in multi-objective deep reinforcement learning. *Neural Computing and Applications*, pp. 1–31, 2025.

John Rawls. *The Theory of Justice*. Havard university press, 1971.

Mathieu Reymond, Eugenio Bargiacchi, and Ann Nowé. Pareto conditioned networks. *arXiv preprint arXiv:2204.05036*, 2022.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average Individual Fairness: Algorithms, Generalization and Experiments. In *NeurIPS*. 2019.

Umer Siddique, Paul Weng, and Matthieu Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *International Conference on Machine Learning*, 2020.

Umer Siddique, Abhinav Sinha, and Yongcan Cao. Fairness in preference-based reinforcement learning. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Umer Siddique, Peilang Li, and Yongcan Cao. Fairness in traffic control: Decentralized multi-agent reinforcement learning with generalized gini welfare functions. In *Multi-Agent reinforcement Learning for Transportation Autonomy*, 2024a.

Umer Siddique, Peilang Li, and Yongcan Cao. Towards fair and equitable policy learning in cooperative multi-agent reinforcement learning. In *Coordination and Cooperation for Multi-Agent Reinforcement Learning Methods Workshop*, 2024b.

Ashudeep Singh and Thorsten Joachims. Policy Learning for Fairness in Ranking. In *NeurIPS*. 2019.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.

Kristof Van Moffaert and Ann Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research*, 15(1):3483–3512, 2014.

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

Paul Weng. Fairness in reinforcement learning. In *AI for Social Good Workshop at International Joint Conference on Artificial Intelligence*, 2019.

J.A. Weymark. Generalized Gini inequality indices. *Mathematical Social Sciences*, 1:409–430, 1981.

R.R. Yager. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Trans. on Syst., Man and Cyb.*, 18:183–190, 1988.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. *NeurIPS*, 32, 2019.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with generalized gini welfare functions. In *Adaptive and Learning Agents (ALA) Workshop*, 2023a.

Guanbao Yu, Umer Siddique, and Paul Weng. Fair deep reinforcement learning with preferential treatment. In *ECAI*, 2023b.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*, 2017.

Xueru Zhang and Mingyan Liu. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*, pp. 525–555. Springer, 2021.

Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *ICML*, 2021.

Supplementary Materials

The following content was not necessarily subject to peer review.

8 Proofs of Technical Analysis

In this section, we provide formal proofs of our technical analysis in detail. For better legibility, we first recall the equations and results that we need for our proofs.

$$\forall \boldsymbol{\omega} \in \Omega, \quad \max_{\pi \in \Pi} \ \phi(\boldsymbol{V}(\pi)), \tag{3}$$

where Ω is the set of valid preference weights sorted in descending order, $\boldsymbol{V}(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}_t]$ is the expected discounted return, and $\phi(\boldsymbol{J}) = \sum_{i=1}^{N} w_i \boldsymbol{V}_{(i)}$ with $\boldsymbol{V}_{(1)} \leq \cdots \leq \boldsymbol{V}_{(n)}$.

Lemma 8.1. *For any MOMDP with linear preferences over objectives, the CCS contains an optimal policy for any linear combination of the objectives.*

Proof. Let $\mathcal{S}$ be the state space, $\mathcal{A}$ be the action space, and $\boldsymbol{r} : \mathcal{S} \times \mathcal{A} \to \boldsymbol{r}^N$ be the vector-valued reward function, where N is the number of objectives. Consider a linear preference vector $\boldsymbol{\omega} \in \Omega$, where $\Omega = \{\boldsymbol{\omega} \in \boldsymbol{r}^N : \sum_{i=1}^{N} w_i = 1, w_i \geq 0\}$. For any policy π, the expected return under a preference $\boldsymbol{\omega}$ is given by $\boldsymbol{\omega}\big(\mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} \boldsymbol{r}(s_t, a_t) \mid s_0 = s\right]\big)$. Thus, the optimal policy $\pi^*_{\boldsymbol{\omega}}$ for preference $\boldsymbol{\omega}$ satisfies

$$\pi^*_{\boldsymbol{\omega}} = \operatorname*{argmax}_{\pi} \boldsymbol{\omega}^T \boldsymbol{V}^{\pi}(s), \ \ \forall s \in \mathcal{S}.$$

By the definition of the CCS, for any $\boldsymbol{\omega} \in \Omega$, there exists a policy $\pi_{\text{CCS}} \in \text{CCS}$ such that

$$\boldsymbol{\omega}^T \boldsymbol{V}^{\pi_{\text{CCS}}}(s) \geq \boldsymbol{\omega}^T \boldsymbol{V}^{\pi}(s), \ \ \forall \pi \in \Pi, \forall s \in \mathcal{S}.$$

To prove the proposition, let's recall the Convex Hull Value Iteration (CHVI) algorithm (Barrett & Narayanan, 2008). Note that the CHVI algorithm iteratively updates the value function for each state by considering the convex hull of the achievable rewards via

$$\boldsymbol{V}(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \boldsymbol{P}(s' \mid s, a) \text{CH}\left(\boldsymbol{r}(s, a) + \gamma \boldsymbol{V}(s')\right),$$

where $\text{CH}(\cdot)$ denotes the convex hull operation. This update rule ensures that the value function $\boldsymbol{V}(s)$ lies within the convex hull of the achievable rewards and the $\text{CH}(\cdot)$ achievable value functions $\boldsymbol{V}^{\pi}(s) \mid \pi \in \Pi$ forms the CCS. Therefore, for any linear preference vector $\boldsymbol{\omega}$, there must exist at least a policy π_{CSS} such that

$$\boldsymbol{\omega}^T \boldsymbol{V}^{\pi_{\text{CCS}}}(s) = \max_{\pi \in \Pi} \boldsymbol{\omega}^T \boldsymbol{V}^{\pi}(s), \ \forall s \in \mathcal{S}.$$

The resulting policies form the CSS, which are sufficient to cover all linear preferences $\boldsymbol{\omega} \in \Omega$. Thus, for any linear combination of objectives, the optimal policy can be found within the CSS, confirming its sufficiency and optimality. □

While GGF introduces non-linear fairness objectives, its piecewise linearity and concavity allow representation as a maximum of linear functions, which ensures that solutions lie within the CCS. The following proposition establishes the sufficiency of the CCS in representing optimal policies for ϕ_{GGF} preference weights.

Proposition 8.1. *For any $s \in \mathcal{S}$ in an MOMDP and a piecewise-linear concave welfare function ϕ_{GGF} (e.g., GGF) that can be represented as, $\phi_{GGF}(\boldsymbol{V}^{\pi}(s)) = \min_{\sigma \in \mathbb{S}_N} \left\{\boldsymbol{\omega}_{\sigma}^{\top} \boldsymbol{V}^{\pi}(s)\right\}$, there exists a policy $\pi^* \in$ CCS such that:*

$$\phi_{GGF}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{GGF}(\boldsymbol{V}^{\pi}(s)) \quad \forall \pi \in \Pi.$$

Proof. Consider an arbitrary permutation $\sigma_A \in \mathbb{S}_N$. Since ϕ_{GGF} is a piecewise-linear and concave function, under a fixed permutation σ_A it becomes:

$$\phi_{\text{GGF}}(\boldsymbol{V}^{\pi}(s)) = \boldsymbol{\omega}_{\sigma_A}^{\top} \boldsymbol{V}^{\pi}(s).$$

Let $\pi_A \in \Pi$ be the policy that maximizes this linear scalarization:

$$\pi_A = \underset{\pi \in \Pi}{\operatorname{argmax}} \, \omega_{\sigma_A}^{\top} \boldsymbol{V}^{\pi}(s).$$

By the definition of CCS and the result from Lemma 8.1, there exist a $\pi^* \in$ CCS such that

$$\phi_{\boldsymbol{\omega}_{\sigma_A}}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{\boldsymbol{\omega}_{\sigma_A}}(\boldsymbol{V}^{\pi_A}(s)).$$

Thus,

$$\phi_{\boldsymbol{\omega}_{\sigma_A}}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{\boldsymbol{\omega}_{\sigma_A}}(\boldsymbol{V}^{\pi}(s)) \quad \forall \pi \in \Pi$$

Because this holds for any permutation $\sigma \in \mathbb{S}_N$, we cam conclude that for any policy $\pi \in \Pi$, there exists a corresponding $\pi^* \in$ CCS such that

$$\forall \pi \in \Pi, \quad \exists \pi^* \in CCS, \quad \phi_{\text{GGF}}(\boldsymbol{V}^{\pi^*}(s)) \geq \phi_{\text{GGF}}(\boldsymbol{V}^{\pi}(s)).$$

□

Fairness of Non-Stationary Policies. In fair MORL, learning non-stationary policies can be particularly beneficial, as they leverage historical information to make more informed decisions and adapt over time (see Section 4).

Proposition 8.2. *Let the reward $\boldsymbol{r}$ be nonnegative, and Π_S and Π_{NS} be the sets of stationary and non-stationary policies, respectively. For any $s \in \mathcal{S}$ in an MOMDP and a given ϕ_{GGF}, there exists a non-stationary policy $\pi_{NS} \in \Pi_{NS}$ that achieves a higher welfare score than any stationary policy $\pi_S \in \Pi_S$, i.e.,*

$$\exists \pi_{NS} \in \Pi_{NS} : \phi_{GGF}(\boldsymbol{V}^{\pi_{NS}}(s)) \geq \max_{\pi_S \in \Pi_S} \phi_{GGF}(\boldsymbol{V}^{\pi_S}(s))$$

Proof. Let the state value function be defined by:

$$\boldsymbol{V}(s) = \mathbb{E}\left[\boldsymbol{G}_t \middle| \, s_t = s\right]$$

where the return $\boldsymbol{G}_t$ is given by:

$$\boldsymbol{G}_t = \sum_{k=0}^{\infty} \gamma^k \, \boldsymbol{r}_{t+k+1}.$$

Suppose an episode begins at time t and terminates at time T_{end}. For any intermediate time T with $t \leq T < T_{\text{end}}$, we can decompose the return into two parts:

$$\boldsymbol{G}_t = \underbrace{\boldsymbol{r}_{t+1} + \gamma \boldsymbol{r}_{t+2} + \cdots + \gamma^{T-t-1} \boldsymbol{r}_T}_{\boldsymbol{G}_t^{(1)}} + \underbrace{\gamma^{T-t} \left(\boldsymbol{r}_{T+1} + \gamma \boldsymbol{r}_{T+2} + \dots\right)}_{\boldsymbol{G}_t^{(2)}}.$$

With above decomposition, We define value function as two parts:

$$\text{Early-period value function: } \boldsymbol{V}_1(s) = \mathbb{E}\left[\boldsymbol{G}_t^{(1)} \middle| \, s_t = s\right]$$

$$\text{Late-period value function: } \boldsymbol{V}_2(s) = \mathbb{E}\left[\boldsymbol{G}_t^{(2)} \middle| \, s_T = s\right]$$

so that

$$\boldsymbol{V}(s) = \boldsymbol{V}_1(s) + \gamma^{T-t}\boldsymbol{V}_2(s)$$

At time T, stationary policy π_S selects action solely based on late period value function $\boldsymbol{V}_2(s)$, while non-stationary policy has access to both early $\boldsymbol{V}_1(s)$ and late period value function $\boldsymbol{V}_2(s)$ and can condition its action selection on the combined information given by two value functions.
Under a stationary policy, The total value can be presented as:

$$\boldsymbol{V}^{\pi_S}(s) = \boldsymbol{V}_1(s) + \gamma^{T-t} \underset{\boldsymbol{V}_2(s)}{\operatorname{argmax}}\{\phi_{\text{GGF}}[\boldsymbol{V}_2(s)]\}$$

In contrast, under a non-stationary policy the total value is given by

$$\boldsymbol{V}^{\pi_{NS}}(s) = \underset{\boldsymbol{V}_1(s),\boldsymbol{V}_2(s)}{\operatorname{argmax}} \{\phi_{\text{GGF}}[\boldsymbol{V}_1(s) + \gamma^{T-t}\boldsymbol{V}_2(s)]\}$$

therefore:

$$\exists \pi_{\text{NS}} \in \Pi_{\text{NS}} : \phi_{\text{GGF}}(\boldsymbol{V}^{\pi_{\text{NS}}}(s)) \geq \max_{\pi_{\text{S}} \in \Pi_{\text{S}}} \phi_{\text{GGF}}(\boldsymbol{V}^{\pi_{\text{S}}}(s))$$

This completes the proof. □

Optimality of Stochastic Policies for Fairness Unlike the single-objective scenario, in MORL, a deterministic policy may not be optimal. A fairer solution can often be achieved through randomization.

Proposition 8.3. *Let Π_{ST} be the set of stochastic policies and Π_D be the set of deterministic policies. For an MOMDP $\mathcal{M}$ and a concave welfare function such as ϕ_{GGF}, there exists a stochastic policy $\pi_{ST} \in \Pi_{ST}$ such that:*

$$\phi_{GGF}(\boldsymbol{V}^{\pi_{ST}}) \geq \max_{\pi_D \in \Pi_D} \phi_{GGF}(\boldsymbol{V}^{\pi_D}).$$

Proof. The key idea here is that a stochastic policy can represent a convex combination of deterministic policies for any concave welfare function ϕ_{GGF} Busa-Fekete et al. (2017). Hence, stochastic policies can achieve outcomes in the objective space that are unattainable by deterministic policies. Specifically, for ϕ_{GGF}, a deterministic policy π_{D} yields a fixed utility vector $\boldsymbol{V}^{\pi_{\text{D}}}$ while a stochastic policy π_{ST} can yield a distribution over utility vectors. Thanks to concavity of ϕ_{GGF}, which makes our problem in 2 convex optimization and Jensen's inequality (Jensen, 1967), we obtain

$$\phi_{\text{GGF}}\left(\mathbb{E}_{\tau\sim\pi}[\boldsymbol{V}^{\pi_{\text{st}}}]\right) \geq \mathbb{E}_{\tau\sim\pi}\left[\phi_{\text{GGF}}(\boldsymbol{V}^{\pi_{\text{st}}})\right]. \tag{4}$$

Since ϕ_{GGF} is a piecewise linear concave function, there exists a stochastic policy π_{st} that is a convex combination of deterministic policies such that

$$\mathbb{E}_{\tau\sim\pi}[\phi(\boldsymbol{V}^{\pi_{\text{st}}})] \geq \max_{\pi_{\text{d}} \in \Pi_{\text{D}}} \phi(\boldsymbol{V}^{\pi_{\text{d}}}). \tag{5}$$

By combining (4) and (5), we can obtain

$$\phi(\mathbb{E}_{\tau\sim\pi}[\boldsymbol{V}^{\pi_{\text{st}}}]) \geq \mathbb{E}_{\tau\sim\pi}[\phi(\boldsymbol{V}^{\pi_{\text{st}}})] \geq \max_{\pi_{\text{d}} \in \Pi_{\text{D}}} \phi(\phi^{\pi_{\text{d}}})].$$

This completes the proof. □

The optimality of stochastic policies implies that restricting the search for fair solutions to deterministic policies is insufficient. Stochastic policies offer a broader range of solutions and may better capture the trade-offs among multiple objectives, enhancing the overall fairness of the policy.

9 Fairness

In a fair single-policy setting, where the goal is to learn a single policy treating all users equally, three fairness principles, efficiency, equity, and impartiality, are defined below.

Definition 9.1. *Efficiency states that among two feasible solutions, if one solution is (weakly or strictly) preferred by all users, then it should be preferred to the other one, e.g.,* $\boldsymbol{u} \succ \boldsymbol{u}' \Rightarrow \phi(\boldsymbol{u}) > \phi(\boldsymbol{u}')$*, where* $\phi(\boldsymbol{u})$ *is the scalar utility function that specifies the value of a solution.*

Intuitively, the efficiency property specifies that given all else equal, one prefers to increase a user's utility. In the MORL setting, the efficiency property simply means Pareto dominance. More specifically, a solution is considered efficient if it is not dominated by any other solution for all objectives.

Next, we discuss the significance of the *equity* property, which is a stronger property than efficiency and is often associated with distributive justice, as it refers to the fair distribution of resources or opportunities. This property ensures that a fair solution follows the *Pigou-Dalton principle* (Moulin, 2004), which states the transferring of rewards from the more advantaged users to the less advantaged users.

Definition 9.2. *A solution satisfies the Pigou-Dalton principle if for all* $\boldsymbol{u}$*,* $\boldsymbol{u}'$ *equal except for* $u_i = u'_i + \delta$ *and* $u_j = u'_j - \delta$ *where* $u'_i - u'_j > \delta > 0$*,* $\phi(\boldsymbol{u}) > \phi(\boldsymbol{u}')$*.*

Finally, we discuss the *impartiality* property. This property is rooted in the principle of "equal treatment of equals", which states that individuals sharing similar characteristics should be treated similarly.

Definition 9.3. *In a system, individuals with similar characteristics should be treated similarly, i.e., the solution should be independent of the order of its arguments* $\phi(\boldsymbol{u}) = \phi(\boldsymbol{u}_\sigma)$*, where* σ *is a permutation and* $\boldsymbol{u}_\sigma$ *is the vector obtained from vector* $\boldsymbol{u}$ *permuted by* σ*.*

9.1 Welfare Function

A welfare function, denoted as $\phi : \mathbb{R}^D \to \mathbb{R}$, aggregates the utilities of all users (or objectives) and offers a metric of the overall desirability of a solution for the entire group, where $\boldsymbol{\omega}$ represents the set of aggregation weights for all objectives. One well-established welfare function used in this paper is the generalized Gini welfare function. The generalized Gini welfare function constitutes a specific instance of the ordered weighted average (OWA)(Yager, 1988). It is a renowned welfare function employed in multi-objective optimization (Weng, 2019; Siddique et al., 2020; Zimmer et al., 2021; Do & Usunier, 2022; Yu et al., 2023a;b; Siddique et al., 2023), initially devised to quantify income distribution inequality in economics (Weymark, 1981). The generalized Gini welfare function is defined as follows:

$$\phi_{\text{GGF}}(\boldsymbol{u}) = \sum_{i=1}^{N} \omega_{\sigma(i)} u = \boldsymbol{w}_\sigma^T \boldsymbol{u}\,, \tag{6}$$

where $\sigma \in \mathbb{S}_N$, which depends on $\boldsymbol{\omega}$, is the permutation that sorts the components of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_\sigma = (\omega_{\sigma(1)}, \ldots, \omega_{\sigma(N)})$. Equation (6) holds as the weights are rearranged based on the utility vector, assigning the largest weight to the smallest component of $\boldsymbol{u}$, the second-largest weight to the second-smallest component of $\boldsymbol{u}$, and so forth.

The generalized Gini welfare function satisfies the three fairness properties. Due to the positive weights, it is monotonically related to Pareto dominance, fulfilling the efficiency property. Moreover, the reordering of the components in the welfare function makes it symmetric with respect to its components, satisfying the impartiality property. Lastly, as the generalized Gini weights are positive and decreasing, it is Schur-concave, meeting the equity property.

Among numerous welfare functions, the generalized Gini welfare function possesses several favorable properties, namely, simplicity as it is a weighted sum in the Lorenz space (Chakravarty, 1990;

Perny et al., 2013), well-understood properties axiomatized by Weymark (1981), and generality. These favorable properties make it a suitable choice for addressing the challenge of finding fair solutions. Moreover, it is notably a concave function, which will make the solution to our problem easier.

To emphasize the versatility of the generalized Gini welfare function, various special cases can be derived by adjusting its weights accordingly. These cases include:

- **Maxmin fairness**: Setting $\omega_1 = 1$ and $\omega_i = 0$ for $i = 2, \cdots, K$ corresponds to the maxmin notion of fairness (Rawls, 1971).
- **Regularized maxmin fairness**: Assigning $\omega_1 = 1$ and $\omega_i = \varepsilon$ for $i = 2, \cdots, K$ aligns with the regularized maxmin notion of fairness.
- **Utilitarian approach**: Setting $\omega_i = 1/K$ represents the utilitarian approach.
- **Leximin fairness**: If the ratio ω_j/ω_{j+1} tends toward infinity, it corresponds to the leximin notion of fairness (Rawls, 1971; Kurokawa et al., 2015).

10 Descriptions of Environments

10.1 Species Conservation

In the field of ecology, the challenge of conserving interdependent endangered species is paramount. The simulation environment focuses on the balance required in the conservation of two such species: the sea otter and the northern abalone, which are currently endangered. The predation relationship between these species, with sea otters feeding on abalones, presents a unique challenge that requires careful consideration of fairness and equity in conservation efforts. Based on the framework in (Chadès et al., 2012), we define the state space as the current population numbers of the sea otters and northern abalones. The action space consists of: introducing sea otters, enforcing antipoaching measures, controlling sea otter populations, implementing a combination of half-antipoaching and half-controlled sea otters, or taking no action. Each action carries significant ecological consequences; for instance, while the reintroduction of sea otters is essential for maintaining the abalone population, it must be carefully managed to prevent the abalone's extinction. Conversely, overlooking other management actions could lead to the demise of either species. The transition function employed in our model accounts for population dynamics, including external threats such as poaching and oil spills. Since our objective is to optimize the population densities of both species, we define the reward function as the densities of both species, i.e., $N = 2$.

10.2 Resource Gathering

In this scenario of resource gathering, we consider a 5×5 grid world domain inspired from (Barrett & Narayanan, 2008). This domain presents a unique challenge centered around the acquisition of three types of resources: gold, gems, and stones, thereby establishing a multi-objective framework with $\mathcal{K} = 3$. The autonomous agent is positioned within this grid world, and resources are distributed randomly across various locations. As a resource is collected by the agent, it is immediately regenerated at a new random location within the grid, ensuring a perpetual availability of resources. In this problem, the state is characterized by the agent's current location on the grid and a cumulative count of each type of resource collected over the course of the agent's trajectory. The agent can navigate the grid through actions aligned with the four cardinal directions: up, down, left, and right, facilitating movement across the grid. To add complexity to the resource management challenge, resources are assigned differing values, reflecting their relative importance. Specifically, gold and gems are attributed a value of 1, underscoring their significance, whereas stones are considered less valuable, with a value of 0.4. This valuation leads to an intentionally uneven distribution of resources within the grid, comprising two stones, one gold, and one gem. This configuration is designed to simulate a scenario where the agent must not only maximize the collection of resources but also achieve a balanced acquisition across the different types of resources. The overarching objective for the agent in this environment is dual: to maximize the total value of resources collected while

ensuring an equitable collection across the various resource types. Achieving this balance is crucial for optimizing the agent's resource-gathering strategy, enhancing its overall utility and adaptability within the dynamic grid world. This nuanced approach to resource management in a simulated environment offers insights into the complexities of resource distribution and acquisition strategies, contributing to the broader discourse on multi-objective optimization in dynamic settings.

10.3 Multi-Product Web Advertising

We now consider the multi-product web advertising (MWP) problem, where an online store offers N distinct types of products for sale and an intelligent agent makes strategic decisions at each timestep about which advertisement to display: a product-specific advertisement for one of the products $i \in [0, ..., N-1]$, or a general advertisement that is not tailored to any specific product. The effectiveness of an advertisement is contingent upon its relevance to the customer's recent web activity, with appropriate advertisements significantly increasing the likelihood of a purchase, whereas inappropriate ones may deter the customer altogether. The state space of this problem is defined by the number of products available in the store, augmented by the number of visits, purchases, and exits. A visit state indicates a customer's interest in a particular product, a purchase state signifies the completion of a transaction, and an exit state occurs when a customer leaves the website without making a purchase. The action space is expanded to $n+1$ actions, where actions 0 through n correspond to displaying advertisements for specific products, and action n represents the option to show a general advertisement that does not target any specific product in the inventory. This additional action introduces an additional layer of complexity, as the agent must decide the optimal moment to transition between states. The reward function is designed such that the agent receives a reward of 1 in the i^{th} dimension of the reward vector if a product of type i is sold after the display of its advertisement. The primary objective of this problem is to maximize the aggregate returns from product sales while striving for an equitable distribution of sales across the different product types. This goal underscores the need for fair solutions that not only optimize overall profitability but also ensure a balanced representation of product sales, thereby addressing the dual challenges of efficiency and equity in this domain.

11 Hyperparameters

To ensure reproducibility, we have meticulously documented all hyperparameters across different environments in Tables 1,2,3, and 4. We utilize the well-known high-quality MORL baselines[1] for implementing baseline algorithms. In these tables, we present the hyperparameters corresponding to Envelope, GPI, PCN, and our proposed algorithms in three distinct environments, namely, species conservation (SC), resource gathering (RC), and multi-web product advertising (MWP).

[1] https://github.com/LucasAlegre/morl-baselines

Table 2: Set of hyperparameters used for training Envelope.

Hyperparameter	SC	RC	MWP
Discount factor (γ)	0.99	0.99	0.99
Learning rate (α)	0.0001	0.0005	0.005
Batch size	64	64	64
Hidden Layers	256 x 256 x 256 x 256	256 x 256 x 256 x 256	256 x 256 x 256 x 256
Buffer Size	50000	50000	50000
Initial Epsilon	1.0	1.0	1.0
Final Epsilon	0.05	0.05	0.05
Epsilon Decay Steps	50000	50000	50000
Learning Starts	100	100	100
Gradient Updates	1	1	5
Max Gradient Norm	1.0	1.0	1.0
Ω Distribution	Gaussian	Gaussian	Gaussian
Tau	0.5	0.5	0.5

Table 3: Set of hyperparameters used for training our proposed methods.

Hyperparameter	SC	RC	MWP
Discount factor (γ)	0.99	0.99	0.99
Learning rate (α)	0.0001	0.0005	0.005
Batch size	64	64	64
Hidden Layers	256 x 256 x 256 x 256	256 x 256 x 256 x 256	256 x 256 x 256 x 256
Buffer Size	50000	50000	50000
Initial Epsilon	1.0	1.0	1.0
Final Epsilon	0.05	0.05	0.05
Epsilon Decay Steps	50000	50000	50000
Learning Starts	100	100	100
Gradient Updates	1	1	5
Max Gradient Norm	1.0	1.0	1.0
Ω Distribution	Gaussian	Gaussian	Gaussian
Tau	0.5	0.5	0.5

Table 4: Set of hyperparameters used for training GPI.

Hyperparameter	SC	RC	MWP
Discount factor (γ)	0.99	0.99	0.99
Learning rate (α)	0.0001	0.0005	0.005
Batch size	128	128	256
Hidden Layers	256 x 256 x 256 x 256	256 x 256 x 256 x 256	256 x 256 x 256 x 256
Num Networks	2	2	2
Buffer Size	50000	50000	50000
Initial Epsilon	1.0	1.0	1.0
Final Epsilon	0.05	0.05	0.05
Epsilon Decay Steps	50000	50000	50000
Learning Starts	100	100	100
Gradient Updates	1	1	5

Table 5: Set of hyperparameters used for training PCN.

Hyperparameter	SC	RC	MWP
Discount factor (γ)	0.99	0.99	0.99
Learning rate (α)	0.0001	0.0001	0.0005
Batch size	128	256	128
Hidden Layers	64 x 64	64 x 64	64 x 64
Desired Return	[1, 1]	[200, 200, 200]	[100, 100, 100, 100, 100]
Buffer Size	500000	500000	1000000
Max Horizon	5000	1000	1000